An introduction to structural equation modelling for language assessment research

Antony John Kunnan *California State University, Los Angeles*

This article provides an introduction to structural equation modelling (SEM) for language assessment research in five sections. In Section I, the general objectives of SEM applications relevant to language assessment are presented. In Section II, a brief overview of SEM that considers the methodology and the statistical assumptions about data that have to be met. In Section III, the commonly used steps and concepts in SEM are presented. In Section IV, SEM application matters with example models are discussed. In Section V, recent critical discussions and some directions for future SEM applications in language assessment research are addressed.

I SEM and language assessment research

1 What are the general objectives?

The general objectives of SEM application in language assessment research have been:

1) research on the exploration of the two-part conceptualization of construct validation of test score-use in order to improve test design: construct representativeness (components, processes and knowledge structures that are involved in test responses; and nomothetic span (relationship of the test to other measures of individual differences) as outlined by Embretson (1983, 1994);
2) research on the exploration of the factor structure of test performance or questionnaires in order to better understand the abilities assessed by tests or test taker characteristics collected through questionnaires of homogeneous and heterogeneous groups of test

Address for correspondence: Dr Antony John Kunnan, Associate Professor, TESOL Program, Charter School of Education, California State University, Los Angeles, Los Angeles, CA 90032-8143, USA; e-mail: akunnan@calstatela.edu

Language Testing 1998 15 (3) 295–332 0265-5322(98)LT153OA

takers or respondents (examples, Muthén, 1989; Kunnan, 1995; Purpura, 1996; Ginther and Stevens, 1998);

3) research on the exploration of the hypothesized relationships among test taker characteristics or background (or external factors), test taking strategies and test performance in a second or foreign language context to better understand the effect of salient test taker characteristics on test performance (examples, Muthén, 1988, 1989; Kunnan, 1995; Purpura, 1996);
4) research on the exploration of hypothesized relationships among test task characteristics and test performance in order to better understand the effect of different test tasks (multiple methods) on test performance. SEM would provide a more powerful mechanism for this type of investigation than regression, which has been previously used by researchers (examples, Freedle and Kostin, 1993; Bachman, Davidson and Milanovic, 1996); and
5) research on the exploration of population heterogeneity among test takers, since this is generally typical of most data sets (including language assessment data sets, especially in large-scale high stakes ESL/EFL tests). For example, in instructional or language assessment settings, widely varying curricula, opportunities to learn, exposure or instruction in target language may require data to be analysed as independent multi-samples (example, Kunnan, 1995) as well as simultaneous multi-samples (example, Ginther and Stevens, 1998).

2 What are the previous research studies that have applied SEM?

The earliest use of SEM in language assessment research was by Bachman and Palmer in three studies of construct validation of the FSI Oral Interview (1981), components of communicative proficiency (1982) and self-ratings of communicative language ability (1989). Other researchers who have used SEM included Swinton and Powers (1980), who examined the component abilities that underlie performance on the TOEFL, Purcell (1983), who investigated models of pronunciation accuracy, Fouly (1985), who investigated the relationships among learner variables and second language proficiency, Wang (1988), who investigated cognitive achievement and psychological orientation among language minority groups, Hale, Rock and Jirele (1989), who studied the factor structure of the TOEFL, and Turner (1989), who investigated second language cloze test performance.

During the 1980s, Gardner and other second language acquisition researchers were using SEM with second language acquisition data (Gardner, Lalonde and Pierson, 1983; Gardner *et al*., 1987; Gardner,

1988; Clement and Kruidenier, 1985; Ely, 1986) to investigate motivation, aptitude, and attitude as factors that affect second language acquisition.

Four recent SEM applications include Sasaki (1993), who investigated the relationships among second language proficiency, foreign language aptitude, and intelligence, Kunnan (1995), who investigated the influence of some test taker characteristics on test performance in tests of English as a foreign language, Purpura (1996), who investigated the relationships between test takers' cognitive and metacognitive strategy use and second language test performance, and Ginther and Stevens (1998), who investigated the factor structure of an Advanced Placement Spanish language examination among four different Spanish-speaking test taking groups.

This short list illustrates that SEM applications in language assessment research have been very few in number and the range of investigations equally small.

II Overview of SEM

In the 50th anniversary issue of *Psychometrika*, Bentler (1986) documented the unusual rapidity of the growth of structural modelling or structural equation modelling (SEM) methodology. This was followed by the Austin and Wolfe (1991) annotated SEM bibliography of 294 technical and substantive applications articles, dating back to an article by Wright in 1921. More recently, as an indicator of the surge of this literature, Tremblay and Gardner (1996) recorded the growth of SEM in psychological journals and Austin and Calderón (1996) annotated 320 new publications to provide an update of the theoretical-technical, applicational, pedagogic and philosophical articles. From these bibliographies, it is clear that researchers in fields such as biology, education, economics, marketing, medicine, psychology and sociology are leading the way in the use of SEM.

A search for applications of SEM in the field of language assessment will certainly not turn up more than a handful of entries at the most (none is listed in the above bibliographies though). This low level of interest in SEM among language assessment researchers is probably due to many reasons, the chief ones being the lack of a pedagogic introduction to SEM for language assessment research, very few examples of SEM application to language assessment data, and very little discussion of the merits and the limitations of SEM for the field of language assessment. This introduction will attempt to fill this gap by providing a pedagogic perspective with examples so that reading the articles that follow and using SEM methodology with language assessment data will hopefully be more attractive.

1 What is structural equation modelling?

SEM can be viewed as a coming together of several models: multiple regression, path analysis and factor analysis. In the regression model, a directional relationship between two sets of measured variables, the dependent variable and a set of regressor variables, is posited and evaluated; the path analysis model tests theoretical relationships among independent measured variables and dependent measured variables and the direct and indirect effects of the independent variables on the dependent variables; the factor analysis model attempts to determine which observed measured variables share common variance–covariance characteristics with a latent construct or factor. SEM is an integration of these models, offering the mechanism to hypothesize relationships between constructs and measured variables and among constructs based on substantive theory. As Bentler (1995) puts it, 'linear structural equation modeling is a useful methodology for statistically specifying, estimating, and testing hypothesized relationships among a set of substantively meaningful variables' (p. ix).

SEM applications are so wide today that Marcoulides and Schumacker (1996) state that 'the use of the term structural equation modeling is broadly defined to accommodate models that include latent variables, measurement errors in both dependent and independent latent constructs, multiple indicators, reciprocal causation, simultaneity and interdependence' (p. 1). The commercial software packages that implement SEM (e.g., AMOS, EQS, LISREL, LISCOMP, to name a few) offer different procedures, such as confirmatory factor analysis, path analysis, multitrait-multimethod matrix analysis, recursive and non-recursive models, multi-sample models, multi-level models, time series models, growth models and covariance structure analysis.

Procedurally, SEM attempts to explain a correlation or a covariance data matrix, derived from a set of *observed or measured variables*, that is hypothesized in a measurement model or a structural model. While the concept of a correlation matrix is well known, a few terms specific to SEM may need brief explanations here: *a covariance matrix* (or, more accurately, a variance–covariance matrix) is made up of variance terms on the diagonal and covariance terms on the off-diagonal; a measurement or a structural model that is hypothesized posits that a few factors or *latent variables*, which are smaller than the number of measured variables, are responsible for the covariance among the measured variables; and latent variables are of two types, *independent* and *dependent*; the former type are hypothesized to influence the latter type of latent variables. Other terms that are central to SEM will be explained during the course of this article.

2 What are the assumptions that have to be met?

The statistical assumptions about the data (measured variables) include the following:

a Level of measurement:[1] All four levels of measurement (categorical or nominal, ordinal, interval and ratio) can be used in SEM but the general recommendation is that these levels of measurement not be mixed in a correlational or covariance matrix. Covariance matrices are generally preferred over correlational matrices; if the latter are used, they must be appropriate to the level of measurement in the data. Correlation matrices consisting of Pearson product–moment correlation coefficients (when both variables are interval), the phi coefficients (when both variables are nominal) and the tetrachoric coefficients (when both variables are dichotomously recorded) are used widely in SEM. Other coefficients such as the biserial (when one variable is interval and the other recoded into a dichotomy), point-biserial (when one variable is interval and the other is dichotomous), the polyserial (when one is an ordinal and the other an interval variable) and the polychoric (when both are ordinal variables) may be required, depending on the combination of the type of measurement used. Many computer programs (for example, EQS, LISREL in PRELIS, Mx, SEPATH) now permit the analysis of models that have categorical variables as long as they are categorized versions of variables that are continuous.

b Distribution of values and normality: Most estimation procedures used for SEM assume that data are normally distributed. Univariate normality can be checked by examining the skewness and kurtosis of the measured variables. If non-normality is observed for a measured variable, it may be due to outliers, and possible solutions for this situation could include transforming the data or editing the data to exclude the outliers (especially if the outliers are less than 2 per cent of the data). Multivariate normality can be checked by observing the skewness and kurtosis for all the measured variables together. Most computer programs can assess normality; EQS, for example, provides two Mardia coefficients and case numbers of the data with the largest contribution to normalized multivariate kurtosis (see Bentler, 1995). If the data are not normally distributed, appropriate statistical estimation methods such as elliptical least squares and arbitrary least squares can be used. Recent simulation studies (e.g., Chou, Bentler and Satorra 1991; Chou and Bentler 1995) have shown that the

[1]Level of measurement is also known as type of scale.

asymptotic distribution free method does not work well except with very large samples. Other remedies that could be used for multivariate non-normality include the use of robust statistics such as Satorra–Bentler scaled test statistic (Satorra and Bentler, 1988, 1994), robust standard errors (Bentler and Dijkstra, 1985) and the use of item parcels and transformation of non-normal variables (see West, Finch and Curran, 1995).

c Linearity: Correlation coefficients assume linear relationships and the size of a correlation coefficient indicates the degree of linear relationships between two measured variables.[2] If the relationship is not linear, for example, curvilinear, correlation coefficients will not reflect this but this could be observed in a scatter plot. If a curvilinear relationship between two measured variables can be explained by a meaningful theory, then a nonlinear approach to SEM as suggested by the Kenny–Judd model (Kenny and Judd, 1984) can be attempted, though Jöreskog and Yang's (1996) article presents many concerns with this approach.[3]

d Sample size: Two issues with reference to sample size for SEM are the representativeness of the sample, which pertain to generalizability of results, and the accuracy and stability of estimates, which pertain to the reliability of results and the model. In order to address both the above issues, SEM researchers generally suggest large sample sizes, though a few of them have come up with actual minimum sample sizes: Ding, Velicer and Harlow (1995) indicate 100 to 150 subjects and Boomsma (1987) recommends 400. Another way of going about sample size would be to consider sample size in relation to the number of variables: Bentler and Chou (1987) suggest a ratio as low as five subjects per variable for normal and elliptical distributions and 10 subjects per variable for other distributions. In practice, however, it seems that sample sizes less than 150 may not ensure stable estimates or for that matter representativeness. Another approach to the problem of sample size has been the recent argument that the evaluation of models should be based on power considerations, rather than on sample size (see Kaplan, 1995).

