10

STRUCTURAL EQUATION MODELING IN L2 RESEARCH

Rob Schoonen

If there is one thing that we know in second language (L2) research, it is that there are many factors involved in L2 learning and use. These factors are found in very complex relationships, which may even change with increasing language proficiency. These relationships are far more complex than what we can describe with the computation of a series of simple bivariate correlations. L2 researchers have to be able to deal with multivariate analyses of data. Structural equation modeling provides a framework to investigate these complex multivariate relationships.

Conceptual Motivation

Structural equation modeling (SEM), also known as causal modeling, covariance structure analysis, or LISREL analysis, has as its distinguishing feature that it requires some sort of modeling. Modeling implies that researchers need to be explicit about the relationships they envisage between measured variables and underlying constructs (i.e., latent variables) and between the constructs themselves. Therefore, a researcher has to think carefully about the hypothesized relationships before embarking on a SEM enterprise. SEM provides the researcher with a toolbox that can uncover complex relationships that go well beyond the bivariate relations as expressed in a correlation or a simple regression, but also beyond the multivariate relationships that are usually addressed in a multiple regression analysis (see Jeon, Chapter 7 in this volume).

SEM can be used at various stages of theory development, ranging from confirmatory testing to exploration. More specifically, Jöreskog and Sörbom (1996) mention three situations for fitting and testing models. First is a strictly confirmative situation, where there is a single model that is put to the test with empirical data. The model is either accepted or rejected. Second is testing alternative or competing models, when a researcher wants to choose between two or three concurrent models on the basis of a single data set. A third use is a model-generating situation, when a researcher starts off with an initial model and then tries to improve it on the basis of (mis)fit results (Jöreskog & Sörbom, 1996, p. 115). The result of a model-generating situation should not be taken as a real statistical testing of the (final) model, and the process of model improvement should not only be guided by statistical outcomes but also by substantive theoretical considerations. The resulting model should then be put to the test anew with different data (creating a new, confirmatory situation).

The possibilities in a SEM analysis seem to be unlimited (see Hancock & Schoonen, 2015), and the flexibility of the approach to address them makes SEM a very attractive analytic framework, leading to an increase in recent years in the use of SEM in L2 research (Plonsky, 2014). However, it is not difficult to imagine that these options also contain a risk for using the technique uncritically (see the "Pitfalls" section in this paper). Therefore, it is crucial that the user has theoretical guidance with respect to the research questions he or she wants to investigate and the analytic choices that need to be made. Lewin's well-known quote that there is "nothing so practical as a good theory" applies here for sure.

SEM is a collection of analyses that can be used to answer many research questions in L2 research. Prominent is the use of SEM to predict (or "explain") complex constructs, such as reading and writing proficiency, or the development of these complex proficiencies, on the basis of scores on component skills. Other studies investigate the complex relations between related constructs, such as motivation and attitude toward foreign languages. At the initial stage, modeling these kinds of relationships, a researcher could start with drawing graphs depicting how constructs influence each other, or how they are related, using unidirectional or bidirectional arrows, respectively, to connect the constructs. To make it more concrete, the constructs could be connected to measured, observed or manifest variables. Conventionally, underlying or latent variables are represented as circles or ovals, and observed variables as rectangles (see Figure 10.2–10.3). SEM is also highly flexible, able to deal with multiple dependent variables and multiple independent variables. These variables can be continuous, ordinal, or discrete, and they can be indicated as observed variables (i.e., observed scores) or as latent variables (i.e., the underlying factor of a set of observed variables) (Mueller & Hancock, 2008; Ullman, 2006). Examples of complex models in L2 studies can be found in, for instance, Gu (2014), Schoonen, Van Gelderen, Stoel, Hulstijn, and De Glopper (2011) or Tseng and Schmitt (2008). Which measured and latent variables, and which relations to include in the SEM analysis is up to the researcher. We should keep in mind that statistical techniques per se cannot make substantive decisions. As is the case with nearly all analyses described in this volume, SEM requires a number of choices to be made by the researcher, and these choices must be made on solid theoretical grounds.

In the remainder of this chapter a number of examples will be presented to illustrate the possibilities of SEM. Furthermore, a more detailed sample analysis will be provided using two different software packages, LISREL and AMOS. Readers interested in other packages or more extensive introductions to the available software are referred to the corresponding manuals or specialized introductions (Byrne, 1998, 2006, 2010, 2012). Readers who want to learn more about SEM than this chapter can offer, or who want to know more about the theoretical underpinnings of SEM, will find suggestions for further reading at the end of this chapter.

Two Parts of a Model: Measurement and Structure

Testing the relationships that one postulates or expects between the theoretical variables (as opposed to measured variables) is just one part of an analysis with SEM, often referred to as the *structural model*. In a structural model one can design hypothesized relations between theoretical variables. For example, does Language Exposure directly influence a language learner's Language Development or is this presumed effect mediated by Working Memory Capacity? Or maybe a researcher wants to compare the tenability of these two concurrent hypotheses (see Figure 10.1). These kinds of research questions relate to the structural part of the model.

However, a researcher can address these issues only provided that he or she has reliable and valid measures for the latent theoretical variables involved. From L2 research we know that adequate measurement of core variables is almost never as straightforward as we would like it to be. An important part of SEM analysis therefore concerns the modeling of the measurement of theoretical variables or constructs. These measurement concerns are addressed in what is referred to as the *measurement model*. The main question here is: What are appropriate measures for the constructs or latent variables one intends to measure? In our example, we need measures for Language Exposure, Language Development and Working Memory to investigate our hypotheses, and thus the model needs to be extended with the measured or observed variables involved (see Figure 10.2). The number of observed variables needed to operationalize a latent variable depends on other features of the model, but three measures will suffice in most cases (see Kline, 2010).

Although the measurement part of the model seems to be a psychometric issue, decisions about construct operationalization get at the heart of validity research, which makes it a substantive issue. For example, in a study about the

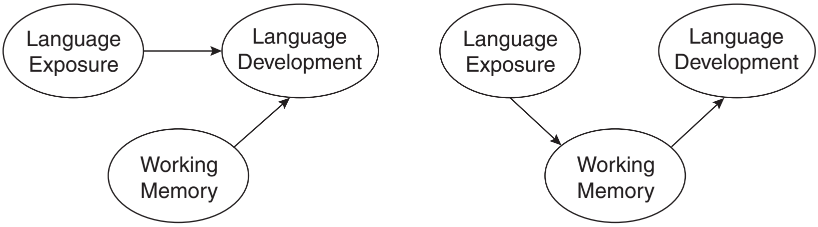


FIGURE 10.1 Two competing structural models

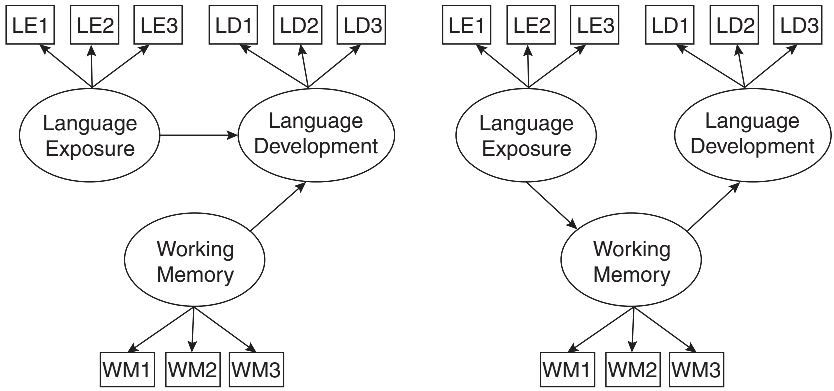


FIGURE 10.2 Two competing structural models with measurement part added

relationship between linguistic ability and some other construct, a researcher has to decide whether linguistic ability can be measured by vocabulary, grammatical knowledge, and pragmatic knowledge together or whether these three domains should be kept separate and should be measured each on their own. This latter type of research question is what is often treated as a confirmatory factor analysis (CFA) problem (see Ockey, 2014). In other words: Do the measures involved measure a single construct or do they measure multiple constructs?

Underlying Factors

When one wants to investigate the underlying structure of a set of variables, for example the subtests of a test battery, one can use SEM to actually test hypotheses about the number of factors that are underlying and also about their interrelations. Key is the testing of hypotheses, which implies that one has a priori one or a few (competing) expectations that can be put to the test. This is different from, for example, exploratory factor analysis (EFA) or principal component analysis (PCA), where in a data-driven way the number of underlying factors (or components) is determined according to a statistical criterion (Ockey, 2014; Loewen & Gonulal, Chapter 9 in this volume). Using SEM, one has to model the relationship between the measured variables and the hypothesized factors (i.e., latent variables) and subsequently test the fit of the model to the empirical data. This makes it a CFA. An advantage of the SEM framework is that the relations between selected factors can be modeled in the structural part of the model.

Imagine, for example, a second-language ability test battery that consists of nine tests: Grammaticality Judgments (V1), Resolution of Anaphors (V2), Understanding of Conjunctions (V3), Vocabulary Size (V4), Depth of Vocabulary Knowledge (V5), Knowledge of Metaphors (V6), Sentence Comprehension (V7), Use of Verb Inflection (V8), and Use of Agreement (V9). A researcher could question, for example, whether the nine test scores are best described (or explained) by one underlying general L2 linguistic skill, or whether a three-factor model with a metacognitive-metalinguistic factor, a lexical-semantic factor, and a morpho-syntactic factor is more plausible.

