2015

Structural Equation Modeling Reporting Practices for Language Assessment

Gary Ockey
Iowa State University, gockey@iastate.edu

Ikkyu Choi
Educational Testing Service

Structural Equation Modeling Reporting Practices for Language Assessment

Abstract
Studies that use structural equation modeling (SEM) techniques are increasingly encountered in the language assessment literature. This popularity has created the need for a set of guidelines that can indicate what should be included in a research report and make it possible for research consumers to judge the appropriateness of the interpretations made from a reported study. This article attempts to fill this void by providing a set of reporting guidelines appropriate for language assessment researchers.

Introduction

Structural equation modeling (SEM), also referred to as analysis of covariance structures and causal modeling, is a family of statistical techniques which includes confirmatory factor analysis (CFA), structural regression, path, growth, multiple-groups, and multi-trait multi-method (MTMM) models. SEM techniques have been used in language testing for various purposes, including to assess a test's (or other measure's) internal structure (Gu, 2014; In'nami & Koizumi, 2012; Sawaki, Stricker, & Oranje, 2009), to assess the effect of test methods on test performance (e.g., Llosa, 2007; Sawaki, 2007), to assess the equivalency of models for different populations (e.g., Llosa, 2005; Purpura, 1998; Shin, 2005), and to understand the effects of test tasks (Carr, 2006) or test taker characteristics such as language exposure (Kunnan, 1994), personality (Ockey, 2011) and strategy use (Purpura, 1997; 1999) on test performance. SEM has also been used by language assessment researchers to investigate properties of questionnaires (Phakiti, 2008; Purpura, 1999, 2004).

Since Kunnan (1998) lamented the dearth of SEM studies in the language testing literature, use of SEM in the field has grown exponentially. Unfortunately, this growth has not been accompanied by a set of clear guidelines for reporting the studies, ones that would make it easy for language testing researchers to replicate or evaluate the research. As a result, it is often difficult for consumers of SEM reports to interpret the findings of the reported studies. This concern is substantiated by research conducted by In'nami and Koizumi (2011). They investigated 50 articles published prior to 2008 that used SEM techniques and were found in language assessment or language learning journals.[i] Their analysis was based on principles that were considered to be best practice at the time they completed their study. While it may not have been completely appropriate to evaluate earlier work with more up-to-date thinking, the findings

suggest that many research reports in applied linguistics and language assessment research have not reported sufficient information for readers to evaluate and verify SEM results. For instance, the study found that 15 of 50 articles did not report the estimation method that was used to obtain a solution. The researchers indicated that this was quite problematic, particularly in cases in which the data required a particular technique for accurate estimation, such as with categorical data. They concluded that without information about the estimation procedure used, it was not possible to judge the validity of or interpret the findings of these studies. The authors found a similar situation with reporting of the extent to which the data were appropriate for SEM analyses. Numerous researchers failed to report on the extent to which data were normally distributed or whether or not they had complete data sets, in spite of the fact that non-normal and missing data are frequently encountered in language assessment studies and important to interpretations of SEM research. In'nami and Koizumi's research also indicated the importance of a set of reporting guidelines for model fit for language assessment researchers. In eight of fifty articles examined, even chi-square statistics for the overall model fit were not provided. In'nami and Koizumi compared the reporting of fit indices to the guidelines recommended by numerous researchers in other fields and found that few of the articles followed any of these guidelines. This finding suggested not only the need for researchers to report sufficient information but also the need for a set of guidelines that language assessment researchers would deem important to follow.

This paper is written for the purpose of providing a set of coherent reporting guidelines for research papers which employ SEM techniques in the language assessment field. Our specific aims for this paper are to: 1) guide researchers on what to report, so that attempts to appropriately interpret the findings and replicate the study can be made, and 2) help reviewers

and consumers of SEM studies judge the quality and legitimacy of the research and its implications.