---

[2]Simply put, a linear relationship between two measured variables, when plotted on a graph, forms a straight line because the rate of change or frequency in one measured variable is consistent with changes in the other measured variable.

[3]References for further study on this subject include Jaccard and Wan (1996) and Schumacker and Marcoulides (1998).

e Stochastic relationships: The relationships between independent latent variables and dependent latent variables may be stochastic, or not fully deterministic. This means that not all of the variation in the dependent latent variables is accounted for by the independent latent variable. The unexplained variance of the dependent latent variable is represented by a stochastic residual associated with each relationship.

III Steps and concepts

1 What are the steps in SEM?

Bollen and Long (1993) list five steps in an SEM application: model specification, model identification, model estimation, testing model fit, and model respecification. But prior to embarking on the five steps, the type of model formulation planned should be clear in the researcher's mind as this will have ramifications for evaluation and interpretation of the model.

Jöreskog (1993) identifies three types of model formulation modes: *Strictly Confirmatory*, in which a researcher formulates one single model and tests this model with empirical data, which should be accepted or rejected based on interpretable parameter estimates and model fit; *Model Comparison*, in which a researcher specifies several alternative models and tests these models with empirical data; and *Model Generating*, in which a researcher specifies a tentative initial model, tests it with empirical data, then respecifies the model based on suggestions from an SEM analysis and substantive theory, tests the respecified model again, respecifies the model again, tests the respecified the model, and so on, until a satisfactory model emerges. With these three types of model formulation modes in mind, Bollen and Long's (1993) steps can be now described.

Step 1: Model specification: The first step refers to the model specification, which a researcher formulates based on a theory or prior research in the substantive area. An SEM model generally consists of two parts, the *measurement model* and the *structural model*. The measurement model specifies the relationships between measured variables and latent variables that are not directly measurable but are specified, and the structural model specifies the direct and indirect relationships among the latent variables.

The researcher also needs to understand the four types of parameters in a model: *paths* between measured variables and latent variables, *uniquenesses* associated with measured variables, *variances* of independent latent variables, and *disturbances* associated with latent variables. Each path will be associated with a parameter, some of which will be free to be estimated and some fixed.

In order to understand this, a brief explanation regarding parameters is essential: all parameters that require specification represent relationships in a model (between measured variables, between measured variables and latent variables) and characteristics of variables themselves, such as means and variances. In a model, parameters can be either *free parameters* to be estimated from the data, *fixed parameters* (typically fixed at 1) and therefore not estimated, or *constrained parameters*, which are constrained to equal one or more other parameters.

As an illustration, consider Figure 1, in which the measured variables, V1 to V3, load on the latent variable, F1.[4] These variables, based on substantive reasoning, were not allowed to load on any of the other latent variables (F2, F3 and F4). The loadings from these

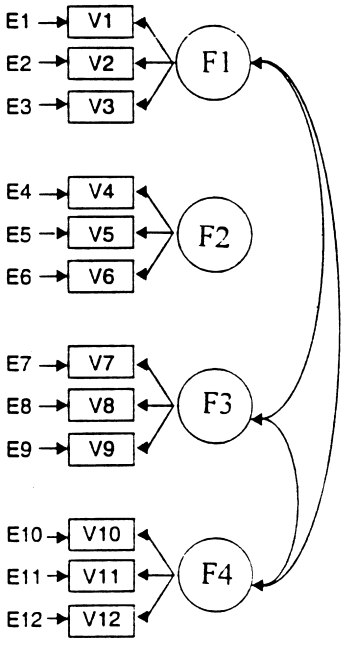

Figure 1 Measurement model 1

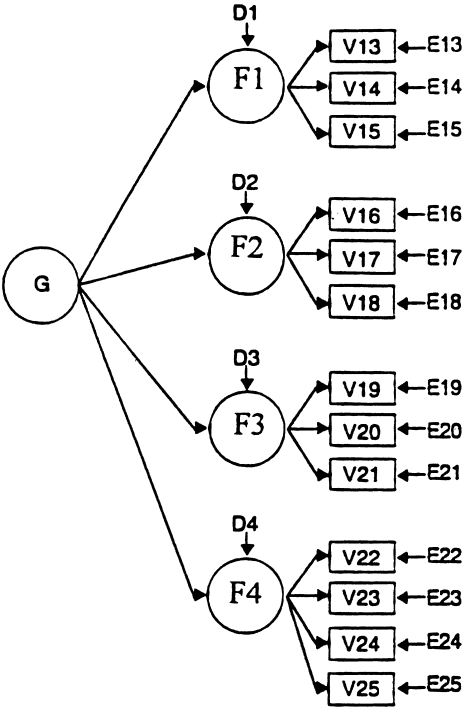

Figure 2 Measurement model 2: higher-order

Figures 1 and 2 reproduced with permission from Kunnan, *Test Takes Characteristics and Test performance: Studies in Language Testing 2*. 1996. Cambridge University Press.

[4]EQS terminology and notations are used here; similar plain English terminology and notations are used in the SIMPLIS Command Language in LISREL 8 (see Jöreskog and Sörbom, 1993) and in some other computer programs.

variables on the latent variables represent free parameters to be estimated and all others that were constrained and not allowed to be estimated represent fixed parameters, fixed to zero. An additional constraint for identification purposes is that one of the measured variables loading on a latent construct must have its factor loading fixed to 1 as a reference indicator so that the measurement scale for each latent construct can be set because of the indeterminacy between the variance of the latent construct and the loadings of the measured variables. In Figure 1, the loadings of V3, V6, V9 and V12 were set fixed to 1, while all the other loadings were free to be estimated.

1) *Measurement models*: Figure 1 presents a measurement model that defines four *independent latent variables* (diagrammed using circles), labelled F1, F2, F3 and F4, each of which is associated with three measured variables (diagrammed using rectangles), labelled V1 to V12. Associated with each measured variable is a uniqueness (diagrammed with arrows pointing from the uniquenesses to the associated measured variable), labelled E1 to E12. Three of the latent variables, namely, F1, F3 and F4, are correlated with one another, and curved arrows indicate these relationships. This model is formulated in a strict confirmatory mode and the main interest in this model is to assess how well the measured variables measure each of the latent variables. This will be indicated by the estimated path coefficients relating the measured variables to the latent variables (corresponding to factors) and their associated standard errors, and the estimated uniquenesses, and their associated errors. The fit indices provide an overall indication of the fit of the model (of the measured variables and their relationships to the latent variable) through the maximum likelihood estimation method.

   Figure 2 presents another measurement model that depicts four independent latent variables and measured variables with a latent variable labelled G for general factor. This type of model is termed a higher-order, second-order, or hierarchical factor model, once again formulated in a strict confirmatory mode. In addition to the other elements of the model, residuals of latent variables F1, F2, F3 and F4, termed disturbances (diagrammed with arrows pointing from the disturbances to the associated latent variables), labelled D1 to D4, are indicated. As always, the main interest in these models is to assess how well the measured variables measure each latent variable.
2) *Structural models*: Figure 3 presents a structural model that depicts the relationships among the independent latent variables (F1, F2, F3 and F4), as in Figure 1, and the dependent latent

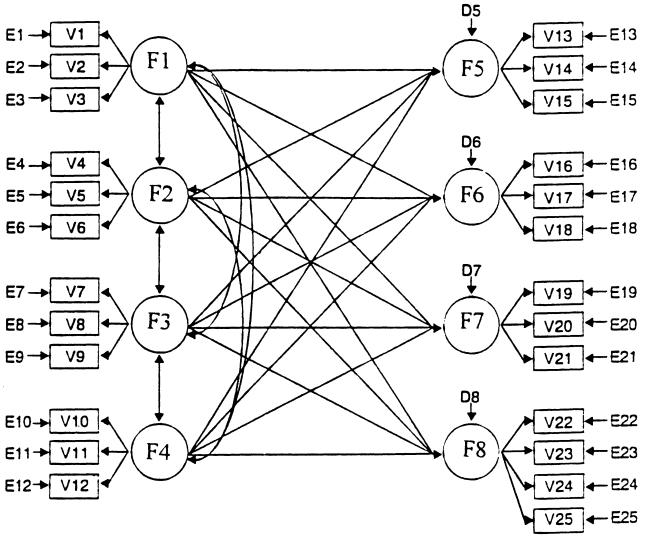

Figure 3 Structural model 1

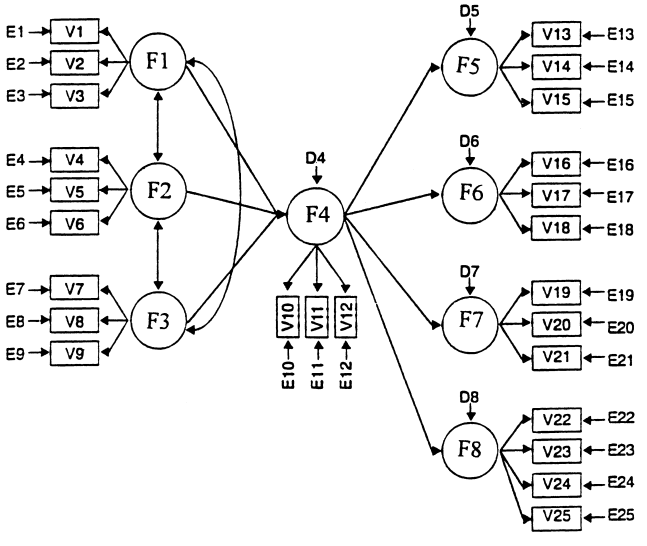

Figure 4 Structural model 2

Figures 3 and 4 reproduced with permission from Kunnan, *Test Takes Characteristics and Test performance: Studies in Language Testing 2*. 1996. Cambridge University Press.

variables (F5, F6, F7, F8 without the G factor) and the correlations among the independent latent variables. Figure 4 presents a different structural model from that in Figure 3, in which three latent independent variables, F1, F2 and F3, influence another

latent variable, F4, which in turn influences the dependent latent variables (F5, F6, F7 and F8). These two models are formulated in the model comparison mode (based on competing hypotheses)

and the main interest in these models is to assess the relationships among the latent variables and examine how well the overall model fits with the data submitted.