Figure 10.3 depicts both competing models. In the first model (left panel), the observed variables V1 through V9 are dependent of one (latent) underlying variable, "General L2 Linguistic Factor," and there is some unexplained residual variance (*e*) in each observed variable as indicated by the arrows coming from *e* (error). In the second model (right panel) three underlying constructs are postulated, namely, (a) metacognitive, (b) lexical-semantic, and (c) morpho-syntactic factor. Scores on three instruments (tests) are considered indicative of metacognitive proficiency: Grammaticality judgments (V1), the resolution of anaphors (V2), and understanding of conjunctions (V3). Vocabulary size (V4), depth of vocabulary knowledge (V5), knowledge of metaphors (V6), and sentence comprehension (V7) are assumed to be typical indicators of lexical-semantic proficiency, and scores on the verb inflection (V8) and the agreement test (V9) are typical indicators of morpho-syntactic proficiency. Ideally, one would prefer more than two measured variables to indicate the latent variable morpho-syntactic proficiency. So, the model postulates that there is reason to assume that there are three underlying latent variables involved in the test performances. The model further indicates that these latent variables are not fully unrelated to each other because there are double-headed arrows indicating covariance between the three factors, covariance being the unstandardized equivalent of correlation. The covariance between residual e_5 and e_7 will be introduced later on.

Advantages of SEM

The example in Figure 10.3 largely deals with the way one defines and measures the theoretical variables (cf. CFA) and as such is considered part of the measurement model. One of the advantages of SEM is that one can test the fit of the hypothesized model against one's data, and one can also compare and test the difference in fit between the two competing models described later in this chapter.

There are at least two other advantages to using SEM in these kinds of analyses. First, researchers are more or less forced to come up with hypotheses about relationships between their measurements (observed scores) and underlying constructs or latent variables. Most hypotheses in L2 research involve variables that are not directly observable, such as language proficiency, working memory capacity, speaking proficiency, and so on. However, in the actual empirical investigation researchers want to test the tenability of their claims about these latent underlying variables. Putting forward a measurement model makes this part of studies more explicit and thus more open for empirical scrutiny and discussion. In some cases theoretically relevant variables can be measured more directly, such as age or parental education. In such cases, the observed and latent variables coincide.

Another advantage of SEM pertains to the more substantive analyses in the structural part of the model. Once one has modeled the collected data in a well-fitting

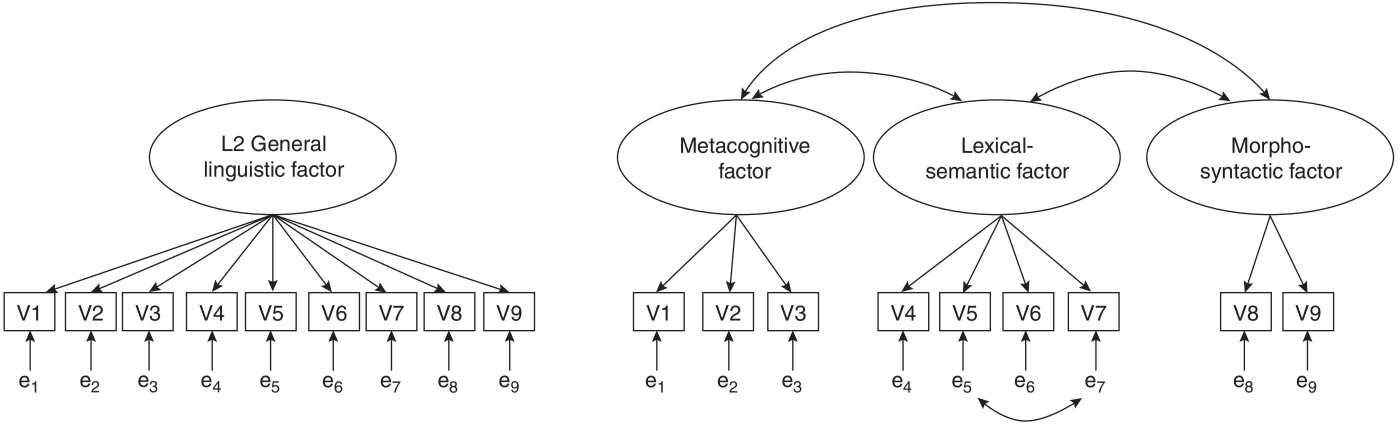


FIGURE 10.3 Two competing models: a one-factor model and a three-factor model

measurement model, one can test substantive hypotheses with latent variables that are so-called error-free. From Figure 10.3 one can see that the latent variables are determined by the covariance of the different measured variables (V1–V9 in the left panel or V1–V3, V4–V7 and V8–V9, respectively, in the right panel) and thus that the idiosyncrasies of the measurements, including measurement error (e_1–e_9), are partialed out (excluded). This way an analysis of the relations of latent variables in the structural model, not being attenuated by measurement error, can provide a clearer picture of what these relations are (see Mueller & Hancock, 2008, for an example). In the structural part of our three-factor model, the researcher can investigate whether the three factors simply covary as depicted in Figure 10.3 or show more specific relations. For example: Is the metacognitive knowledge the result of lexical-semantic and morpho-syntactic proficiency? To test such a hypothesis the relationship between the three factors should be modeled as regressions (with one-directional arrows) in which metacognitive knowledge is the dependent variable and lexical-semantic and morpho-syntactic proficiency are the independent variables (analogous to Figure 10.2; see also Jeon, Chapter 7 in this volume). Alternatively, one could also claim that the three factors are unrelated. This would lead to a model without any connections between the three factors, or—in other words—covariances of 0. Comparison of the fit of the various models to the available data as described later in this chapter will suggest which model is most plausible.

The previous example is—for practical reasons—kept simple, but numerous multiple regression models with single as well as multiple dependent variables in all kinds of different configurations can be analyzed if there are good substantive reasons to do so (see Tseng & Schmitt, 2008; Schoonen et al., 2003; Gu, 2014). One could say that SEM elegantly combines factor-analytic procedures with regression-analytic ones (and many more, see Hancock & Schoonen, 2015, for examples in the L2 field; in addition, Rovine & Molenaar, 2003, show all kinds of variance-analytic applications of SEM). However, this flexibility requires substantial sample sizes, data that meet certain requirements, and a clear plan for the analyses, because the number of possibilities for the analyses are sometimes overwhelming. In the next section, we will go into more detail as we discuss SEM analyses step by step. First, we will focus on general principles and considerations at the successive stages in SEM analyses. Second, we will have a closer look at what an analysis looks like in two of the available packages for SEM analyses (see the next section): LISREL, being one of the earlier and well-developed packages, and AMOS, being part of the IBM SPSS family of packages.

General Considerations in SEM Analyses

In the previous section, it was said that SEM can combine factor-analysis, regression-analysis, and much more. In this chapter we confine ourselves to modeling relationships between measured variables and latent variables and latent variables among each other; we will ignore the possibility of modeling (latent) mean scores. Furthermore, we will focus on the modeling of interval or continuous data, such as test scores and reaction time data. For other applications we refer to more extensive introductions as mentioned at the end of this chapter. Hancock and Schoonen (2015) discuss a number of possible applications in the field of second language acquisition and applied linguistics.

Data Preparation

The data for the SEM analysis have to meet certain requirements for a straightforward analysis. For the procedures to work well and for the testing and parameter estimation to be reliable, the continuous variables should be multivariate normally distributed. Among other things (see Kline, 2010), this means that the individual variables are univariate normally distributed. So, initial screening of the data is relevant for a valid interpretation of the outcomes of a SEM analysis. This includes checks on skewness and kurtosis of variables, but also outliers can affect an analysis in a detrimental way. Bivariate plots for pairs of variables give a first impression of possible violations of a multivariate normal distribution. For an overview of multivariate assumptions and data preparation, see Jeon (Chapter 7 in this volume). If data violate assumptions for SEM, especially multivariate normality, the researcher can resort to other estimation methods within the SEM framework or apply corrections to the outcome statistic (χ^2) and the standard errors for the estimated parameters (Satorra-Bentler's scaled version). See West, Finch and Curran (1995) or Finney and DiStefano (2013) for an extensive discussion about the assumptions in SEM and possible alternatives in case these assumptions are violated.

In L2 research, as in other empirical domains, data sets are seldom complete. There are several ways to deal with missing data, such as listwise deletion of cases with missing data or estimation of a missing score on the basis of available scores. Listwise deletion avoids controversial imputation of estimated scores. This approach, however, is advisable only in cases where (a) data are assumed to be missing completely at random and where (b) the sample is large enough to endure the resulting loss of statistical power. Imputation of missing values can be a good alternative, but has its drawbacks as well. For example, replacing the missing score by the sample mean will reduce the score variance, an important source of information in modeling. Fortunately, there are more advanced procedures for dealing with missing data. Most software packages for SEM have their own provisions for handling missing data that are very sophisticated, so it might be wise to consider their options (Kline, 2010; for a more thorough discussion see Enders, 2013). Working with incomplete data implies that one works with the raw data (including missing value codes), and not with just a correlation or covariance matrix as input data. However, using a correlation or covariance matrix as the input data for an analysis is a viable option if one wants to replicate analyses from the literature and only a covariance matrix or a correlation matrix (preferably with corresponding means and standard deviations) is available (see the next section and Discussion Question 8).

Designing a Model

After preparing the data, the most exciting part of the analysis begins: designing the model. This process should be guided by theoretical considerations and expectations, and can best be split into two stages (Mueller & Hancock, 2008). The first stage involves testing the measurement model, which helps us determine whether the presumed latent variables are measured by the observed test scores in the expected way. At this stage no constraints are implemented regarding the relationships among the latent variables, so that any misfit of the model is due to the way the latent and observed variables were presumed to be related in the model.