We assume that readers of this paper are familiar with SEM concepts and techniques. We underscore the fact that this paper is not meant to be taken as an introductory account of SEM or a general review of SEM techniques. Moreover, the paper is not meant for an advanced SEM user audience seeking new or novel techniques or theory. The purpose of this paper is to provide guidelines for reporting SEM analyses in language assessment research. For an introduction to SEM for language testing researchers, readers might find it useful to read *An introduction to structural equation modeling for language assessment research* (Kunnan, 1998) or *Exploratory factor analysis and structural equation modeling* (Ockey, 2013). For a more in depth, yet highly accessible introduction to SEM research, we recommend Kline (2011). We note, however, that some basic SEM concepts important in the language testing field are not covered in this text. We provide some context for our guidelines along with the recommendations and justification for these recommendations to make them more readable and meaningful than a mere list of do's and don'ts. In particular, we include reminders of selected topics important for conducting SEM research and highlight a few issues that we believe are of particular relevance to reports prepared by language testing researchers. To achieve this aim, we provide a bit more context for some topics than others, recognizing that some readers would prefer more context while others would prefer less. Our hope is that we have provided sufficient context to make it clear what needs to be reported. We attempt to provide context in a conceptual manner without delving into technical details. This approach might be regarded as vague and sometimes even inaccurate for more technically oriented readers, but we believe that accessibility should be our priority in presenting this set of guidelines.

In developing these guidelines, we reviewed a number of existing recommendations for reporting analyses and results in other fields which use SEM techniques, including: *Reporting structural equation modeling results in Psychology and Aging: Some proposed guidelines* (Raykov, Tomer, & Nesselroade, 1991), *Writing about structural equation models* (Hoyle, & Panter, 1995), *Reporting analysis of covariance structures* (Boomsma, 2000), *Principles and practice in reporting structural equation analyses* (McDonald & Ho, 2002), *A beginner's guide to structural equation modeling* (Schumacker & Lomax, 2010), and *Reporting results from structural equation modeling analyses in Archives of Scientific Psychology* (Hoyle & Isherwood, 2013). Having reviewed these and other sources, we discovered that while these sources largely agree on what to report in a paper that uses SEM techniques, there are notable exceptions. This disagreement likely stems at least in part, from the field specific purposes of the research. This convinced us even more of the need for language testing to have its own set of guidelines for reporting research which uses SEM techniques. To present a coherent set of recommendations for reporting an SEM analysis, we frame our reporting guidelines under familiar SEM headings, namely, 1) model proposal, 2) model identification, 3) data, 4) parameter estimation, 5) model fit, 6) model interpretation, and 7) alternative models. We believe that framing our guidelines this way will also help researchers determine an appropriate place for reporting particular aspects of their research.

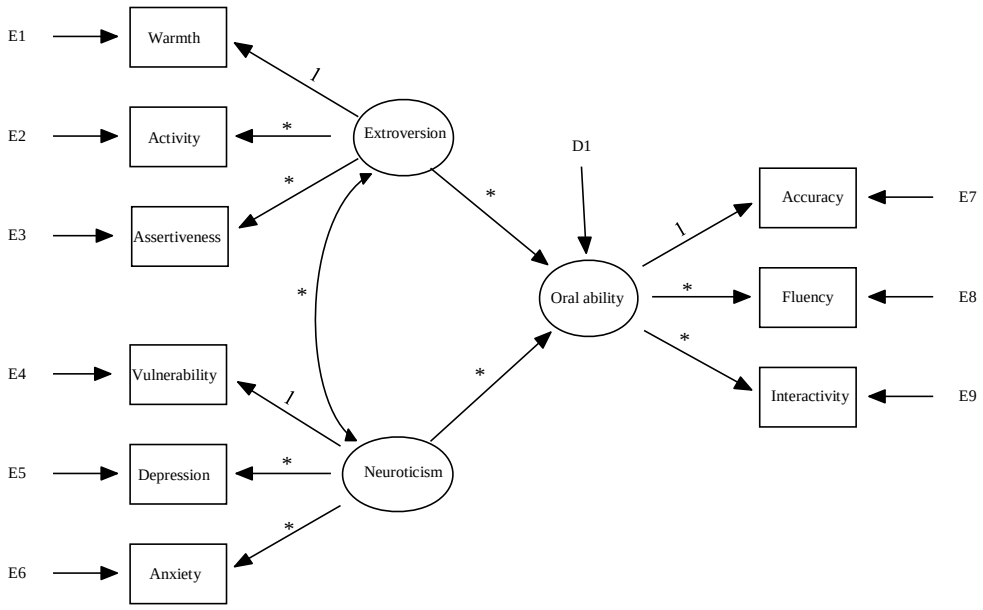
Model proposal

Proposed SEM models should be presented in path diagram form following established conventions. We recommend that SEM software programs, which include diagramming applications, be used for this purpose. As an illustration, Figure 1 presents an SEM model diagram which hypothesizes a relationship among some personality variables and second

language oral ability. The information provided in the diagram along with the explanations in the caption should be included in SEM reports.