Step 2: Model identification: After a model is specified, an important consideration is the identification of the model. As Hoyle (1995) states, 'Identification concerns the correspondence between the information to be estimated – the free parameters – and the information from which it is to be estimated – the observed variances and covariances' (p. 4). According to Schumacker and Lomax (1996), 'model identification … depends on the specification of parameters as free, fixed or constrained. Once the model is specified and the parameter specifications are indicated, the parameters are combined to form one and only one model-implied variance–covariance matrix' (p. 100). Identification problems may still exist and it depends in part on the amount of information in the covariance matrix (derived from the empirical data) necessary for the unique estimation of the parameters of the model.

Three levels of identification are possible: a model can be said to be *under-identified*, if one or more parameters were not uniquely estimable from the covariance matrix; a model can be said to be *just-identified* if all the parameters were uniquely estimable; and a model can be said to be *over-identified* when there is more than one way to estimate the parameter(s). An under-identified model is one that should not be trusted, as the parameter estimates may be unstable, but if additional constraints are imposed, such a model could become identified, either just-identified or over-identified, both of which are acceptable identification levels. The t-Rule is a simple way to examine model identification (i.e., $\frac{1}{2}$ p (p + 1) nonredundant elements in the variance–covariance matrix versus the number of estimated parameters, as a necessary condition of estimation (see Marcoulides and Hershberger, 1997). The difference between the number of nonredundant elements in the variance–covariance matrix and the number of model parameters to be estimated is the degrees of freedom (df). For example, in a multiple regression model with five predictors (i.e., Y = 1, X = 5), there are 6(7)/2 = 21 nonredundant elements in the variance–covariance matrix and 21 parameters that need to be estimated (i.e., 1 intercept, 5 regression coefficients, and 15 covariances among predictors). Thus, with df = 0, the model is just-identified.[5]

---

[5]I am thankful to an external reviewer for information regarding the t-Rule.

Two practical problems of identification that are typically encountered include the sample covariance matrix not being *positive definite*, a problem caused by some measured variables being perfectly predicted by others, and *nonconvergence* in the estimation process. Remedies for both these situations are better model specification and better start values or initial estimates for the estimation process (Chou and Bentler, 1995).

Step 3: Model estimation: The primary purpose of model estimation is to obtain estimates for all the parameters to be estimated. Several estimation methods are currently available in the commercial computer packages: the maximum likelihood (ML) and generalized least squares (GLS) estimation methods, which assume multivariate normality distribution, elliptical least squares (ELS) for elliptical distribution theory and arbitrary least squares (ALS) for arbitrary distribution theory.[6]

If the measured variables are interval-scaled and the data are multivariate normal, ML estimates, standard errors, and χ^2 values and goodness of fit summaries would be appropriate. But if these statistical assumptions cannot be assumed, then the ML estimates may not be appropriate. The Robust option available in EQS provides the Satorra–Bentler scaled χ^2 statistic that is designed to have a distribution that is more closely approximated by χ^2 than the usual statistic and robust standard errors that are correct in large samples even if the distributional assumptions are not met.

A standardized solution to the model is available in some computer packages. In EQS, standardization is done for all the Vs (variable) and Fs (factor) model variables, including errors and disturbances. All loadings, therefore, have a similar interpretation. One important feature of this standardization process is that the parameters that were fixed will also receive new estimates.

Step 4: Testing model fit: When a model is evaluated for fit, both global model fit and individual parameter fit need to be examined. Global fit can be evaluated through a judicious choice of statistics and although several measures of model fit are available, researchers may have particular difficulty in choosing model fit statistics because, as argued by Tanaka (1993), model fit should be multifaceted. The two most popular ways of evaluating model fit, as Hu and Bentler (1995) state, 'are those that involve the χ^2 goodness-of-fit statistic and the so called fit indices that have been offered to supplement the

[6]More estimation procedures are available in the SEM packages but only the most popular are mentioned in this article.

χ^2 test' (p. 76). These indices can be used to test models that have been formulated based on prior research or substantive literature or models that are respecified and submitted based on model estimates and model interpretability.

The χ^2 statistic, the most widely used single statistical test in SEM, assesses the magnitude of the difference between the observed sample covariance matrix and the reproduced, or model-implied, covariance matrix. The χ^2 statistic is presented along with a probability level that indicates whether the statistic is statistically significant. A statistically *significant* χ^2 *test* indicates that the difference between observed and estimated matrices is due to sampling variation, while a statistically *nonsignificant* χ^2 *test* indicates that there is model fit though there is no certainty that other models might not have similar model fits. Thus, a researcher would be interested in obtaining a nonsignificant χ^2 with associated degrees of freedom.

Two popular approaches that estimate a best-fitting solution and an evaluation of model fit when the data are multivariate normal are: maximum likelihood (ML) and generalized least squares (GLS). Schumacker and Lomax (1996) state that 'the ML estimates are consistent, unbiased, efficient, scale-invariant, scale-free, and normally distributed' (p. 125) if the data meet the multivariate normality assumption. The GLS estimates are similar to ML estimates though the estimates are best when data meet the assumptions of an elliptical distribution theory. When non-normality is assumed, two other estimates are recommended: Browne's (1984) asymptotical distribution free (ADF) criterion and Satorra–Bentler scaled χ^2 statistic (Satorra and Bentler, 1990).[7]

Other goodness-of-fit (GFI) indices, generally formulated to range in value from 0 (no model fit) to 1.0 (perfect model fit), that should be consulted in addition to the χ^2 statistic include: the comparative fit index (CFI) or the goodness of fit index (GFI) and the Tucker–Lewis index (TLI) or the Bentler–Bonnet normed fit index (BBNFI). Generally, if any of these indices are above .90, the rule of thumb is that there is a recommendation from the indices that there is model fit, pending examination of the χ^2 statistic and model interpretability. Schumacker and Lomax (1996) present a table of goodness of fit criteria and acceptable fit interpretation (p. 121, Table 7.1) for these and other indices.

One practical concern that might be critical is the effect of sample size on the goodness of fit indices. Saris, Ronden and Satorra (1987)

---

[7] Recent simulation studies (e.g., Chou, Bentler and Satorra, 1991; Chou and Bentler, 1995) have shown that the asymptotic distribution free method does not work well except with very large samples.

found that the χ^2 statistic may be acceptable only in the case of large samples, and Bearden, Sharma and Teel (1992) found that the mean of NFI values tend to be far less than 1.0 when sample size is small. An index suggested by Wheaton *et al*. (1977) as a way of dealing with the effect of large sample sizes on the χ^2 statistic is the χ^2/df (degrees of freedom) ratio. Based on their experience, they suggested that a ratio of around 5.0 was reasonable, but Stage (1990) argues that 2.5 or less is an indication of model fit. Because of all this confusion, Schumacker and Lomax (1996) state that 'it is prudent to report multiple measures rather than to rely on a single choice' (p. 134).

Hatcher (1996) presents a short summary of results when a model provides an ideal fit. With a slight modification, the following are the characteristics of ideal model fit with attention to individual parameters:

1) Very few standardized residuals should exceed 0.02 or 0.03.
2) The *p* value associated with the model fit χ^2 test should exceed .05; the closer to 1.00, the better.
3) The CFI and the non-normed fit index (NNFI) should both exceed .90; the closer to 1.00, the better.
4) The R^2 value for each dependent factor should be relatively large, in terms of the size of what is typically obtained in the field.
5) The *t* statistics for each path coefficient and the standardized path coefficients (from Vs to Fs and Fs to Fs) should be significant (exceed .05) and meaningful.
6) Overall model fit obviously should be assessed by considering both statistical fit and model interpretability.

Step 5: Model respecification: When a model is specified and submitted for evaluation, it may or may not have goodness of fit indices that suggest model fit. Such models are considered mis-specified models and they will have biased parameter estimates known as specification errors. Detection of such specification errors, which is known as specification search, could result in information that could be used in model respecification.

The first set of estimates to examine would be the parameter estimates, especially for their expected direction and statistical significance. If a parameter is found to be nonsignificant, the parameter could be fixed to 0 in respecifying the model, but this should be done only if this is meaningful and is substantively supported. Another set of estimates to examine are the residual matrix for the magnitude of the values. If a given variable has large residual values, this indicates that the variable is mis-specified, but if many variables have large values, this indicates that there is a general mis-specification of the model.

Modification indices are often used to evaluate hypotheses concerning whether a restriction is statistically inconsistent with the data (see Sörbom, 1989). Other methods involve evaluating the expected parameter change (EPC) that a specific parameter may take if freed (see Saris, Satorra and Sörbom, 1987), and examining the vanishing tetrads using the TETRAD program (see Glymour *et al*., 1987).[8]

The Lagrange Multiplier and Wald statistic provided in the EQS software program offer further help in model respecification. The Lagrange Multiplier statistic indicates what effect the freeing of fixed parameters would have on the model; in other words, it suggests the addition of parameters. The Wald statistic, in contrast, indicates which parameters should be dropped from the model. But caution should be exercised here as only additions or deletions of parameters that are substantively justifiable should be added or dropped.

Once model respecification is accomplished and the resubmitted model is found to have model fit, cross-validation of this final model with another sample (or another random half of the same sample) is necessary before the final model can be considered an accurate representation of a phenomenon.