The latent variables being latent, do not have a scale of themselves. To solve this, one can either standardize the latent variable by fixing its variance at 1 (cf. *z*-values) or equate the scale to that of one of the observed variables, a so-called reference variable. In the latter case the regression weight for the observed variable on the latent variable is fixed at a value of 1. Both solutions are equivalent.

If the fit of the measurement model is satisfactory (that is, the model fits well) and all observed measures can—to a reasonable extent—be explained by their underlying variables, one can move on to the second stage: modeling the relationships among the latent variables. However, if the measurement model does not fit satisfactorily, the relations between the measured variables and the underlying variables needs to be reconsidered. A variable might not be related to the underlying variable(s) in the expected way, or a variable may show only a weak relation to the underlying variable(s). Validity and/or reliability issues could be involved if a measured variable does not fit the hypothesized relations.

At the second stage, when the structural model is developed, one can test the substantive hypotheses about the theoretical constructs, either as a single model or as competing models that can be compared to select the best model. There are often many possibilities for modeling relationships between variables, especially in complex data sets. Therefore it is wise to make a plan for the analyses beforehand to avoid getting side-tracked or to avoid the risk of "overfitting" (i.e., continuously adjusting the model to the data). There is a thin line between testing models and exploring for new ones. One easily enters the phase of explorations in which test statistics lose their original interpretation and outcomes require replication.

The building blocks of a model are its parameters and they basically consist of variances and covariances (i.e., correlations and regressions). When modeling a parameter, a researcher has three options. The first option is to fix a parameter at a certain value; for example, a covariance can be set at 0 when it is hypothesized that there is no covariance between two variables and the parameter does not need to be estimated, or a variance can be set at 1 when one wants to standardize a latent variable. If one wants to equate a latent variable's scale to that of a reference variable, the regression ("factor loading") of that particular observed variable on the latent variable can be set at 1 to achieve that. As a second option, the researcher can model a parameter to be "free" and the program will estimate the value of the parameter such that it fits the data best. This may be the case when, for example, it is assumed that there is a relationship between latent variables (e.g., Metacognitive knowledge, Lexical-semantic knowledge, and Morpho-syntactic knowledge in the earlier example), and we want an estimate of the size of the covariance. In such cases, the covariance parameter will be modeled as a free parameter. A third way in which a parameter can be modeled is to constrain it to be equal to another parameter. One can postulate that covariances, regressions, and/or variances are equal. These options for modeling parameters apply to the structural and measurement part of a model alike. For example, in a test development project a researcher could be interested in the question of whether tests A and B are parallel in a psychometric sense. This—among other things—means that the error variance in A and B and the regressions for A and B on the latent variable are equal to each other, respectively (cf. Bollen, 1989; see Schoonen, Vergeer, & Eiting, 1997 for an application).

Fitting and Evaluating a Model

Once a researcher has operationalized the hypotheses in a model, she or he can put this model to test by fitting it to the data. Essentially, on the basis of the model specifications, the SEM analysis reproduces or estimates a covariance matrix of observed variables that would accommodate the model specifications best, and this reproduced covariance matrix is compared to the actual covariance matrix of the input data. In this process, initial estimates or starting values for the free and constrained parameters are computed by the program. Based on the differences between the observed sample covariance matrix and the reproduced or estimated matrix, these initial values are adjusted in a second iteration to minimize the difference between the reproduced and observed covariance matrix. In successive iterations the program will estimate the optimal parameter values such that the difference between observed and reproduced matrix is minimal according to—for instance—a maximum likelihood (ML) function and further iterations do not lead to better fit. The program will stop its iterations and report the achieved results. Researchers have several options for the fit functions, such as ML and general or unweighted least squares (GLS and ULS, respectively). See Bollen (1989) for an extensive treatment of the different procedures. Software packages as LISREL and AMOS provide ML estimates by default. Different software packages for SEM may use slightly different procedures to compute starting values and algorithms to minimize fit functions, and therefore the same analysis on the same data set may sometimes lead to slightly different parameter estimates, but usually the general results will be the same. These packages are constantly updated to meet new requirements and insights. Ullman (2007) provides a comparison of a few packages at that point in time.

Once the SEM iterations have converged to a solution, the researcher will have to evaluate whether the model satisfactorily fits the data. This is not a simple yes/no matter, because there are multiple ways of evaluating the fit of a model. There is a statistical way and there are many descriptive ways. The analysis gives a chi-square (or related) statistic with a corresponding *p*-value and degrees of freedom (*df*). In conventional null hypothesis testing, researchers usually want to reject the null hypothesis (e.g., $p < .05$). However in SEM analyses, most of the time one does not want to reject the model. This raises the question of whether *p*-values simply greater than .05 suffice. This issue is further complicated by the fact that the chi-square in SEM analyses is sensitive not only to sample size, but also to the number of parameters that had to be estimated. Most researchers use the chi-square statistic as a more descriptive indicator of model fit than as a serious statistical significance test. A ratio of less than 2 for χ^2 / df is considered a good fit (Kline, 2010; Ullman, 2007). The degrees of freedom are derived from the number of observed variables in the input and the number of parameters estimated in the model, and as such they are also a good check on the model specification. One should be able to forecast the degrees of freedom for one's model in a SEM analysis. If the data set under investigation consists of *m* variables, the covariance matrix consists of $m\ (m + 1) / 2$ elements. From this number, the number of parameters has to be subtracted to get the degrees of freedom. Of course, two parameters set to be equal count as a single estimated parameter. Predicting the degrees of freedom of one's model before actually running the analysis is thus a check of the correct implementation of the model.

In addition to a chi-square value, a SEM analysis will provide the researcher with many more descriptive fit indices. Some are based on the differences (residuals) between the input covariance matrix and the reproduced covariance matrix (e.g., standardized root mean square residual, or SRMR). Other indices take the number of estimated parameters into account as well; the more parsimonious the model is (i.e., the fewer estimated parameters), the better (e.g., the root mean square error of approximation, or RMSEA). Others are based on a comparison between the fit of the tested model and a basic or "null" model that assumes the variables to be unrelated (e.g., the nonnormed fit index, or NNFI, also known as the Tucker-Lewis index, and the comparative fit index, or CFI). Different fit indices weight different aspects of the model (sample size, number of parameters, residuals, etc.) differently (see Kline, 2010). For most of these fit indices both lenient and strict cutoff criteria can be found in the literature (Hu & Bentler, 1999). As a rule of thumb, the *SRMR* should be lower than .08, the *RMSEA* lower than .06, and the *CFI* higher than .95 (Hu & Bentler, 1999). As with determining the number of factors in EFA or the number of clusters in a cluster analysis (see Loewen & Gonulal, Chapter 9 in this volume, and Staples & Biber, Chapter 11 in this volume), multiple fit indices should be taken into account to avoid overprioritizing one particular criterion.

A third (additional) evaluation of a model consists of the inspection of the model parameters themselves and the residuals. It could well be the case that, generally speaking, a complex model fits the data well, but that at the same time some "local" misfit exists. Therefore, a check of the residuals and of the meaningfulness of individual parameter estimates is advisable. Eyeballing the standardized residuals (i.e., the standardized differences between the observed covariances of the input variables and the reproduced covariances) may show outlying residuals that indicate local misspecifications. In a similar vein, parameter estimates that are illogical (such as a negative variance or a correlation out of the −1 to 1 range) could flag a local misfit as well.

Pitfalls

One of the risks of using SEM is that researchers endlessly tweak a model, helped by the so-called modification indices that indicate how the chi-square will change if a certain fixed parameter is set free (Lagrange Multiplier test) or if a free parameter is set fixed (Wald test). It is very tempting to attune a model according to these indices and in such a way to strive for more acceptable fit statistics. However, this is also a risky enterprise because researchers are often inclined to include relationships that are not theoretically supported, and after a number of modifications the significance testing can no longer be seen as real hypothesis testing and *p*-values become meaningless. The researcher might end up with a hybrid model that most likely will not be replicable. If analyses cannot be replicated, the study "might as well be sent to the *Journal of Irreproducible Results* or to its successor, *The Annals of Improbable Research*," according to Boomsma (2000, p. 464).

A more interesting and useful approach is to compare two competing models, preferably representing two stances in a theoretical debate. A comparison of the fit of the two models could point to the model and the theoretical stance that deserves our support. Consider, for example, the unitarian holistic view on language proficiency versus the componential view mentioned earlier. A SEM analysis of test scores could show that a multiple-factor model fits the data much better than a one-factor model, and that multiple latent variables (components) should be distinguished, favoring the componential view. Models that are hierarchically nested (i.e., the parameters of one model (A) form a subset of the parameters of the other (B)), can be compared statistically by the chi-square difference test. The difference in the two models' chi-squares is a new chi-square with as the degrees of freedom the difference in *df*s of the two compared models ($\Delta\chi^2 = \chi^2_A - \chi^2_B$; $\Delta df = df_A - df_B$).

In all cases, it is considered best practice to report the steps taken in the development of the ultimate model, which parameters were set to be fixed at a certain value, which ones were freely estimated, and which ones were constrained to be equal to another parameter (Mueller & Hancock, 2008). A model's replicability is one of the points that is stressed by Boomsma (2000), quoting Steiger's (1990) adage: "An ounce of replication is worth a ton of inferential statistics" (p. 176).