Figure 1

Hypothesized relationships among personality factors and oral ability



Rectangles represent observed variables. Ovals indicate latent variables.
One-headed arrows indicate an expected directional relationship between two variables.
Two-headed arrows indicate covariation between two variables.
The asterisks on the arrows indicate that these relationships are freely estimated parameters.
The numbers on the arrows indicate that these relationships are fixed (at those numbers).
E indicates errors and D disturbances. Although not indicated in the diagram, the variances of the errors and the disturbance are all free parameters.

Additional information needed to show relationships among specialized models, such as parameter constraints in a multiple-groups model, should be indicated by symbols which are defined in the text. When models are large, it may be appropriate to separately display measurement models (i.e., the part of the overall model that relates measured variables to latent variables), and structural models (i.e., the part of the model that shows the relationship among

the latent variables). When this strategy is used, it should be made clear in both the text and the figure captions that these models are parts of a larger model. If variables, which are deemed to measure the same construct, are bundled together (e.g., item parcels), this should also be indicated by the figure and described in the text. Prose that provide logical justification for the bundling scheme should accompany the figures.

Theory for both modeled and unmodeled relationships to justify the hypothesized model should be provided. For example, the model in Figure 1 does not have an arrow between *Warmth* and *Activity*, demonstrating the assumption that any relationship between the two observed variables is due to the common factor, *Extroversion*. Theory and findings from prior research should be provided to support not only the hypothesized relationship between extroversion and warmth and activity but also to indicate why no direct relationship between warmth and activity is hypothesized. It is usually not necessary to discuss each relationship separately, however.

Model identification

Reporting details about the model specifications is important for interpreting SEM research. Identification involves specifying conditions to ensure a unique set of solutions for the statistical model under question. In practice, model identification comes down to satisfying the following conditions: have more data points (i.e., unique elements of an observed variance-covariance matrix) than freely estimated parameters; have more than one observed variable per latent variable; and assigning a scale to each latent variable.[ii] It is not necessary to report on satisfying the second condition since it should be evident from the path diagram whether or not it is met. The first condition can be checked by examining model degrees of freedom, which is equal to the difference between the number of data points and the number of free parameters. To

ensure that this condition has been met, model degrees of freedom should be reported. The third condition is usually satisfied by either fixing the variance of a latent variable or the factor loading on one of the observed variables at a convenient value (often at one). To show that this condition has been satisfied, researchers should show in the model or text whether a parameter is fixed or not. Given the large number of variables involved in an SEM study, it can quickly become difficult for a reader to identify free and fixed parameters based solely on model descriptions. To mitigate this problem, we recommend using different legends for free and fixed parameters, or using a legend for one and not for the other, in a path diagram. For example, Figure 1 shows that the loading from the oral ability factor to accuracy was fixed at one for identification, with the other two loadings from that factor freely estimated, as indicated by asterisks. If multiple models are considered while only one model is provided in a path diagram, the degrees of freedom of all considered models should be made clear by describing their deviations from the model in the path diagram. When models are highly complex, we concur with Boomsma (2000) that supplemental materials to provide this information be used. In any case, enough detail of each model should be made available so it is possible for readers to verify the reported degrees of freedom by counting the number of free parameters and unique data points.