2 What are the considerations when interpreting SEM results?

a Limitations: The following limitations need to be considered in the interpretation of results:

1) *Incompleteness of a model*: This refers to difficulty in knowing whether a model is complete or not and whether additional measured variables would improve model fit or not. Researchers would need to depend on the substantive literature or prior research for guidance in this matter.
2) *Undecidability of best model*: This refers to difficulty in deciding which model is superior when two or more models have the same number of parameters and have equally good model fit. These 'models obviously cannot be distinguished mathematically' but they might be different 'only in terms of their substantive meaning and the interpretability of solutions obtained when they are fit to the data' (MacCallum, 1995: 30). Other analytical procedures that could be used in the evaluation of equivalent models are 1) examining the values for the squared multiple correlations

[8] A more recent approach based upon a Tabu search procedure has been introduced for conducting specification searches in SEM (see Marcoulides, 1998; Marcoulides, Drezner and Schumacker, in press).

of equations from a set of equivalent models to arrive at a preferred model (Jöreskog and Sörbom, 1989) and 2) using the new complexity criterion for model evaluation termed ICOMP developed by Bozdogan (1988) and discussed in some detail in Williams, Bozdogan and Aiman-Smith (1996).

b Philosophical: The following philosophical considerations need to be understood before a model is interpreted:

1) *Objective state of affairs*: The point here is that a structural model is a mathematical model that represents an objective state of affairs, written in the language of objects which have properties that can cause, determine or influence other objects (see Mulaik, 1994). Thus, measured variables have to be considered objective objects with inherent properties such as values that can influence, however small, the values of other measured variables.
2) *Defeasible reasoning*: Following Mulaik and James (1995), this type of reasoning refers to how conclusions regarding a structural model could be justified but that those conclusions could be 'defeated' by acquiring new and relevant data (Pollock, 1986). Reasoning of this type could also be dialectical, in which a researcher can have a dialectic with other researchers, between himself or herself and nature, and within the researcher, and with reasonable certainty forming the logic of discovery, in contrast to Popper's (1959) assertions in *The logic of discovery*.
3) *Causality*: Though SEM had been cast in a strict causal framework by some researchers in the past (e.g., Blalock, 1971; James, Mulaik and Brett, 1982), most recent authors and editors (e.g., Bollen, 1989; Bollen and Long, 1993; Hoyle, 1995; Marcoulides and Schumacker, 1996; Schumacker and Lomax, 1996) prefer the term *structural equation modelling* or *covariance structure analysis*, indicating clearly that establishing causality from correlations is not the focus of SEM. As Muthén (1992) argues, 'In my view, these statistical analyses have very little to do with causality … these techniques are not devices for rigorous testing of causal theories, but merely a powerful way of analyzing covariance structures. It would be very healthy if more researchers abandon thinking of and using terms such as *cause* and *effect*' (p. 82). Along the same lines, de Leeuw (1985) writes, 'the cause–effect terminology cannot be defended, except in those rare cases (such as Mendelian genetics)' (p. 372).

3 What is multi-sample structural modelling?

The five steps outlined above are useful for SEM applications that involve either a single sample or multiple samples. Multi-sample structural modelling is often necessary because even though data are frequently obtained as if they were from a single population, a closer examination of the subjects might uncover population heterogeneity that rules out a single population. Muthén (1989) argues that 'in educational achievement modeling with factor analysis and item response theory, the homogeneity assumption is unrealistic when applied to a sample of students with varying instructional background' (p. 558). He goes on to give examples from modelling of mathematics achievement for US eighth grade students, where opportunity to learn test items in algebra and geometry may be different due to varying curricula or tracks, from survey research, where the validity and reliability of some items can be expected to vary by race, gender, region and issue salience, and from psychiatric epidemiology, where surveys are concerned with data from a mixture of 'normal' and 'abnormal' subjects.

Similarly, in cancer and arthritis research, data are analysed separately for ethnic groups. Two recent studies are: Stein *et al*., who examined the factor structure of barriers to the use of mammography among Black, White and Hispanic women, and Coulton *et al*., who found factor invariance across three ethnic groups for Arthritis Impact Measure Scales (both cited in Bentler and Stein, 1992). In the field of education, Bryk, Lee and Holland (1993) show how multi-sample Catholic and public school data need to be used in order to examine the different effects the school systems have on student achievement.

Multi-sample data can be created by sorting cases in a data set into separate samples based on salient characteristics of the subjects such as personal attributes (e.g., age, gender, ethnicity, native language), economic background (socioeconomic status), educational background (school grade level, university level, achievement or ability level), psychological background (motivation level, attitude), citizenship or residence (nation, state or region), culturally or socioeconomically different groups, or groups receiving different treatments. Obviously, not all the above characteristics will be salient in any given research context, so prior research or substantive theory should be the guiding force for the creation of multiple samples.

Three types of multi-sample analysis have been identified:

1) *Independent multi-sample analysis*: In this type of analysis, multiple samples are examined separately and the parameter estimates and goodness of fit indices between the samples are 'eyeballed' as statistical comparison of equivalence of parameters cannot be made between the samples nor can mean differences

between the samples be estimated. Schumacker and Lomax (1996) state that this type of analysis is not strictly multi-sample modeling because only one sample is evaluated at a time, but this approach may be useful in establishing baseline models for the samples of interest prior to simultaneous multi-sample analysis.

2) *Simultaneous multi-sample analysis*: In this type of analysis, multiple samples are examined simultaneously and the analysis can statistically determine whether certain parameters or parameter matrices are equivalent across the samples for any of the measurement and structural models.
3) *Covariance structure analysis*: In this type of analysis, multiple samples are examined simultaneously and the analysis can statistically determine whether there are mean differences for the variables and/or the structural equations across the samples.

Three practical problems, however, may hinder multi-sample analysis. First, large sample sizes and distribution of values for each sample that meet the assumption of multivariate normality are absolutely crucial for stable variances and covariances. Secondly, membership in the samples has to be based on variables (such as age, gender, ethnicity) that do not vary across variables (such as opportunity to learn, which may vary across test items). Thirdly, when multi-sample data are created, the data may be hierarchical in nature. For example, students may be sampled on a test from some classrooms which may be sampled on the test from some schools. Two alternative methods provide solutions in these situations: MIMIC (multiple indicators, multiple causes) modelling (Muthén, 1988, 1989) and multi-level modeling (Muthén and Satorra, 1989; Raudenbush and Bryk, 1988; McArdle and Hamagami, 1996), though more applications are needed before these methods become routinely used.

IV SEM application matters

1 Software programs

Several commercial software programs implement SEM. The most popular are EQS developed by Bentler (1995) and Bentler and Wu (1995), LISREL 8-SIMPLIS (developed by Jöreskog and Sörbom, 1993), WINAMOS (developed by Arbuckle, 1995) and LISCOMP (developed by Muthén, 1987). Other SEM programs include Mx, SAS PROC CALIS and STATISTICA-SEPATH (see Schumacker and Lomax, 1996, for more information). Byrne (1994) demonstrates the EQS approach; in Byrne (1995) she compares EQS and LISREL approaches and in Byrne (1998) she presents applications and programming in LISREL, PRELIS and SIMPLIS.

2 Preliminary analyses

The following analyses should be completed as preparation so that baseline information about the data is available.

1) *Frequency distributions and descriptive statistics*: Frequencies and descriptive statistics need to be examined for discrepancies with the data, such as incomplete cases and outliers. If such discrepancies exist, the data need to be cleaned up by removing cases with missing values or by imputing values for the variable and by examining cases that are outliers and dropping them where justified. Most software programs need complete data sets for conducting SEM, though the WINAMOS program can handle missing data directly using its unique ML algorithm (see Arbuckle, 1996).

   Frequencies and descriptives also need to be examined for both univariate and multivariate normality of distributions by examining kurtosis and skewness of the data. If values for kurtosis or skewness of the individual variables of interest exceed ± 2.0, the data are not normally distributed. Most software programs provide information indicating which cases are contributing to the non-normal distribution; at that stage, one option is to delete those cases from the data set when there is evidence that there has been a computer scanning error, improper marking on the scantron card, or other justifiable reason. Extreme caution needs to be exercised in the deletion of cases, however, as unjustifiable deletions of cases could lead to a biased data set. If the data still turn out to be non-normal, estimation methods and fit statistics that are appropriate should be used (such as ML Robust and Satorra–Bentler scaled statistics and robust standard errors).
2) *Correlations and exploratory factor analysis*: Correlations need to be examined for the strength of relationships among variables of interest. If correlations among variables of interest are high, then at the factor analysis stage, the oblimin (correlated) solution should be used; if correlations are low, the orthogonal (uncorrelated) solution should be used. If correlations among variables are linear (after examining scatter plots), the variables are acceptable for SEM; if the correlations among variables are nonlinear, appropriate data transformations for those variables would be necessary prior to structural modeling.

   Exploratory factor analysis (EFA) results need to be examined for the factor structure of the variables of interest. If the correlated solution looks interpretable, then factors should be correlated at the modeling stage; if the uncorrelated solution looks

interpretable, then factors should be uncorrelated at the modelling stage. It might be necessary to run the factor analysis procedure several times before a final solution that is interpretable is arrived at. Each EFA analysis might need to be varied in terms of number of factors extracted as well as the type of solution sought (correlated or uncorrelated).

3) t-*tests and ANOVA*: *t*-tests and/or ANOVA results need to be examined to determine whether there are statistically significant differences among any categorical grouping variables of interest, such as gender, native language and test preparation. If one of the above tests shows that the sample mean of one group is statistically significantly different from that of another sample group of interest, this finding should be used in planning multi-sample structural modelling, particularly if means structures are to be analysed (see, Byrne, 1994 for worked example). If statistically significant mean differences are not found, then multi-sample structural modelling may not be meaningful.

3 The modelling process

The first aim of the researcher should be to evaluate the measurement models, and only after these have been evaluated, to evaluate the structural models. If multi-sample models are considered necessary, these should be evaluated next. Jöreskog and Sörbom (1993) provide invaluable advice on this point: 'The testing of the structural model, i.e., the testing of the initially specified theory, may be meaningless unless it is first established that the measurement model holds. If the chosen indicators for a construct do not measure that construct, the specified theory must be modified before it can be tested. Therefore, the measurement model should be tested before the structural relationships are tested. It may be useful to do this for each construct separately, then for the constructs taken two at a time, and then for all constructs simultaneously' (p. 113). This approach is often referred to as the multi-stage approach to model evaluation (see Anderson and Gerbing, 1988). Unfortunately (or fortunately), not everyone agrees with this approach and a one-step approach to model evaluation has been recently proposed (see Hayduk, 1996).[9]

The researcher also needs to keep in mind at this stage Jöreskog's (1993) model formulation modes, if only as a way of clarifying the overall goal of the research that is being planned. The researcher

[9] An upcoming issue of the journal *Structural Equation Modeling* is devoted entirely to this debate.

needs to ask then whether the overall goal of the research is to strictly confirm a model, to compare models or to generate models.