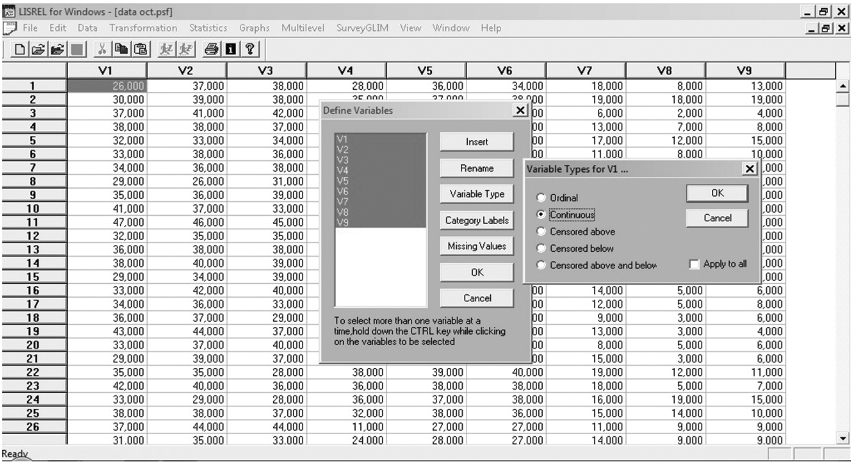
A SEM Analysis Step by Step

Later in this chapter a study (Gu, 2014) that uses SEM in various ways is briefly introduced and discussed (see "Text Box 6"). In this section we show the steps a researcher has to take to perform a SEM analysis. SEM is a rich toolbox with all kinds of options and possibilities, many more than can be illustrated in a single chapter or a single example. Readers who want more extensive introductions and examples are referred to "Tools and Resources" and "Further Readings." This introductory example will illustrate the use of LISREL and AMOS, respectively. The introduction to LISREL will refer to two modes of working with LISREL, i.e., using the SIMPLIS syntax and using the program menus. The introduction to AMOS will be brief to avoid overlap with the introduction to LISREL. Both packages can take different data file formats as input quite easily, for example raw data files and SPSS data files.

The example concerns data that allow us to test the models depicted in Figure 10.3. The data are fictitious: nine variables, *N*=341. If these data (for example as an SPSS file) are imported in LISREL (8.80 Student version), this will prompt PRELIS 2.80 for—among other things—data screening (e.g., evaluation of distributions, multivariate plots). The researcher will be prompted to save the data as a PRELIS data file (*.psf) that can be used for the SEM analyses. Note that imported data are by default considered to be ordinal; these can easily be changed into continuous by clicking **Data** > **Define Variables**, selecting the variables you want to change and then selecting *Continuous* and *OK* (see Figure 10.4).

Command lines for a SEM analysis in LISREL are straightforward. Researchers can use the matrix notation, the SIMPLIS language, and/or a graphical interface (Jöreskog & Sörbom, 1996–2001). For this example, the SIMPLIS language was used (see Text Box 1). A new "SIMPLIS project" file can be opened by selecting **File** > **New** from the top bar of the LISREL program. In this case we named the file SAMPLEDATA.spj (*.spj is the extension LISREL adds). In this new screen (see Figure 10.5), you can either key in your commands or paste them from a menu (similar to working in SPSS via the menu options versus working in a SPSS syntax file). The options under **Setup** (in the top bar) can be helpful in building a setup for the analyses.

After entering the Title, and in this case ignoring the definition of groups (since our data pertain to a single group of participants), we can read the data for the analysis by clicking *Add/Read Variables* in the *Variables* menu, selecting *PRELIS System File* from the drop-down menu, and then browsing to the path where we have saved the *PRELIS System File* (SAMPLEDATA.psf). Click *OK*, and the nine variables (V1 to V9) and a constant are available for model specification. In the right-hand panel we can add the latent variables that we assume to underlie our measured variables. In our first model we hypothesize one general factor: L2 proficiency (L2Prof). Entering this label (see Figure 10.6) and a clicking on *OK* will bring us back to the setup screen. Select *Setup* again and click *Build SIMPLIS Syntax*, and the setup so far appears in the upper panel. This setup can


LISREL for Windows - [data oct.psf]
File Edit Data Transformation Statistics Graphs Multilevel SurveyGLIM View Window Help
Define Variables
V1
V2
V3
V4
V5
V6
V7
V8
V9
Insert
Rename
Variable Type
Category Labels
Missing Values
OK
Cancel
Variable Types for V1 ...
OK
Cancel
Ready

	V1	V2	V3	V4	V5	V6	V7	V8	V9
1	26,000	37,000	38,000	28,000	36,000	34,000	18,000	8,000	13,000
2	30,000	39,000	38,000	35,000	37,000	38,000	19,000	18,000	19,000
3	37,000	41,000	42,000			00	6,000	2,000	4,000
4	38,000	38,000	37,000			00	13,000	7,000	8,000
5	32,000	33,000	34,000			00	17,000	12,000	15,000
6	33,000	38,000	36,000			00	11,000	8,000	10,000
7	34,000	36,000	38,000			00			,000
8	29,000	26,000	31,000			00			,000
9	35,000	36,000	39,000			00			,000
10	41,000	37,000	33,000			00			,000
11	47,000	46,000	45,000			00			,000
12	32,000	35,000	35,000			00			,000
13	36,000	38,000	38,000			00			,000
14	38,000	40,000	39,000			00			,000
15	29,000	34,000	39,000			00			,000
16	33,000	42,000	40,000			00	14,000	5,000	6,000
17	34,000	36,000	33,000			00	12,000	5,000	8,000
18	36,000	37,000	29,000			00	9,000	3,000	6,000
19	43,000	44,000	37,000			00	13,000	3,000	4,000
20	33,000	37,000	40,000			00	8,000	5,000	6,000
21	29,000	39,000	37,000			00	15,000	3,000	6,000
22	35,000	35,000	28,000	38,000	39,000	40,000	19,000	12,000	11,000
23	42,000	40,000	36,000	36,000	38,000	38,000	18,000	5,000	7,000
24	33,000	29,000	28,000	36,000	37,000	38,000	16,000	19,000	15,000
25	38,000	38,000	37,000	32,000	38,000	36,000	15,000	14,000	10,000
26	37,000	44,000	44,000	11,000	27,000	27,000	11,000	9,000	9,000
	31,000	35,000	33,000	24,000	28,000	27,000	14,000	9,000	9,000

To select more than one variable at a time,hold down the CTRL key while clicking on the variables to be selected

○ Ordinal
● Continuous
○ Censored above
○ Censored below
○ Censored above and below
□ Apply to all

FIGURE 10.4 PRELIS data definition options

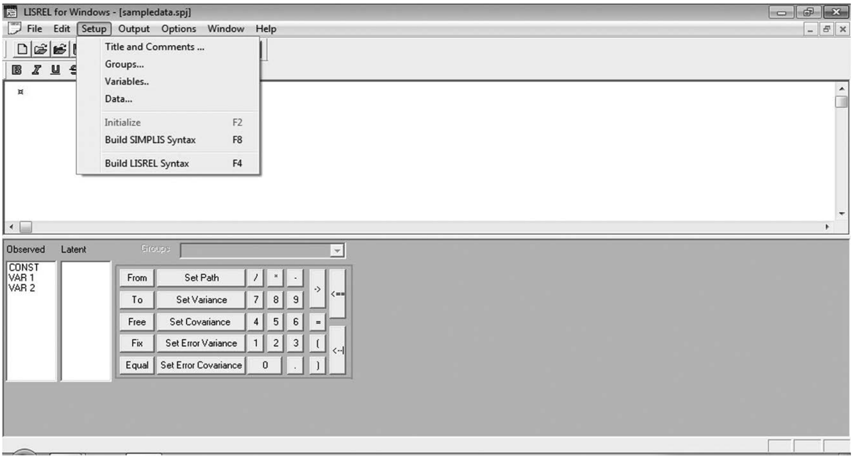
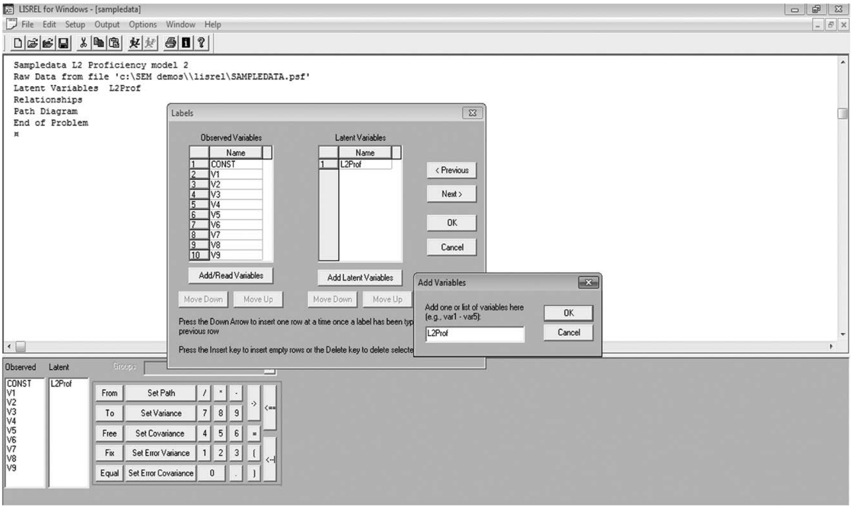


FIGURE 10.5 Starting to build command lines

be extended either by typing additional model specifications (such as the ones in Text Box 1), or by selecting keywords from the lower panel and dragging variable names from the lower panel to the upper panel. Clicking *Build SIMPLIS Syntax* again will check and add default information, such as the variance of latent variables. Since the latent variables have no predefined scale it is assumed that they have a variance of 1. In Text Box 1, the actual data are entered as a covariance matrix as an alternative way of importing data, which might be convenient if data are not available as raw data, but are—for instance—derived from published work. Sample size and the names of the observed (measured) variables need to be mentioned explicitly where they are implied when one uses the PRELIS system file.