Data

For a defensible SEM solution, it is crucial that data be appropriately prepared, and for a reader to judge the findings of an SEM study, it is necessary for the writer to provide details of this data preparation. Given that In'nami and Koizumi (2011) found that failure to report on the appropriateness of data for SEM analyses was a major cause for concern, we devote a fair amount of space to this topic. It is essential, that the characteristics of the sample data be

adequately described in an SEM report. For language assessment research, this would include relevant information about test takers, including gender, language background, and other variables of interest in the analysis. Sample size, descriptive statistics, and reliability estimates should be reported for each of the observed variables. We emphasize that a correlation or covariance matrix of all observed variables, which was used for the actual analysis, be provided, so consumers of the research can attempt to replicate the findings.

Large enough sample sizes and the absence of outliers are important for an acceptable SEM analysis. A typical application of SEM involves a number of parameters. In addition, standard errors and test statistics for SEM are obtained assuming a very large sample size. Therefore, sample size is an important factor that determines the quality of an SEM study. There are several rules of thumb about the minimum required sample size. For example, a common recommendation is to have at least ten test takers per estimated model parameter (Raykov & Marcoulides, 2006; Ullman, 2001). However, MacCallum, Widaman, Zhang, and Hong (1999) show that such global recommendations of sample size based on the number of factors or parameters are not necessarily valid, and other model elements such as the level of communalities can play an important role in determining an appropriate sample size. Given the lack of agreement and clarity on what constitutes appropriate sample size in SEM research, it is crucial that researchers carefully consider various aspects of hypothesized models and provide justification of their sample size for each analysis by indicating the rule of thumb (or other recommendation) that they are following, such as 10 test takers per estimated model parameter. Maybe a more defensible option to justify one's sample size is to conduct a power analysis, following the procedures proposed by Satorra and Saris (1985) or MacCallum, Browne, and Sugawara (1996). When sample size does not meet the common rule of thumb or the result of a

power analysis is not satisfactory, the researcher should caution the reader to consider the small sample size when interpreting the data.

SEM parameters are often estimated assuming that data are from an underlying multivariate normal distribution. If data are multivariate normal, individual variables are normally distributed, the joint distribution of all pairs of variables is bivariate normal, and each bivariate distribution shows a linear trend. While univariate normality does not guarantee multivariate normality, it is helpful to inspect the distribution of each variable to pinpoint variables that exhibit large deviations from a normal distribution. Therefore, we recommend univariate normality be reported based on an inspection of Q-Q plots or histograms and skewness and kurtosis values. Reports should include these values and the criteria that the researcher deems appropriate for meeting this data requirement. For instance, researchers could indicate that the skewness and kurtosis values are all within the guidelines set by Kline (2011): no skewness values exceed the absolute value of three and no kurtosis values exceed the absolute value of ten. The assumption of bivariate linearity is often inspected with scatter plots and therefore should be reported based on a brief description of the plots. We recommend researchers report Mardia's (1970) normalized coefficient as an indicator of the extent to which the data are multivariate kurtotic because it is commonly used by language assessment researchers, easy to interpret, and designed to closely follow the standard normal distribution in a large sample.

Because this assumption is commonly violated (a normalized Mardia value which exceeds three: Bentler, 2008), it is not unusual for data to be remedied prior to an SEM analysis. Researchers often remove a few outliers when a small portion of the data set substantially impacts its properties, while variable transformation is common when excessive univariate kurtosis or skewness is identified. When data contain a number of outliers and/or deviate much

from multivariate normality, more complicated procedures such as differential weighting of each data point might be employed. When any of these remedies are used, justification for the particular technique selected, how use of the technique is likely to impact the results, and details about how it was used should be reported.

Another important concern in many language assessment studies is missing data, and as In'nami and Koizumi (2011) found, few SEM research studies report the amount of missing data or ways in which missing data are managed. Given this concern, we briefly mention approaches to dealing with missing data and provide an indication of the importance of reporting what researchers have done when data are missing.[iii] The simplest way of handling missing data is listwise deletion, but this technique has the disadvantage of ignoring a portion of available data, which can have an impact on resulting parameter estimates unless the underlying missing data mechanism is missing completely at random (MCAR). Pairwise deletion preserves more data than listwise deletion, but can affect standard error estimates. When missing at random (MAR) can safely be assumed, full-information maximum likelihood (FIML) estimation is appropriate. Another defensible option is multiple imputation of missing data. This involves creating multiple "complete" datasets, each of which contains imputed missing values under a missing data model. Each such dataset is then analyzed as if it were a complete dataset, and their results are combined, with adjustments to account for the missing data, to make inferences.[iv] Different ways of handling missing data will lead to different results. Therefore, researchers should report which of these approaches (or another) they used and justify how they handle missing data. If no data are missing, researchers should also make it clear that they had a complete data set for the analysis.