4 Preparing a model[10]

1) *Draw a basic model*: A model should be formulated based on substantive theory in the field, prior research or by relying on results of the exploratory factor analysis, provided that the factor structure is substantively meaningful.
2) *Assign names to variables and factors*: The Vs and Fs that are part of the model need to be assigned names or labels that match the variables in the data file and the factors from substantive theory or exploratory factor analysis.
3) *Specify which path coefficients are to be estimated (from Fs to Vs and Fs to Fs)*: These paths are the ones of interest to the researcher and in EQS these paths can be specified in two ways, by using the diagrammer or by writing equations in the program. With the diagrammer, the paths are drawn with a one-way arrow between Vs and Fs; corresponding Es and Ds associated with Vs and Fs, respectively, will be drawn by the diagrammer automatically. In addition, one of the paths from an F to a V must be fixed at a value of 1.0 for identification purposes and to allow the other paths to be free (either with an * or with an * and a start value like 1 or .50 or so; an asterisk without a start value will mean that the software program will determine its own start values). Alternatively, these paths can be specified by writing equations in the program itself. For the measurement model presented in Figure 1, these take the form:

```
V1  = 1*F1 + E1 ;
V2  = 1*F1 + E2 ;
V3  = 1  F1 + E3 ;
V4  = 1*F2 + E4 ;
V5  = 1*F2 + E5 ;
V6  = 1  F2 + E6 ;
V7  = 1*F3 + E7 ;
V8  = 1  F3 + E8 ;
V9  = 1*F3 + E9 ;
V10 = 1*F4 + E10;
V11 = 1*F4 + E11;
V12 = 1  F4 + E12;
```

[10]EQS 5.5 was used for this section.

For the structural model presented in Figure 3, the equations take the form:

```
F5 = 1*F4 + D5;
F6 = 1*F2 + 1*F4 + D6;
F7 = 1*F1 + 1*F2 + 1*F3 + D7;
F8 = 1*F1 + 1*F2 + 1*F3 + D8;
```

4) *Specify variances and covariances to be estimated among Fs*: If Fs are not to be correlated, no arrows between or among the Fs should be drawn. Fs need to be linked with a two-way curved arrow if the Fs are going to be correlated factors. In addition, Fs need to have fixed variances (with 1.0 value) and generally free covariances among Fs (with an * or start value, as preferred).
5) *Run model, estimate parameters and evaluate model*: Once the specifications are done and the model looks like the model anticipated, the program should be run, parameters estimated and model fit evaluated (at the global and individual parameter levels).
6) *Choose LM test and/or Wald test*: Model modification can be carried out with the help of the two tests: the Lagrange multiplier (LM) test evaluates whether freeing one of the fixed parameters will improve the model, and indicates which parameters should be added to the model, while the Wald test evaluates whether the free parameters in a model are necessary from a statistical point of view, and indicates which parameters should be dropped from the model. Though these tests are able to suggest individual parameter modifications, there are problems with modifying the model one parameter at a time. Furthermore, these model modifications suggested by the above tests must be substantively meaningful for the modifications to be made and for the modified models to be evaluated.

5 Reviewing the output file

The output file needs to be reviewed very carefully. With more experience, routine messages can be skipped. The following is the order of a standard output file in EQS 5.5:

1) *Program control information*: This is a repeat of the program submitted or drawn with the diagrammer and the Build EQS procedure. This information should be checked to confirm whether what was requested is reflected in the program.
2) *Error messages*: Error messages indicating serious problems will be listed here. If EQS cannot correct these problems, it will not proceed to the computational sections. Errors need to be

corrected and the program resubmitted for another job run. Minor errors listed here also need to be corrected, although these may not interrupt the job run.

3) *Univariate statistics*: These are basic descriptive statistics that can help you check the means, standard deviations, kurtosis and skewness for all variables. Univariate normality of the data set through kurtosis and skewness values need to be checked; these values should generally be under ± 2.
4) *Multivariate statistics*: These are the test results for multivariate normality; values should be under 5.0. If they are not, cases that contribute to this non-normality may need to be dropped or ML Robust needs to be used as an estimation procedure. The case numbers with the largest contribution to normalized multivariate kurtosis (also known as outliers) help identify the cases to be dropped; EQS permits the deletion of a maximum of 10 cases through the DELETE command in the program for each job run.
5) *Covariance matrix*: This is the matrix that EQS generates from the raw data, and which is the basis for EQS estimates and runs. The Bentler–Weeks structural representation provides the list of dependent and independent variables in the program submitted for the job run. Check for accurate representation.
6) *Model statistics and condition codes*: A list of model statistics that the program provides is presented here. In addition, the condition codes information is particularly important: the condition code that indicates that all is well with the model is: '*Parameter estimates appear in order. No special problems were encountered during optimization*'. If condition codes such as 'linear dependencies', 'constraint problems', and 'convergence problem' appear, model respecification may be required.
7) *Residual matrices*: These matrices show the residual values for each variable. The standard residual matrix needs to be checked for any variables with high residuals. The average absolute standardized residuals should be small; nearly .01 or so. The list of the largest standardized residuals will indicate which variables are problematic. The distribution of these standardized residuals in a graph shows how symmetrical the distribution is. The more the residuals of the variables are clustered around 0, the better, and this is an indicator of model fit.
8) *Fit indices*: There is a lot of information here, so focus on a few indices such as the χ^2, degrees of freedom, probability value, and the four indices at the end, the Satorra–Bentler scaled χ^2 (SB χ^2), the Bentler–Bonett normed fit index (BBNFI) and the comparative fit index (CFI). With regard to the χ^2, the lower

this figure, the better the model fit; the higher the probability value (particularly above 0.01), the better the model fit. With regard to the last three indices, the closer they are to 1.00, the better the model fit. The CFI is the most trusted and, therefore, this index is the preferred indicator, but model fit should be multifaceted and therefore one index is not enough indication of model fit or misfit. In addition, statistical model fit is only one aspect of structural modelling; model interpretability should never be overlooked.

9) *Iterative summary*: This provides the number of iterations it took the program to 'converge', or to minimize the function; the fewer the iterations, the better.
10) *Unstandardized and standardized solutions*: This provides a list of all Vs, Fs and Es, first with unstandardized estimates, standard errors and test statistics, and then a standardized solution. The standardized solution is the one to examine; high V and F values and low error values are what is expected for model fit.
11) *Additional estimates*: If commands such as constraints, inequalities, the LM test and the Wald test were requested in the program, additional estimates would be provided here.

6 Example models

The examples presented here are from Kunnan (1995; see the study for the theoretical foundation and methodological procedures), which investigated the influence of test taker characteristics on test performance in tests of English as a foreign language by exploring the relationships between the two groups of variables. The measurement models of the test taker characteristics and the test performances will be presented first, followed by the structural models linking both these measurement models.

a Measurement model for test-taker characteristics: Figure 5 (outlined previously in Figure 1) presents the measurement model of four independent latent variables that make up the test taker characteristics: HCF (Home Country Formal Instruction), HCI (Home Country Informal Exposure), ESC (English Speaking Country) and MON (Monitoring) based on 12 measured variables. This measurement model (formulated in the confirmatory mode) was based on results from an exploratory factor analysis and the model was also substantively meaningful, based on prior research in SLA on formal and informal instruction and self-monitoring. Parameter estimates for all measured variables and correlations among the latent variables are presented. Table 1 presents the goodness-of-fit statistics and these indicate that the model is acceptable.

Table 1 Goodness-of-fit indices for measurement model for test taker characteristics (N = 380)

χ^2	67.90 df = 51
p <	0.057
χ^2/df	1.33
SB χ^2	67.85
BBNFI	0.95
CFI	0.99

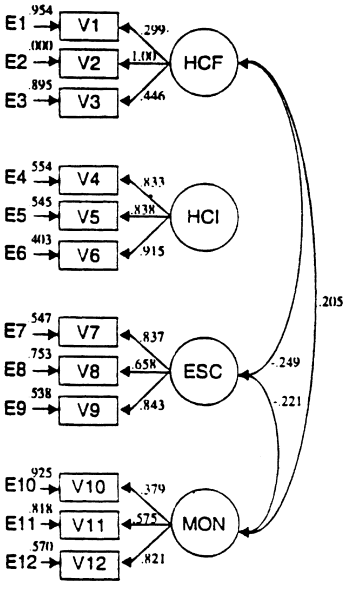

Figure 5 Measurement model for test taker characteristics with standardized estimates

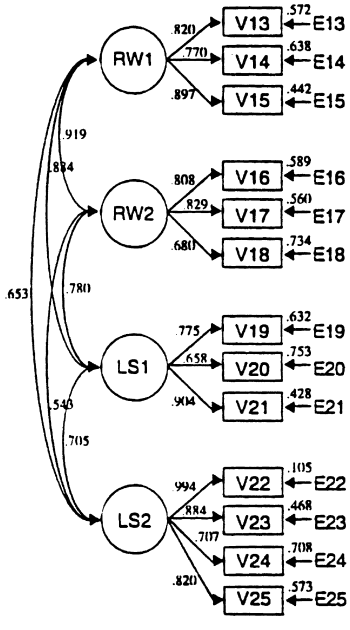

Figure 6 Measurement model for test performance with standardized estimates

Figures 5 and 6 reproduced with permission from Kunnan, *Test Takes Characteristics and Test performance: Studies in Language Testing 2*. 1996. Cambridge University Press.