LISREL for Windows - [sampledata]
File Edit Setup Output Options Window Help

```
Sampledata L2 Proficiency model 2
Raw Data from file 'c:\SEM demos\\lisrel\SAMPLEDATA.psf'
Latent Variables  L2Prof
Relationships
Path Diagram
End of Problem
```

Press the Down Arrow to insert one row at a time once a label has been typ previous row

Press the Insert key to insert empty rows or the Delete key to delete selecte

Add one or list of variables here (e.g., var1 - var5):

L2Prof

FIGURE 10.6 Adding latent variable command lines

TEXT BOX 1: COMMAND LINES FOR LISREL ANALYSIS (ONE-FACTOR MODEL)

```
Ti L2 Proficiency
Observed variables
V1 -V9
covariance matrix
42.039
32.026 41.285
30.452 33.114 53.178
14.603 9.254 11.165 56.655
12.818 8.959 8.700 41.741 41.969
11.251 7.340 10.574 34.607 29.376 33.313
6.825 5.476 4.992 21.428 21.028 14.994 12.726
21.101 18.174 21.764 23.986 21.238 18.994 11.174 32.831
19.748 17.708 22.026 20.805 17.889 17.877 9.591 28.147 29.678
SAMPLE SIZE is 341
Latent variables
L2Proficiency
Relationships
V1-V9 =L2Proficiency
Path diagram
End of problem
```

In the command lines, the equals sign (=) can be read as "is determined by." The pre-final line in Text Box 1 will result in a path diagram that depicts the hypothesized model and as such provides a nice check on the specification of the model. By default the program will provide ML estimates. However, data requirements such as multivariate normality need to be met to get trustworthy estimates (Kline, 2010). The estimation procedure can be changed from ML to, for example, GLS by adding an extra SIMPLIS command line: Method of Estimation: General Least Squares just above or under Path diagram in Text Box 1, or by selecting **Output** > **Simplis outputs**. This leads us to options for the method of estimation and other output features. Of course, there are many more options for analyses and kinds of output LISREL can produce than can be demonstrated here (see Jöreskog & Sörbom, 1996–2001 for more detailed descriptions).

The analysis is run by clicking the *Run LISREL* button in the top bar. If there are no serious misspecifications or syntactical errors, the model will show the path diagram with the estimates. One can switch to the output file with all the details by means of the *Window* button. The LISREL output file that results from the analysis echoes the command lines and the covariance matrix for reference. The most important part of the outcomes consists of the parameter estimates with their standard errors and the indices for model fit. In this example, fit indices as reported in Text Box 2 indicate that the model should be rejected and does not fit the data very well. None of the aforementioned fit indices that are reported for the one-factor model comes close to the recommended cutoff for good fit.

TEXT BOX 2: EDITED PART OF THE LISREL OUTPUT (ONE-FACTOR MODEL)

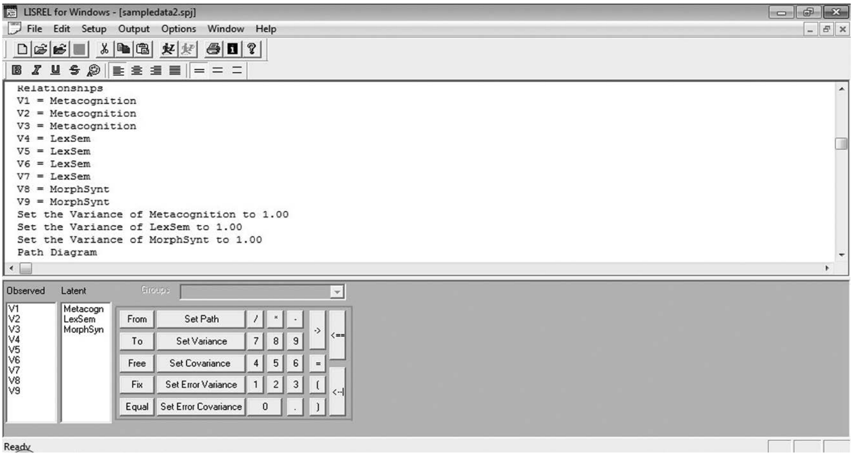
```
                 Goodness of Fit Statistics

                  Degrees of Freedom = 27
      Minimum Fit Function Chi-Square = 1177.53 (P = 0.0)
                          (. . .)
    Root Mean Square Error of Approximation (RMSEA) = 0.38
   90 Percent Confidence Interval for RMSEA = (0.36 ; 0.40)
      P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00
                          (. . .)
     Chi-Square for Independence Model with 36 Degrees of
                     Freedom = 3764.64
                          (. . .)
                Normed Fit Index (NFI) = 0.69
              Non-Normed Fit Index (NNFI) = 0.59
           Parsimony Normed Fit Index (PNFI) = 0.52
              Comparative Fit Index (CFI) = 0.69
              Incremental Fit Index (IFI) = 0.69
                          (. . .)
```

```
             Root Mean Square Residual (RMR) = 9.00
                    Standardized RMR = 0.22
                           (. . .)
```

In a similar way one can build a three-factor model; that is, one has to replace the last six lines of the setup as represented in Text Box 1 and introduce three latent variables (instead of one): Metacognition, Lexical-Semantic, and Morpho-Syntactic Knowledge (see Text Box 3). Working with the LISREL menu, one can add and rename labels for latent variables via **Setup** > **Variables** as illustrated earlier, and then redesign the model accordingly in the upper panel (see Figure 10.7). This model specification can be fitted to the data by clicking the *Run LISREL* button in the top bar. The results show that a three-factor model is far more realistic and that it fits the data much better, although still not very well yet. The fit indices (see Text Box 4) come close to the required level for good fit. Statistically speaking, the model has to be rejected ($\chi^2 = 120.28$, $df = 24$), but it constitutes an enormous improvement compared to the first model ($\chi^2 = 1{,}177.53$, $df = 27$). At the "cost" of three extra estimated parameters (these are the covariances between the latent variables), the reduction in chi-square is remarkable and statistically significant ($\Delta\chi^2 = 1{,}057.25$, $\Delta df = 3$, $p < .001$), which means that the less restrictive three-factor model is preferred. The RMSEA, which reduced from .38 to .12, however, indicates that the model fit is still not satisfactory. The normed fit index (NFI) and the CFI both show a noticeable increase (from .69 to .96) and both are satisfactory. The SRMR dropped from .22 to .041, which is in the range of acceptable models.

TEXT BOX 3: COMMAND LINES FOR LISREL ANALYSIS (THREE-FACTOR MODEL)

```
Ti L2 Proficiency (3 factors)
 (. . .)
Latent variables
Metacognition LexSem MorphSynt
Relationships
V1-V3 = Metacognition
V4-V7 = LexSem
V8-V9 = MorphSynt
Path diagram
End of problem
```



```
Relationships
V1 = Metacognition
V2 = Metacognition
V3 = Metacognition
V4 = LexSem
V5 = LexSem
V6 = LexSem
V7 = LexSem
V8 = MorphSynt
V9 = MorphSynt
Set the Variance of Metacognition to 1.00
Set the Variance of LexSem to 1.00
Set the Variance of MorphSynt to 1.00
Path Diagram
```

FIGURE 10.7 Setup for the three-factor model

TEXT BOX 4: EDITED PART OF THE LISREL OUTPUT (THREE-FACTOR MODEL)

```
                 Goodness of Fit Statistics

                  Degrees of Freedom = 24
    Minimum Fit Function Chi-Square = 120.28 (P = 0.00)
                        ( . . . )
  Root Mean Square Error of Approximation (RMSEA) = 0.12
90 Percent Confidence Interval for RMSEA = (0.097 ; 0.14)
   P-Value for Test of Close Fit (RMSEA < 0.05) = 0.00
                        ( . . . )
  Chi-Square for Independence Model with 36 Degrees of
                   Freedom = 3764.64
                        ( . . . )
               Normed Fit Index (NFI) = 0.97
            Non-Normed Fit Index (NNFI) = 0.96
        Parsimony Normed Fit Index (PNFI) = 0.65
            Comparative Fit Index (CFI) = 0.97
            Incremental Fit Index (IFI) = 0.97
                        ( . . . )
        Root Mean Square Residual (RMR) = 1.56
                  Standardized RMR = 0.041
                        ( . . . )
```

The comparison of a one-factor model to a three-factor model as a research goal could have been theoretically underpinned. The three-factor model seems to be the better one, but is not yet completely satisfactory. It will depend on the specific research context whether the researcher can defend additional theoretically supported model improvements, or whether he or she enters the phase of explorations.