The construction of measurement models implies close relationships among the indicators of a latent variable. However, severe multicollinearity may cause problems in

parameter estimation, and therefore theoretically and/or empirically redundant variables should not be included in a model. For this reason, we recommend that researchers report the variance inflation factor (VIF), the ratio of the total standardized variance divided by the unique variance. When multicollinearity is severe (for example, VIF is greater than 10: Kline, 2011), researchers should report the steps they take to mitigate this problem, such as excluding a variable or combining two or more variables.

In addition, any software error or warning messages indicating the deviation of data from the model assumptions should be reported. When data are remedied, the descriptive statistics of both original and remedied variables should be provided along with justification for remedying the data and the specific procedures that were used. Transformed variables, if any, should be described and interpreted with the transformation in mind.

Parameter estimation

To ensure valid interpretations of SEM results, reports must include details about the parameter estimation methods. Estimation is a process of finding the best solution for every free parameter based on a given model specification. Normal theory-based maximum likelihood (ML) is the most commonly used estimation method among language assessment researchers (and is the reason to ensure multivariate normality, as described in the previous section). Many methods make strong distributional assumptions, including ML and generalized least squares (GLS) estimation, which can also be found in language assessment research studies.

In the previous section, we underscored the importance of reporting the ways in which non-normal data are remedied prior to parameter estimation. There are also a number of approaches to address non-normal data by making adjustments at the estimation stage, such as

distribution-free estimation methods that are appropriate when large samples are available (Browne, 1984), and the approach commonly used in language assessment research, the Satorra-Bentler (1988; 1994) corrected normal theory method. In this latter approach, the original data are analyzed with a normal theory method with adjustments to both standard errors and test statistics for model fit. The adjusted test statistic is often called the Satorra-Bentler chi-square statistic.

Different estimation methods in general yield different parameter estimates and test statistics. Moreover, commonly used SEM software packages in the language assessment field, such as Amos (Arbuckle, 2006), EQS (Bentler, 2008), LISREL (Jöreskog & Sörbom, 2006), and Mplus (Muthén & Muthén, 1998-2012), have different default settings for different types of analyses and results output. Therefore, the methodology section of an SEM study should clearly state which estimation method and software package was employed in estimating the model parameters. The use of a method should be justified based on the description of distributional characteristics of the data, sample size, and other relevant factors. When multiple sets of estimation methods are used, the agreement between the resulting estimates from different initial values or estimation methods should be reported.

Model fit

Reporting model fit evaluation

Reports must include appropriate model fit indices to be validly interpreted. The model must be shown to reasonably fit the data before claims can be made about the relationships among the variables. Indices of model fit indicate the degree to which the proposed model can reasonably account for the observed data. Various fit indices exist, and each provides a

somewhat different estimation of the extent to which the model can reasonably explain the relationships among the data. Therefore, it is important that researchers report appropriate fit indices, rather than selecting ones that best support their model. Given the findings of In'nami and Koizumi (2011), which suggest lack of agreement in which indices to report, we provide some explanation of the types of indices along with justification for which ones should be reported.

Three types of indices have been used to investigate model fit: absolute fit indices, adjusted for parsimony indices, and relative fit indices.[v] Absolute indices assess the extent to which the model-implied and actual data variance-covariance matrices are plausibly the same. In other words, these indices provide an indication of the extent to which the observed data matrix and the hypothesized model data matrix are the same. The less similar they are, the worse the fit. Examples of absolute fit indices include the chi-square statistic, the standardized root mean square residual (SRMR; Bentler, 1995), and the goodness of fit index (GFI; Jöreskog & Sörbom, 1981). Adjusted for parsimony indices differ from absolute fit indices in that they penalize models with a larger number of free parameters. For example, other things being equal, the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) increases (smaller RMSEA is preferred) as the number of free parameter increases, favoring more parsimonious models. Another example of adjusted for parsimony indices includes the adjusted goodness of fit index (AGFI; Jöreskog & Sörbom, 1981). Relative fit indices compare the improvement in fit of the proposed model to a baseline model in which covariances among all variables are hypothesized to be zero. The more the improvement of the proposed model as compared to the baseline model, the better the fit. Relative fit indices can be used to compare the fit of a proposed model to the observed data or to compare the relative fit of two competing models to the observed data. The