Table 2 Goodness-of-fit indices for measurement model for test performance (N = 380)

χ^2	221.70 df = 59
p <	0.001
χ^2/df	3.76
SB χ^2	220.78
BBNFI	0.94
CFI	0.95

b Measurement model for test performance: Figure 6 (outlined above in Figure 2) presents the measurement model of four independent latent variables that make up EFL test performance, RW1 (Reading-Writing 1-FCE papers), RW2 (Reading-Writing 2-TOEFL sections 2, 3; TEW), LS1 (Listening-Speaking 1-interactional; FCE papers 4 and 5 and TOEFL 1) and LS2 (Listening-Speaking 2-non-interactional; SPEAK), based on 13 measured variables. This measurement model (formulated in the confirmatory mode) was based on results from an exploratory factor analysis and the model was also substantively meaningful, based on prior language assessment research that has generally indicated that language proficiency is multicomponential. The model is presented with parameter estimates for all measured variables and correlations among the latent variables. Table 2 presents the goodness-of-fit statistics and these indicate that the model is acceptable.

c Structural model 1 for test taker characteristics and test performance: Figure 7 (outlined previously in Figure 3) presents the first structural model for test taker characteristics and test performance

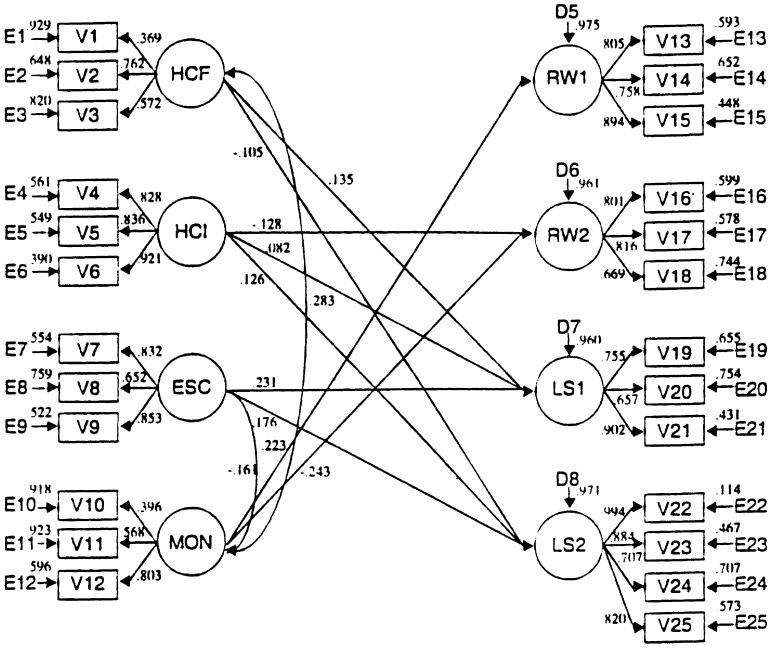

Figure 7 Structural model 1 for test taker characteristics and test performance with standardized estimates

Figure 7 reproduced with permission from Kunnan, *Test Takes Characteristics and Test performance: Studies in Language Testing 2*. 1996. Cambridge University Press.

linking both the independent and dependent variables and their associated measured variables and errors. This model is based on the hypothesis that these latent variables representing test taker characteristics influence latent variables representing EFL test performance and this is a substantively meaningful model in the field of SLA. Parameter estimates for all measured variables and correlations among the latent variables, and path coefficients between the independent and dependent latent variables, are presented. Table 3 presents the goodness-of-fit statistics and these indicate that the model is reasonably acceptable.

d Structural model 2 for test taker characteristics and test performance: Figure 8 (outlined previously in Figure 4) presents the second structural model for test taker characteristics and test performance linking both the independent and dependent variables and their associated measured variables and errors, as the first was only reasonably acceptable. This model is a variation of Gardner's (1988) socio-educational model in which Gardner posits social milieu (cultural beliefs) as the first group of variables, followed by individual difference variables (intelligence, language aptitude, motivation and situational anxiety), then followed by SLA contexts (formal and informal language training), and finally linguistic and nonlinguistic outcomes. In Figure 8, language training variables (HCF, HCI and ESC), analogous to Gardner's SLA contexts, influence monitoring (MON), which in turn influences test performance (RW1, RW2, LS1 and LS2) analogous to Gardner's linguistic outcomes. Substantively, this model was considered more meaningful than structural model 1 (presented in Figure 7).

Parameter estimates for all measured variables and correlations among the latent variables, and path coefficients between the independent and dependent latent variables, are presented. Table 4 presents the goodness-of-fit statistics and these indicate that the model is reasonably acceptable.

Model comparison between structural models 1 and 2 indicate that while both models are statistically (based on individual parameter

Table 3 Goodness-of-fit indices for structural model 1 (N = 380)

χ^2	577.66 df = 258
$p <$	0.001
χ^2/df	2.24
SB χ^2	568.07
BBNFI	0.89
CFI	0.94

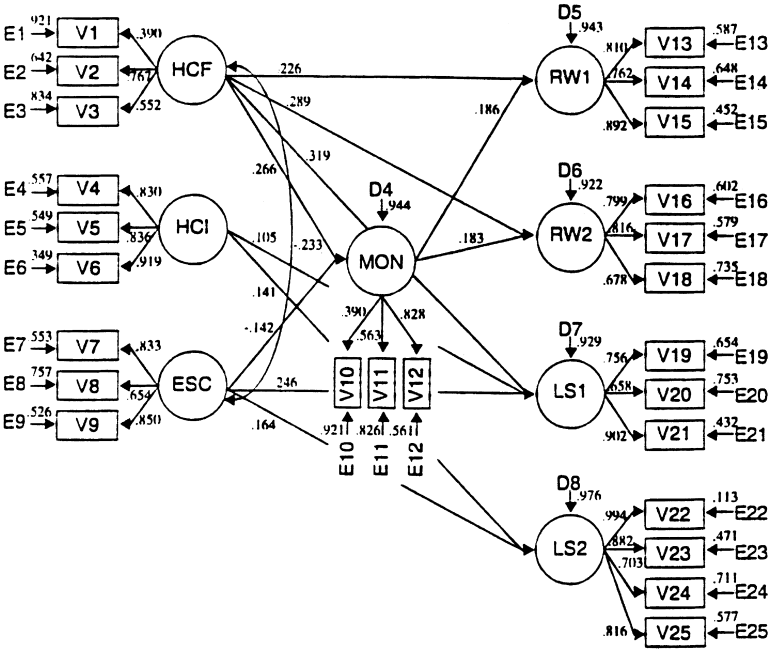

Figure 8 Structural model 2 for test taker characteristics and test performance with standardized estimates

Figure 8 reproduced with permission from Kunnan, *Test Takes Characteristics and Test performance: Studies in Language Testing 2*. 1996. Cambridge University Press.

estimates and goodness-of-fit indices) and substantively acceptable, structural model 2 is the preferred model. This is based on the consideration that though the χ^2 difference between the two models is 21.28 and highly significant ($p < .001$), the second model, based on Gardner's prior research in SLA, provides clearer interpretation that formal and informal instruction and exposure influence self-monitoring and self-monitoring in turn influences EFL test performance.[11]

Table 4 Goodness-of-fit indices for structural model 2 (N = 380)

χ^2	556.38 df = 257
$p <$	0.001
χ^2/df	2.17
SB χ^2	398.77
BBNFI	0.90
CFI	0.94

[11]For more discussions on these and alternative models that were evaluated, see Kunnan (1995).

V Critical discussions and the future

1 What are the recent critical discussions regarding SEM?

Despite the major technological advances and the impressive number of SEM applications in many fields, the methodology is not without its detractors. In an introduction to a recent debate on SEM organized with 12 researchers, editor Shaffer (1992) positions a major critique – Freedman's (1992) – of SEM thus: 'During the last 30 years, it [SEM] has become a popular methodological tool for social scientists. Yet the apparent precision of the inferences in a path analysis depends on a network of highly restrictive stochastic assumptions. David Freedman explicates those assumptions, specifies their implications, and points out how difficult they are to verify and how their violation can have devastating effects on the resulting inferences … He notes that his critique applies even more strongly to the popular latent structure or covariance structure modeling, which includes an extra layer of abstraction and additional strong assumptions due to the latent, indirectly measured, variables in the causal chains' (p. x).

The basis of Freedman's (1992) critique is Hope's study (1984) of the relationship between schooling and social mobility in Scotland and the US. In this study, Hope asks whether Scotland is indeed the meritocracy it is often said to be and whether it is more so than the US. His Scottish data were from the Scottish Mental Survey drawn in 1947 from all 11-year-old boys born on the first day of every other month and followed until 1964, with data available on nearly 600 boys on four variables – father's occupation, boy's IQ, boy's education and boy's occupation. His US data were drawn from Jencks *et al*. (1972). The equations for his four-variable path model, for both groups, as stated by Freedman (p. 19), are as follows:

education = $a \times$ IQ + $b \times$ (father's occupation) + error
occupation = $c \times$ (education) + $d \times$ IQ + $e \times$ (father's occupation) + error

The path model is illustrated in Figure 9.

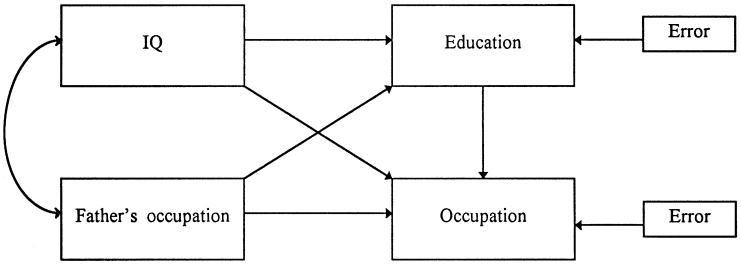

Figure 9 Hope's (1984) path model, reproduced from Freedman (1992)

Freedman critiques the study and stresses its weaknesses. First, he points out the difficulties in measuring IQ and success in life and the difference in measuring the school variable in Scotland (based on track system) and in the US (based on number of years). Secondly, he criticizes the study for its lack of attention to linearity, the autonomy coefficient, and omitted variables such as whether social stratification in the Highlands of Scotland is the same as in the cities or whether schools in Edinburgh are similar to those in Glasgow. Finally, he asks, 'given such problems, what connects the model to reality?' (p. 22). Freedman (1992) sums up his critique thus: 'My opinion is that investigators need to think more about the underlying social processes, and look more closely at the data, without the distorting prism of conventional (and largely) stochastic models' (p. 27).

Bentler (1992) dismisses Freedman's critiques by stating: '[his] critiques remind me primarily of the many published discussions of the misuse of a variety of basic statistical methods (e.g., *t* test, χ^2, etc.) that have graced applied statistics across the decades. It is almost certain that structural modeling will be misused at least as much as these older methods, since the techniques are more complex, and also require substantially more substantive theory and greater insight into data by the practitioner' (p. 53).