For the sake of demonstration, let us assume that all but one test score is derived from separate test administrations. The exception pertains to V5 and V7, which are subtest scores derived from one and the same test. As a consequence, disturbances during that test will affect both scores. In other words, there might be so-called correlated error. This phenomenon can be modeled by allowing covariance between the two residuals concerned (e_5 and e_7); in other words, add the line Let error covariance between V5 and V7 be free in the model specification. A final analysis shows that this extra free parameter in the model substantially improves fit ($\chi^2 = 71.16$, $df = 23$, RMSEA = .08, NFI = .98, CFI = .99, SRMR=.034). Not all indices are completely satisfactory for this model ($\chi^2 / df > 2$, RMSEA = .08) but if there are no more plausible parameters to add, the researcher might want to stop here and inspect the parameter estimates. When the parameter estimates are logical and within the normal ranges (for example, no negative estimates of variance), then the researcher can start the substantive interpretation. In this simple model it is important that the nine observed variables are explained to a large extent by the three presumed latent variables. The coefficients of determination (R^2) range from .62 to .93. which is reasonably good (see Text Box 5). From a theoretical point of view the correlations between the latent variables are interesting: How high are they? Are they different from 0 and—at the other end—sufficiently different from 1? In this case, LISREL reports .31 (.05), .65 (.04) and .63 (.04) with the corresponding standard errors between brackets for CIs and/or significance testing. When one takes the standard errors into account it can be concluded that the estimates are (statistically) different from 0 and 1. In this example, the focus was on the underlying latent variables of the nine observed variables. In a next step or dealing with different research questions, one could investigate whether claims about "causal" relations between the three latent variables of the kind illustrated in Figure 10.2 can be maintained. One may want to test whether metacognitive knowledge is the result of lexical-semantic and morphosyntactic knowledge. To address that question the regression of Metacognitive Knowledge on the Lexical-Semantic and the Morphosyntactic factors should be specified.

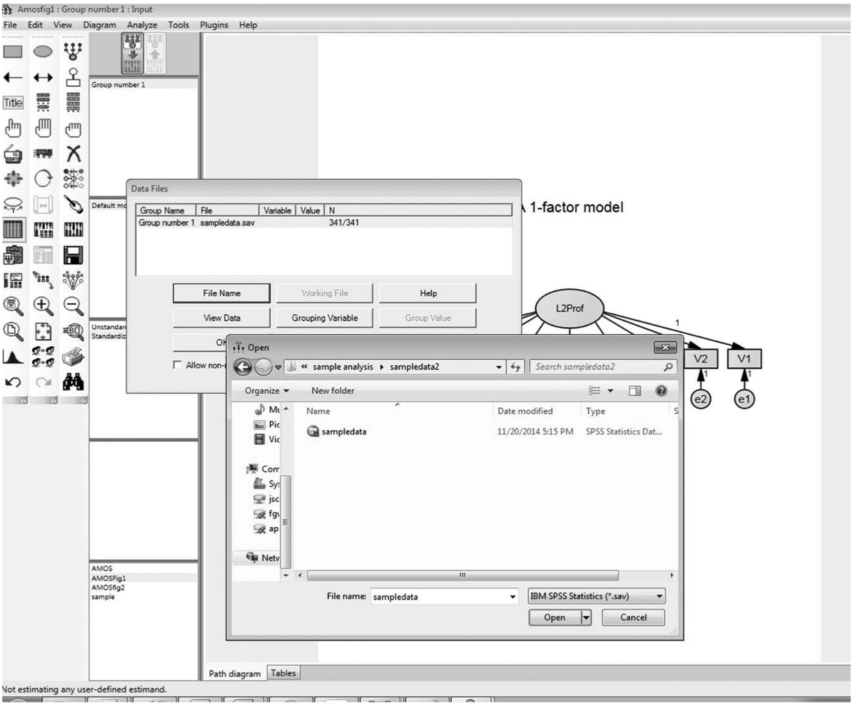
TEXT BOX 5: EDITED PART OF THE LISREL OUTPUT (THREE-FACTOR MODEL WITH CORRELATED ERROR)

```
LISREL Estimates (Maximum Likelihood)
Measurement Equations

    V1 = 5.57*Metacogn, Errorvar.= 11.07, R² = 0.74
        (0.29)                    (1.33)
        18.92                      8.34
    V2 = 5.69*Metacogn, Errorvar.= 8.95, R² = 0.78
        (0.29)                    (1.25)
        19.81                      7.14

    V3 = 5.74*Metacogn, Errorvar.= 20.26, R² = 0.62
        (0.34)                     (1.94)
        16.69                      10.46

    V4 = 6.97*LexSem, Errorvar.= 8.06, R² = 0.86
        (0.32)                  (1.12)
        22.08                    7.19

    V5 = 5.96*LexSem, Errorvar.= 6.50, R² = 0.85
        (0.27)                  (0.85)
        21.80                    7.65

    V6 = 4.97*LexSem, Errorvar.= 8.56, R² = 0.74
        (0.25)                  (0.83)
        19.65                   10.38

    V7 = 3.06*LexSem, Errorvar.= 3.38, R² = 0.73
        (0.16)                  (0.35)
        19.30                    9.79

    V8 = 5.53*MorphSyn, Errorvar.= 2.24, R² = 0.93
        (0.24)                    (0.70)
        23.52                      3.22

    V9 = 5.09*MorphSyn, Errorvar.= 3.78, R² = 0.87
        (0.23)                    (0.64)
        22.24                      5.90

Error Covariance for V7 and V5 = 2.82
                                (0.47)
                                 5.95

Correlation Matrix of Independent Variables

                  Metacogn   LexSem   MorphSyn
                  --------   ------   --------
Metacogn              1.00
  LexSem              0.31     1.00
                     (0.05)
                      5.64
MorphSyn              0.65     0.63        1.00
                     (0.04)   (0.04)
                     17.80    17.73
```

```
                    Goodness of Fit Statistics

                      Degrees of Freedom = 23
        Minimum Fit Function Chi-Square = 71.16 (P = 0.00)
                              (. . .)
   Root Mean Square Error of Approximation (RMSEA) = 0.080
   90 Percent Confidence Interval for RMSEA = (0.060 ; 0.10)
    P-Value for Test of Close Fit (RMSEA < 0.05) = 0.0079
                              (. . .)
    Chi-Square for Independence Model with 36 Degrees of
                         Freedom = 3764.63
                              (. . .)
                  Normed Fit Index (NFI) = 0.98
                Non-Normed Fit Index (NNFI) = 0.98
             Parsimony Normed Fit Index (PNFI) = 0.63
                Comparative Fit Index (CFI) = 0.99
                Incremental Fit Index (IFI) = 0.99
                              (. . .)
              Root Mean Square Residual (RMR) = 1.32
                    Standardized RMR = 0.034
                              (. . .)
```

The same analyses can be done in AMOS by drawing the required model with the tools provided in the program. The opening screen of AMOS (Graphics) consists of three parts, with the left-most panel showing a toolbox for model drawing. Holding the cursor on an icon in the toolbox will show its function. From this panel one can select the tools needed for drawing the model: circles and boxes for latent and measured variables, respectively; single- and double-headed arrows, but also a tool to add measured variables to a latent variable ; and an eraser to delete parts of a model. Once a model is designed, one can import the data by clicking , *Filename* and then browsing the computer for the right data file (see Figure 10.8)—by default an *SPSS* file, but other formats can be read as well. All variables in the model need to be named, and the measured variables in the model need to be linked to variables in the data file. Double-clicking circles lets you key in names for latent variables. Note that the "errors" need to be named as well, for example E1 through E9, because they are treated as latent variables in AMOS. Desired features of the analysis or outcomes, such as ML and standardized parameters, can be handled in the Analysis Properties menu, which you can access by clicking this button . If the model is fully designed, the data and variables are included, and the features for the analysis are set, the *Calculate* button can be clicked. The two top buttons in the middle panel now allow the researcher to toggle between a representation of the model as designed (i.e., input) and a representation of the model with parameters (i.e., output). However, the details of the analysis such as fit indices, standard errors, and possible warnings are provided in text. Clicking *View Text* ( ) provides access to the text file with a table of contents at the left (*navigation tree*) and at the right the corresponding results. Figure 10.9 shows the fit indices for our model with three factors and correlated error. The chi-square was identical to that of the LISREL analysis (71.16), as are the fit indices. In AMOS fit indices are reported next to the fit of an independence model and a saturated model. The model of interest is the Default model, which is labeled this way because we did not enter a name for it.

This is a very superficial introduction of the possibilities of AMOS and LISREL. Readers who wish to embark on SEM sessions will best familiarize themselves with the software manual, which is usually embedded in the package in the Help area, or consult more extensive introductions aiming at a certain packages (see Byrne, 1998, 2010).



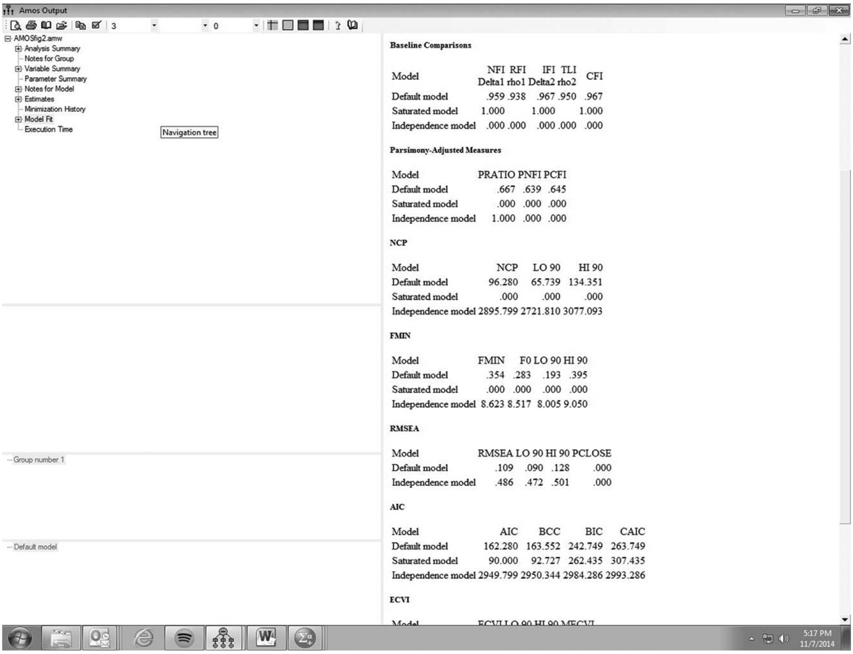
Group Name	File	Variable	Value	N
Group number 1	sampledata.sav			341/341

Name	Date modified	Type
sampledata	11/20/2014 5:15 PM	SPSS Statistics Dat...

File name: sampledata

IBM SPSS Statistics (*.sav)

FIGURE 10.8 Importing data for one-factor model in AMOS



```
Baseline Comparisons

                    NFI   RFI    IFI   TLI
Model                                        CFI
                  Delta1 rho1 Delta2 rho2
Default model        .959 .938   .967 .950   .967
Saturated model     1.000       1.000       1.000
Independence model   .000 .000   .000 .000   .000

Parsimony-Adjusted Measures

Model             PRATIO PNFI PCFI
Default model        .667  .639  .645
Saturated model      .000  .000  .000
Independence model  1.000  .000  .000

NCP

Model                  NCP    LO 90    HI 90
Default model        96.280   65.739  134.351
Saturated model        .000     .000     .000
Independence model 2895.799 2721.810 3077.093

FMIN

Model              FMIN    F0 LO 90 HI 90
Default model       .354  .283  .193  .395
Saturated model     .000  .000  .000  .000
Independence model 8.623 8.517 8.005 9.050

RMSEA

Model              RMSEA LO 90 HI 90 PCLOSE
Default model        .109  .090  .128   .000
Independence model   .486  .472  .501   .000

AIC

Model                   AIC      BCC      BIC     CAIC
Default model        162.280  163.552  242.749  263.749
Saturated model       90.000   92.727  262.435  307.435
Independence model 2949.799 2950.344 2984.286 2993.286

ECVI

Model              ECVI LO 90 HI 90 MECVI
```

FIGURE 10.9 Output file three-factor model with correlated error in AMOS

In Text Box 6 we briefly present parts of a recent study that uses SEM in various ways. Here we focus on the underlying structure of the TOEFL iBT that Gu investigated as part of her doctoral dissertation. In the dissertation and the article (Gu, 2014), a multigroup analysis was conducted to investigate whether the underlying structure holds for two different groups, and whether level of performance was related to studying abroad.

TEXT BOX 6: A SAMPLE STUDY

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing, 31*(1), 111–133.

This study addresses three research questions. For reasons of space, we do not go into the analyses for questions 2 and 3.

Background

Gu (2014) investigated the structure of scores on the Internet-based Test of English as Foreign Language (TOEFL-iBT). This study combines several applications of SEM: a factor-analytic application to investigate the underlying structure of TOEFL-iBT; a multisample analysis to evaluate whether the underlying structure that was found holds for two subpopulations of test takers; and an investigation of the so-called mean structure to compare score differences on the latent variables instead of on the observed scores. It is beyond the scope of this chapter to discuss all these uses of SEM, but Gu's study nicely shows the flexibility of SEM.

Research Questions

1) Is the factorial structure of academic language ability the same for students who have studied abroad and students who have not done so (a study-abroad group versus a home-country group)?
2) Do the two groups differ in their scores on the underlying factors (i.e., latent variables) of academic English?
3) Is there a relationship between length of study abroad and the level on the underlying factors?

Here we focus on Research Question 1.

Method

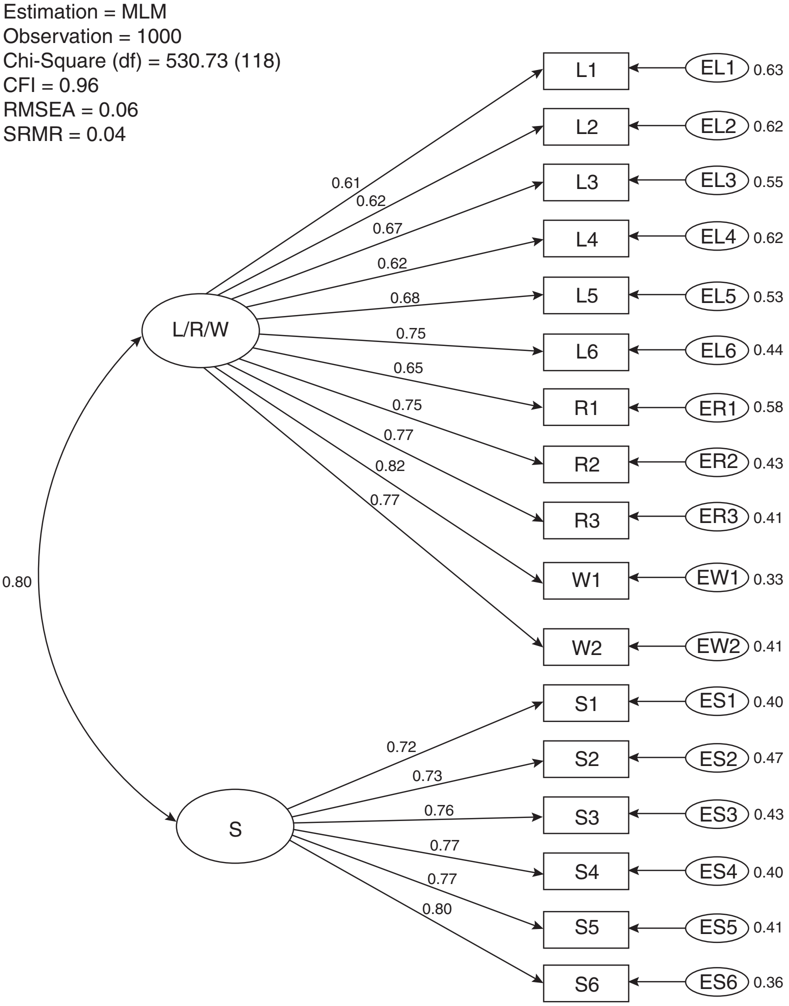
The data consisted of the test scores and questionnaire responses of 1,000 and 370 test takers, respectively. The subsample that answered the questionnaire was split in two groups: (a) never lived in an English speaking environment (*n*=124) and (b) have lived in such an environment (*n*=246). Data for the present analysis were based on test scores of 1,000 candidates for listening, reading, writing, and speaking. From the questionnaire data, Gu derived information about exposure to English language and instruction.

Using the Mplus SEM package (Muthén & Muthén, 2010), Gu explicates the check of relevant assumptions such as normality. Since some score distributions deviated from normality, Gu opted for an adjusted estimation of chi-square, derived indices, and standard errors of parameters (the Satorra-Bentler correction). The scale for each latent variable is determined by using a reference variable and fixating its loadings on the latent variable relative to 1.

Results

Gu postulated three plausible models for the structure of the four skills. The fit of these models and the comparison thereof was used to choose the best model. Model 1 follows the scoring procedure of TOEFL-iBT and previous research. It consists of four factors representing the four skills and one higher-order, overarching factor ("Language Ability") that is supposed to capture the correlations between the four skills. Model 2 is a straightforward four-factor model with intercorrelated factors, one for each skill. Model 3 consists of two factors: "Speaking" on the one hand and "Reading, Writing, and Listening" on the other. This latter model is based on previous research, but is theoretically speaking less transparent (see Gu's Figure 4, reproduced below).

Model fit was evaluated in several ways, as it should be: overall fit (chi-square test, CFI, RMSEA, SRMR), evaluation of parameter estimates, and parsimony for equally well fitting models. The SEM analyses showed reasonable fit for all three models, Model 3 being somewhat less well fitting.



GU'S FIGURE 4: Correlated two-factor model with standardized estimates (Gu, 2014, p. 123) in *Language Testing, 31*(1), 111–133, copyright © 2012 by author. Reprinted by permission of Sage.

Evaluating the parameters, Gu discovered that in Model 2 Listening and Writing showed extremely strong correlation (.97) and that in Model 1 the factor loadings for Listening and Writing on the higher-order general factor were exceptionally high. Both observations indicate that Listening and Writing ability are difficult to distinguish empirically. This prompted Gu to opt for Model 3 with two factors: Speaking with six indicators, and non-Speaking (LWR) with the remaining 11 indicators. The standardized factor loadings were represented in the visual representation of the model (see Figure 4). The selected model was then tested successfully on the subsample of 370 test takers as well.

Readers interested in the solutions to research questions 2 and 3 are referred to Gu (2014). The article includes the descriptives and the correlation matrix of the seventeen variables (*N*=1,000). Furthermore, Gu's study shows a few more (common) applications of SEM in a clear and well-reported way.

In Sum

SEM is a flexible approach to data analysis, especially for larger data sets that represent more complex relationships. The possibilities to apply SEM are enormous, but the substantive interpretation of models and parameter estimates depends heavily on carefully conducted analyses, taking into account data requirements and the risk of overfitting the model.

Tools and Resources

SEM researchers have different software packages at their disposal. Each possess a unique set of strengths and weaknesses in use and in the way they can deal with special cases (Ullman, 2007). Most commercial packages also have demo versions for a limited number of variables and/or participants and/or for a limited period of time that allow the new user to explore the possibilities of the software.