Muthén (1992) suggests that 'the field can produce interesting and useful studies using path analysis and more general structural equation models, but only when carried out by skillful practitioners. The practitioners need better methodological training and statisticians should contribute to this process' (p. 86). And, in a rejoinder to all the commentators, Freedman concedes that 'models may have heuristic value' (p. 123).

In a more recent article, Mueller (1997) cautions SEM users by quoting Duncan and Wolfe: 'The study of structural equation models can be divided into two parts: the easy part and the hard part' (Duncan, quoted by Mueller, p. 355). 'The easy part is mathematical. The hard part is constructing causal models that are consistent with sound theory' (Wolfe, quoted by Mueller, p. 355). Mueller (1997) goes on to assert that Duncan and Wolfe 'pointed to *the* fundamental truth in SEM: No matter how technically sophisticated the employed statistical techniques, SEM analyses can only be beneficial to the researcher if a strong substantive theory underlies the initially hypothesized model(s)' (p. 355).

These critiques and cautions underscore the point that unless researchers use SEM with utmost care and skill they will succumb to the usual difficulties associated with this methodology: inaccurate and imprecise measurement of variables, omitted variables, simplistic

SEM models, and lack of connection between models and substantive theory and reality.

2 What are some directions for SEM applications in language assessment research?

The main question is how can language assessment researchers best apply SEM methodology? One line of research could be the focus on the factor structure of test performance through penetrating investigations. Test performance data from many but similar language tests should advance our knowledge and understanding of the factor structure and componentiality of language ability.

A second promising line of research could be the focus on testing hypothesized relationships among some salient test taker characteristics and test performance. Characteristics such as personal attributes (gender, age, ethnicity, native language, impairment), educational background (level of education, parents' education, ability level), home background (socioeconomic status, parents' occupation and income), as well as psychological characteristics (achievement motivation, attitude, personality, anxiety, risk-taking), and cognitive characteristics (learning style and strategies, field in/dependence) are research areas awaiting investigations, first through measurement models and then with structural models where the salient characteristics can be modelled with test performance. Of course, not all the above characteristics will translate to meaningful models in every situation but much can be learned from careful selection of characteristics and well-designed investigations. Purpura's study (in this volume) is an example of this line of research.

A third line of research that is much more practical would be a multifaceted approach towards construct validation of test-score interpretation that would include the use of SEM techniques along with factor analysis and the multi-trait multi-method approaches.

A fourth line of research is an expansion of any of the first three by penetrating more deeply into test performance data and creating multi-sample analyses based on salient personal attributes (such as gender, age, ethnicity, native language, impairment). These multi-sample studies may be of special interest to researchers who are concerned with the equal validity, reliability and fairness of the test-score interpretation. Bae and Bachman's study (in this volume) is an example of this line of research.

Two lines of research that need the attention of researchers are: investigations of the mental processes that test takers employ when they encounter a task, and investigations of personal growth in ability through longitudinal studies that can identify the route and patterns

of growth in ability. SEM growth models can provide powerful analytic tools for such studies.

VI Conclusion

In this article, an attempt has been made to pedagogically introduce SEM through a presentation of SEM steps and concepts, examples of SEM applications to language assessment data, and a discussion of methodological and substantive issues.

Many more areas that are not unimportant have had to be left out of this article for reasons of space. These include topics such as nonlinear models, multi-level models and growth models, data manipulation for simulation studies such as Monte Carlo, bootstrapping and jackknife, and cross-validation approaches, all of which are introduced in Schumacker and Lomax (1996) and discussed in Marcoulides and Schumacker (1996). Some key conceptual and philosophical issues that could not be touched on include the theory-laden nature of observations that is foundational to SEM, since the way we see 'objects' (measured variables) may be based on our own training and disposition, and thus models may be suspected to be mere artefacts; the problem of underspecification and generalizability, and the thorny issue of the nature of the inferences that can be made on the basis of SEM. These issues are dealt with in Britt (1997), Bechtel (1988), Blalock (1982), and James, Mulaik and Brett (1982).

Help with pull-down menus with particular software programs (example, EQS, LISREL) was not provided in this article. But new texts (Byrne, 1994, 1995, 1998) and program manuals (Bentler and Wu, 1995; Jöreskog and Sörbom, 1994) have excellent step by step plans with click and drag displays for diagrams and a nonmatrix-based, English-like command language (instead of writing program code) that make the mechanical part of structural modelling relatively easy. In addition to these resources, *Structural Equation Modeling: A Multidisciplinary Journal* (published by Lawrence Erlbaum Associates, Inc.) and SEMNET, the SEM special interest group on the Internet, should help researchers keep abreast with technological and conceptual advances.[12]

[12]To subscribe to SEMNET, send the following message:
SUBSCRIBE SEMNET first-name last-name
to the Internet address: listserv@ua1vm.ua.edu

VII References

Anderson, J.C. and **Gerbing, D.W.** 1988: Structural equation modeling in practice: a review and recommended two-step approach. *Psychological Bulletin* 103, 411–23.

Arbuckle, J.L. 1995: *AMOS*. Chicago: Smallwaters Corporation.

—— 1996: Full information estimation in the presence of missing data. In Marcoulides, G.A. and Schumacker, R.L., editors, *Advanced structural equation modeling: issues and techniques*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Austin, J.T. and **Calderón, R.F.** 1996: Theoretical and technical contributions to SEM: an updated bibliography. *Structural Equation Modeling* 3, 105–75.

Austin, J.T. and **Wolfe, L.M.** 1991: Annotated bibliography of SEM: technical work. *British Journal of Mathematical and Statistical Psychology* 44, 93–152.

Bachman, L.F., Davidson, F. and **Milanovic, M.** 1996: The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing* 13, 125–50.

Bachman, L.F. and **Palmer, A.** 1981: The construct validation of the FSI oral interview. *Language Learning* 31, 67–86.

—— 1982: The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 449–65.

—— 1989: The construct validation of self-ratings of communicative language ability. *Language Testing* 6, 14–29.

Bearden, W.O., Sharma, S. and **Teel, J.R.** 1992: Sample size effects on chi-square and other statistics used in evaluating causal models. *Journal of Marketing Research* 19, 425–530.

Bechtel, W. 1988: *Philosophy of science*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bentler, P.M. 1986: Structural modeling and *Psychometrika*: an historical perspective on growth and achievements. *Psychometrika* 51, 35–51.

—— 1992: Structural modeling and the scientific method: comments on Freedman's critique. In Shaffer, J., editor, *The role of models in nonexperimental social sciences*, Washington, DC: AERA/ASA, 53–59.

—— 1995: *EQS: structural equations program manual*. Encino: CA, Multivariate Software, Inc.

Bentler, P.M. and **Chou, C.-P.** 1987: Practical issues in SEM. *Sociological Methods and Research* 16, 78–117.

Bentler, P.M. and Dijkstra, T. 1985: Efficient estimation via linearization in structural models. In Krishnaiah, P.R., editor, *Multivariate analysis VI*. Amsterdam: North-Holland, pp. 9–42.

Bentler, P.M. and **Stein, J.** 1992: Structural equation models in medical research. *Statistical Methods in Medical Research* 1, 159–81.

Bentler, P.M. and **Wu, E.** 1995: *EQS for Windows: user's guide*. Encino, CA: Multivariate Software, Inc.

Blalock, H.M. 1971: *Causal models in the social sciences*. Chicago: Aldine-Atherton.

—— editor, 1982: *Conceptualization and measurement in the social sciences*. Newbury Park, CA: Sage.

Bollen, K.A. 1989: *Structural equations with latent variables*. New York: Wiley.

Bollen, K.A. and **Long, J.S.,** editors, 1993: *Testing structural equation models*. Newbury Park, CA: Sage.

Boomsma, A. 1987: The robustness of maximum likelihood estimation in structural equation models. In Cuttance, P. and Ecob, R., editors, *Structural modeling by example*, Cambridge: Cambridge University Press, 160–88.

Bozdogan, H. 1988: ICOMP: a new model selection criterion. In Bock, H.H., editor, *Classification and related methods of data analysis*, Amsterdam: North-Holland, 599–608.

Britt, D.W. 1997: *A conceptual introduction to modeling: qualitative and quantitative perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Browne, M.W. 1984: Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* 37, 62–83.

Bryk, A.S., Lee, V.E. and **Holland, P.B.** 1993: *Catholic schools and the common good*. Cambridge, MA: Harvard University Press.

Byrne, B.M. 1994: *Structural equation modeling with EQS and EQS/Windows*. Thousand Oaks, CA: Sage.

—— 1995: One application of SEM from two perspectives: exploring the EQS and LISREL strategies. In Hoyle, R., editor, *Structural equation modeling*, Thousand Oaks, CA: Sage, 138–57.

—— 1998: *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Chou, C.-P. and **Bentler, P.** 1995: Estimates and tests in SEM. In Hoyle, R., editor, *Structural equation modeling*, Thousand Oaks, CA: Sage, 37–55.

Chou, C.P., Bentler, P. and **Satorra, A.** 1991: Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a Monte Carlo study. *British Journal of Mathematical and Statistical Psychology* 44, 345–57.

Clement, R. and **Kruidenier, B.G.** 1985: Aptitude, attitude and motivation in second language proficiency: a test of clement's model. *Journal of Language and Social Psychology* 4, 21–37.

Cuttance, P. and **Ecob, R.,** editors, 1987: *Structural modelling by example*. Cambridge: Cambridge University Press.

de Leeuw, J. 1985: Review of books by Long, Everitt, Saris and Stronkhorst. *Psychometrika* 50, 371–75.

Ding, L., Velicer, W.F. and **Harlow, L.L.** 1995: Effects of estimation methods, number of indicators per factor, and improper solutions on SEM fit indices. *Structural Equation Modeling* 2, 119–43.

Ely, C.M. 1986: Language learning data: a description and causal analysis. *Modern Language Journal* 70, 28–35.

Embretson, S. 1983: Construct validity: construct representation versus nomthetic span. *Psychological Bulletin* 93, 179–97.

—— 1994: Applications of cognitive design systems to test development. In Reynolds, C., editor, *Cognitive assessment*, New York: Plenum, 107–35.