- *R* (Fox, 2006): a freeware statistical package. Rosseel (2012) has developed a special package for R users called *lavaan* (latent variable analysis).
- *AMOS* (Arbuckle, 2012): This package is related to *SPSS* and has a graphical interface (see also Byrne, 2010).
- *LISREL*: This package was originally developed by one of the founding fathers of SEM, Karl Jöreskog. It started with a matrix-oriented interface, but now has several interfaces including a visual one and the *SIMPLIS* language (see Jöreskog & Sörbom, 1996–2001)
- Other more or less specialized packages include *Mplus* (Muthén & Muthén, 2010) and *EQS* (Bentler, 2006).

A number of additional online resources and communities can also provide assistance:

- SEMNET (The Structural Equation Modeling Discussion Network): A listserv and discussion board for all things SEM: http://www2.gsu.edu/~mkteer/semnet.html
- The website of David A. Kenny: http://davidakenny.net/cm/causalm.htm
- A thorough set of video lectures on SEM: http://www.ats.ucla.edu/stat/seminars/muthen_08/default.htm

Further Reading

There are many different introductions and advanced volumes dealing with SEM. A good starting point could be the manual of the software package that one wants to use. The manual can provide a quick introduction into the theoretical considerations, many of which are only touched upon here. Byrne (1998, 2006, 2010, 2012) wrote different introductions for different software packages (*Mplus, LISREL*, *EQS*, *AMOS*). More general introductions include Raykov and Marcoulides (2006), Kline (2010), Mueller & Hancock (2008) and Ullman (2007). These volumes also cover some of the more advanced applications, such as multigroup analysis in which models are fitted simultaneously in two (or more) groups (for example, boys and girls, L1 and L2 speakers, or study-abroad and study-home as in Gu's study), or latent growth modeling in which different curves of development can be modeled and related to predictor variables. Hancock and Mueller (2013) provide in their edited volume what they call a "second course," that is, the contributions take the applications a step further and deal with topics like missing data, categorical data, power analysis, and so forth.

There is also a journal dedicated to structural equation modeling that publishes applications from all fields, discusses methodological issues, and has a "teacher's corner" that presents brief instructional articles on SEM-related issues*:* *Structural Equation Modeling: A Multidisciplinary Journal* (ISSN 1070–5511 [Print], 1532–8007 [Online]).

There are also a number of introductions and applications in the field of applied linguistics and language assessment; see Hancock and Schoonen (2015), Kunnan (1998), Schoonen (2005), In'nami and Koizumi (2011, 2012), and Ockey (2014).

Discussion Questions

1. Select a study that uses SEM and read the abstract, introduction, and research questions. On the basis of your reading, draw the model you expect the researchers to test. In what respect does your model diverge from the model actually tested? To what extent can you understand the differences between your model and the author's? Are there any unexpected differences and are these motivated (a priori or post hoc)? How logical are the unexpected differences?

2. Select a study that uses SEM and that postulates correlated error. Are these parameters well explained in terms of the measurement procedures?
3. Select two SEM studies. What criteria do they use for model fit? Do they use criteria from different families of fit indices? Are there any other differences between the two studies? If you would apply the criteria from one study to the other, would that affect the model selection (and conclusions) in the other study? How so?
4. It is claimed that the correlations between latent variables are not attenuated by measurement error. Can you corroborate that on the basis of the data in Text Box 1? What is the average correlation between the observed variables for Metacognitive Knowledge (V1–V3) and observed variables for Morphosyntactic Knowledge (V8–V9)? How does that compare to the .65 reported for the correlation between the latent variables?
5. Using the data set made available along with this chapter (http://oak.ucc.nau.edu/ldp3/AQMSLR.html), explore whether another structural model for the three latent variables in the sample analysis is plausible (e.g., Metacognitive Knowledge as the result of the two latent linguistic variables). How plausible is a model with Metacognitive Knowledge independent of the two latent linguistic variables? Try to model these "hypotheses" and fit the models to the data.
6. How could you test whether the two latent linguistic variables coincide? In other words, test a two-factor model with a metacognitive factor (V1–V3) and a linguistic factor (V4–V9). How does this model compare to the one-factor model? To the three-factor model?
7. SEM and factor analysis have a lot in common. What similarities and differences between the two approaches can you think of? When would one approach be more appropriate or informative than the other?
8. Gu (2014) provides the correlation matrix of the measured variables involved in the models, as well as descriptives statistics. By doing so, the author allows you to replicate her analysis (consult the AMOS manual for importing a matrix). You can start a LISREL analysis with the setup provided in Text Box 1, and then continue by adjusting it. Choose your own title, define the observed variables (L1–W2), insert "correlation matrix" and replace the matrix with Gu's matrix, change sample size, define your latent variables and specify the relations (see also Text Box 3). As you probably know, correlations are standardized covariances, and the standardization is based on the standard deviations of the two variables involved (see Kline, 2010). LISREL can derive the covariances from the correlations on the basis of the standard deviations. So add another command, just above or under the correlation part, that starts with "Standard deviations" and then on the next line list all the standard deviations. Now replicate models 2 and 3 from Gu's study (i.e., the correlated four- and two-factor models).[1] What do you find? There will be small differences due to slightly different algorithms, but the overall outcome should be highly similar. The difference in chi-square is also due to a correction Gu applied to account for the slightly nonnormal data she had. It is beyond the scope of this chapter to go into the details.

Note

1. If you work with LISREL's student version, then you are restricted to 16 observed variables where Gu (2014) has 17.You could either delete the first variable L1 for Listening, or resort to the 15-day trial version of LISREL. If you delete L1, your results will of course differ, as well as the degrees of freedom. Can you predict *df*?

Acknowledgment

The author wishes to thank Jan Hulstijn, Camille Welie, Luke Plonsky, and two anonymous reviewers for their helpful comments. All remaining errors are the author's.

References

Arbuckle, J.L. (2012). *IBM® SPSS® AMOS™ 21 User's Guide*. Chicago: IBM Software Group.

Bentler, P.M. (2006). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal,* 7(3), 461–483.

Byrne, B.M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Byrne, B.M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Byrne, B.M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed). New York: Taylor & Francis.

Byrne, B.M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Taylor & Francis.

Enders, C.K. (2013). Analyzing structural equation models with missing data. In G.R. Hancock & R.O. Mueller (Eds.), *Structural equation modeling. A second course* (2nd ed., pp. 493–519). Charlotte, NC: Information Age Publishing.

Finney, S.J. & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G.R. Hancock & R.O. Mueller (Eds.), *Structural equation modeling. A second course* (2nd ed., pp. 439–492). Charlotte, NC: Information Age Publishing.

Fox, J. (2006). Structural equation modeling with the sem package in R. *Structural Equation Modeling, 13*(3), 465–486.

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing, 31*(1), 111–133.

Hancock, G.R. & Mueller, R.O. (Eds.) (2013). *Structural equation modeling. A second course* (2nd ed.). Charlotte, NC: Information Age Publishing.

Hancock, G.R., & Schoonen, R. (2015). Structural equation modeling: Possibilities for language learning researchers. *Language Learning, 65*: Suppl 1, 158–182.

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

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly, 8*(3), 250–276.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing, 29*(1), 131–152.

Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (1996–2001). *LISREL 8: User's Reference Guide* (2nd ed.). Lincolnwood, IL: Scientific Software International.

Kline, R.B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: The Guilford Press.

Kunnan, A.J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing, 15*(3), 295–332.

Mueller, R.O. & Hancock, G.R. (2008). Best practices in structural equation modeling. In J. Osborne (Ed.). *Best practices in quantitative methods* (pp. 488–508). Thousand Oaks, CA: Sage.

Muthén, L.K., & Muthén, B.O. (2010). *Mplus user's guide. Statistical analysis with latent variables* (6th ed.). Los Angeles: Muthén & Muthén.

Ockey, G.J. (2014). Exploratory factor analysis and structural equation modeling. In A.J. Kunnan (Ed.), *The companion to language assessment. Vol. III: Evaluation, Methodology, and Interdisciplinary Themes* (pp. 1224–1244, Part 10, Chapter 73). Malden, MA: John Wiley & Sons.

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal, 98*, 450–470.

Raykov, T., & Marcoulides, G.A. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah, NJ: Erlbaum.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Rovine, M.J., & Molenaar, P.C.M. (2003). Estimating analysis of variance models as structural equation models. In B.H. Pugesek, A. Tomer, & A. von Eye (Eds.), *Structural equation modeling: Applications in ecological and evolutionary biology* (pp. 235–280). Cambridge: Cambridge University Press.

Schoonen, R. (2005). Generalizability of writing scores. An application of structural equation modeling. *Language Testing, 22* (1), 1–30.

Schoonen, R., Van Gelderen, A., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: the role of linguistic fluency, linguistic knowledge and metacognitive knowledge. *Language Learning, 53*(1), 165–202.

Schoonen, R., Van Gelderen, A., Stoel, R., Hulstijn, J., & De Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary-school students. *Language Learning, 61*, 31–79.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14*(2), 157–184.

Steiger, J.H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*(2), 173–180.

Tseng, W.-T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning, 58*(2), 357–400.

Ullman, J.B. (2006). Structural equation modeling: Reviewing the basics and moving forward. *Journal of Personality Assessment, 87*(1), 35–50.

Ullman, J.B. (2007). Structural equation modeling. In B.G. Tabachnick & L.S. Fidell (Eds.), *Using multivariate statistics* (5th ed., pp. 676–780). Boston: Pearson/Allyn and Bacon.

West, S.G., Finch, J.F., & Curran, P.J. (1995). Structural equation models with nonnormal variables. Problems and remedies. In R.H. Hoyle (Ed.), *Structural equation modeling. Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.