Fouly, K. 1985: A confirmatory multivariate study of the nature of second language proficiency and its relationships to learner variables. Unpublished Ph.D. dissertation, University of Illinois, Urbana.

Freedle, R. and **Kostin, I.** 1993: The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing* 10, 131–70.

Freedman, D. 1992: As others see us: a case study in path analysis. *Journal of Educational Statistics* 12, 101–28. [Reprinted in Shaffer, J., 1992, *The role of models in nonexperimental social sciences*, Washington, DC: AERA/ASA, 3–30.]

Gardner, R.C. 1988: The socio-educational model of second language learning: assumptions, findings and issues. *Language Learning* 38, 101–26.

Gardner, R.C., Lalonde, R.N., Moorcraft, R. and **Evers, F.T.** 1987: Second language attrition: the role of motivation and use. *Journal of Language and Social Psychology* 6, 1–47.

Gardner, R.C., Lalonde, R.N. and **Pierson, R.** 1983: The socio-educational model of SLA: an investigation using LISREL causal modeling. *Journal of Language and Social Psychology* 2, 1–15.

Ginther, A. and **Stevens, J.** 1998: Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish language examination. In Kunnan, A.J., editor, *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 169–94.

Glymour, C., Scheines, R., Spirtes, P. and **Kelley, K.** 1987: *Discovering causal structure: artificial intelligence, philosophy of science, and statistical modeling*. San Diego, CA: Academic Press.

Hale, G.A., Rock, D.A. and **Jirele, T.** 1989: *Confirmatory factor analysis of the TOEFL*. TOEFL Research Report 32. Princeton: Educational Testing Service.

Hatcher, L. 1996: Using SAS PROCALIS path analysis: an introduction. *Structural Equation Modeling* 3, 176–92.

Hayduk, L.A. 1996: LISREL: issues, debates and strategies. Baltimore, MD: The Johns Hopkins University Press.

Hope, K. 1984: *As others see us: schooling and social mobility in Scotland and the United States*. New York: Cambridge University Press.

Hoyle, R., editor, 1995: *Structural equation modeling*. Thousand Oaks, CA: Sage.

Hu. L.-T. and **Bentler, P.M.** 1995: Evaluating model fit. In Hoyle, R., editor, *Structural equation modeling*, Thousand Oaks, CA: Sage, 76–99.

Jaccard, J. and **Wan, C.K.** 1996: *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage.

James, L.R., Mulaik, S.A. and **Brett, J.M.** 1982: *Causal analysis: assumptions, models and data*. Beverly Hills, CA: Sage.

Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H., Heyns, B. and **Michelson, S.** 1972: *Inequality: a reassessment of the effect of family and schooling in America*. New York: Basic Books.

Jöreskog, K.G. 1993: Testing structural equation models. In Bollen, K. and Long, J.S., editors, *Testing structural equation models*, Newbury Park: CA: Sage, 294–316.

Joreskog, K.G. and **Sörbom, D.** 1989: *LISREL 7: a guide to the program and applications* (2nd edn). Chicago: SPSS.

—— 1994: *LISREL 8: a guide to the program and applications*. Chicago: Scientific Software.

Jöreskog, K.G. and **Yang, F.** 1996: Nonlinear structural equation models: the Kenny-Judd model with interaction effects. In Marcoulides, G.A. and Schumacker, R.E., editors, *Advanced SEM: issues and techniques*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 57–88.

Kaplan, D. 1995: Statistical power in structural equation modeling. In Hoyle, R., editor, *Structural equation modeling*. Thousand Oaks, CA: Sage, pp. 100–17.

Kenny, D.A. and **Judd, C.M.** 1984: Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin* 96, 201–10.

Kunna, A.J. 1995: *Test taker characteristics and test performance: a structural modelling approach*. Cambridge: Cambridge University Press.

MacCullum, R.C. 1995: Model specification. In Hoyle, R., editor, *Structural equation modeling*, Thousand Oaks, CA: Sage, 16–36.

Marcoulides, G.A. 1998: *Modern methods for business research*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Marcoulides, G.A., Drezner, Z. and **Schumacker, R.E.** in press: Model specification searches in structural equation modeling using Tabu search. *Structural Equation Modeling*.

Marcoulides, G.A. and **Hershberger, S.L.** 1997: *Multivariate statistical methods: a first course*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Marcoulides, G.A. and **Schumacker, R.E.,** editors, 1996: *Advanced SEM: issues and techniques*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McArdle, J.J. and **Hamagami, F.** 1996: Multilevel models from a multiple group structural equation perspective. In Marcoulides, G.A. and Schumacker, R.E., editors, *Advanced SEM: issues and techniques*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 89–124.

Mueller, R.O. 1997: Structural equation modeling: back to basics. *Structural Equation Modeling* 4, 353–69.

Mulaik, S.A. 1994: Kant, Wittgenstein, objectivity, and SEM. In Reynolds, C.R., editor, *Cognitive assessment: a multidisplinary perspective*, New York: Plenum, 209–36.

Mulaik, S.A. and **James, L.R.** 1995: Objectivity and reasoning in science and SEM. In Hoyle, R., editor, *Structural equation modeling*, Thousand Oaks, CA: Sage, 118–37.

Muthén, B.O. 1987: *LISCOMP: analysis of linear structural equations with a comprehensive measurement model*. Moorsville: Scientific Software.

—— 1988: Some uses of SEM in validity studies: extending IRT to external variables. In Wainer, H. and Braun, H., editors, *Test validity*, Hillside, NJ: Lawrence Erlbaum Associates, Inc., 213–38.

—— 1989: Some uses of SEM in validity studies: extending IRT to external variables. In Wainer, H. and Braun, H., editors, *Test validity*, Hillside, NJ: Lawrence Erlbaum Associates, Inc., 213–38.

—— 1992: Response to Fredman's critique of path analysis: improve credibility by better methodological training. In Shaffer, J., editor, *The role of models in nonexperimental social sciences*, Washington, DC: AERA/ASA, 80–86.

Muthén, B.O. and **Satorra, A.** 1989: Multilevel aspects of varying parameter in structural models. In Bock, H., editor, *Multilevel analysis of educational data*, San Diego: Academic Press, 87–99.

Pollock, J.L. 1986: *Contemporary theorems of knowledge*. Totowa, NJ: Rowman & Littlefield.

Popper, K. 1959: *The logic of discovery*. London: Hutchinson. [Original work published in 1935.]

Purcell, E.T. 1983: Models of pronunciation accuracy. In Oller, J.W., editor, *Issues in language testing research*, Rowley, MA: Newbury House, 133–53.

Purpura, J.E. 1996: Modeling the relationships between test takers' reported cognitive and metacognitive strategy use and performance on language tests. Unpublished Ph.D. dissertation, University of California, Los Angeles.

Raudenbush, S. and **Bryk, A.** 1988: Methodological advances in studying effects of schools and classrooms on student learning. *Review of Research in Education* 15, 423–76.

Saris, W.E., de Ronden, J. and **Satorra, A.** 1987: Testing structural equation models. In Cuttance, P. and Ecob, R., editors, *Structural modeling by example*, Cambridge: Cambridge University Press, 202–30.

Saris, W.E., Satorra, A. and **Sörbom, D.** 1987: The detection and correction of specification errors in structural equation models. In Clogg, C., editor, *Sociological methodology*, San Francisco, CA: Jossey-Bass, pp. 105–29.

Sasaki, M. 1993: Relationships among second language proficiency, foreign language aptitude and intelligence: a structural equation modeling approach. *Language Learning* 43, 313–44.

Satorra, A. and **Bentler, P.M.** 1988: Scaling corrections for chi-square statistics in covariance structure analysis. *Proceedings of the American Statistical Association*, 308–13.

—— 1990: Model conditions for asymptotic robustness in the analysis of linear relations. *Computational Statistics and Data Analysis* 10, 235–49.

—— 1994: Corrections to test statistics and standard errors in covariance structure analysis. In von Eye, A. and Clogg, C.C., editors, *Latent variables analysis: applications for development research*. Thousand Oaks, CA: Sage, pp. 399–419.

Schumacker, R.E. and **Lomax, R.** 1996: *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Schumacker, R.E. and **Marcoulides, G.A.,** editors, 1998: *Interaction and non-linear effects in structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Shaffer, J., editor, 1992: *The role of models in nonexperimental social sciences*. Washington, DC: AERA/ASA.

Sörbom, D. 1989: Model modification. *Psychometrika* 54, 371–84.

Stage, F.K. 1990: LISREL: an introduction and applications in higher education research. In Smart, J., editor, *Higher education handbook of theory and research 6*, New York: Agathon Press.

Swinton, S.S. and **Powers, D.E.** 1980: *Factor analysis of the TOEFL*. TOEFL Research Report 6. Princeton, NJ: Educational Testing Service.

Tanaka, J.S. 1993: Multifaceted conceptions of fit in structural equation models. In Bollen, K. and Long, J.S., editors, *Testing structural equation models*, Newbury Park, CA: Sage, 10–39.

Tremblay, P.F. and **Gardner, R.C.** 1996: On the growth of SEM in psychological journals. *Structural Equation Modeling* 3, 93–104.

Turner, C. 1989: The underlying factor structure of L2 cloze test performance in francophone, university-level students: causal modeling as an approach to construct validation. *Language Testing* 6, 172–97.

Wang, L.-S. 1988: A comparative analysis of cognitive achievement and psychological orientation among language minority groups: a LISREL approach. Ph.D. dissertation, University of Illinois, Urbana.

West, S.G., Finch, J.F. and **Curran, P.J.** 1985: SEM with nonnormal variables: problems and remedies. In Hoyle, R., editor, *Structural equation modeling*, Thousand Oaks, CA: Sage, 56–73.

Wheaton, B., Muthén, B., Alwin, D.F. and **Summers, G.F.** 1977: Assessing reliability and stability in panel models. In Haise, D.R., editor, *Sociological methodology 1977*, San Francisco: Jossey-Bass, 84–136.

Williams, L.J., Bozdogan, H. and **Aiman-Smith, L.** 1996: Inference problems with equivalent models. In Marcoulides, G.A. and Schumacker, R.E., editors, *Advanced SEM: issues and techniques*, Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 279–314.

Wright, S. 1992: Correlation and causation. *Journal of Agricultural Research* 20, 557–85.