comparative fit index (CFI; Bentler, 1990), normed-fit index (NFI; Bentler & Bonett, 1980), incremental fit index (IFI; Bollen, 1989a) and Tucker-Lewis index (Tucker & Lewis, 1973) are examples of relative fit indices. Kline (2011) provides a non-technical discussion of fit indices.

Based on the recommendations of Bentler (2008), Brown (2006), In'nami and Koizumi (2011) and Kline (2011), and our observation of the indices most commonly used in by language assessment researchers, we recommend that two absolute fit indices, the chi-square statistic and the SRMR; one adjusted for parsimony index, preferably the RMSEA and its confidence interval; and one relative fit index, preferably the CFI (or Tucker-Lewis index) be reported. When other fit indices are reported, a clear justification for their selection should be provided.

The chi-square statistic is interpreted using the chi-square distribution, and therefore should be accompanied by degrees of freedom and p-values. As discussed in the previous section, the Satorra-Bentler chi-square statistic (Satorra & Bentler, 1988, 1994) is commonly used to evaluate model fit by language testing researchers when the multivariate normality of data assumption is not tenable. When this statistic is reported, it should be noted along with an explanation that it is interpreted in the same manner as the chi-square statistic.

Because fit indices are differentially affected by a host of factors associated with a particular analysis including sample size, model complexity, estimation method, amount and type of misspecification, normality of data, and type of data (Brown, 2006), there are no agreed upon guidelines for what constitutes acceptable model fit. However, a few researchers have proposed guidelines that have some empirical support based on simulations, such as Hu and Bentler (1999). For good fit, they suggest that SRMR values should be close to .08 or below, RMSEA values should be close to .06 or below, and CFI/Tucker-Lewis index values should be

close to .95 or above. When these values are met, it may not be necessary for researchers to provide further statistical justification for their model fit. In cases in which a researcher feels that slightly lesser fitting models are defensible, the report should include appropriate justification for different cut-off values. Reports should also include a reminder of an acceptable value for each fit index.

Reporting data-driven specification searches

Detailed reporting of the procedures used when data-driven specification searches are employed is crucial to interpreting the results of research that employ such techniques. When a proposed model does not fit the data, or some parts of a model appear unnecessarily complicated, some SEM users move from confirmatory mode to exploratory mode. We want to be clear that we are not advocates of such an exploratory approach, but we do feel it is important to provide guidance for how to report use of the approach given that it is commonly encountered in language assessment research. An exploratory approach uses data-driven specification searches (Long, 1983), involving the Lagrange multiplier (LM) test used to identify the contribution of relationships between variables that are not included in a given model to the overall model fit and/or the Wald (W) test used to identify modeled relationships that are not supported by the data. Because the risk of capitalization on chance is inherent in any data-driven specification search, and simulation studies have shown that specification searches relying on statistical means would lead to highly unjustifiable results (e.g., MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992), we strongly recommend that when exploratory model building is conducted, clear reporting of the steps taken and justification for these steps be provided. It should be made clear in the report that the originally proposed model was changed because it did not fit the data satisfactorily. A description of the step-by-step procedures used to modify the model, such as the

order of conducting the LM and/or W tests, is vital to evaluate the credibility of the final model, and therefore, should be provided in detail. Substantive theory, which justifies the data-driven modifications, should be included in the report, and changes to the model that are not consistent with previous research findings should be accompanied by especially strong rationale. When results are based on models that have used specification searches, it is essential that reports make it clear that the findings are reliant on data-driven specification searches and should be replicated with other data before any claims are made on the basis of the results of the study.

Reporting the comparison of nested models

When a model can be obtained by imposing one or more constraints on the parameters of another model, the model with the parameter constraints is nested within the model without the constraints. It is straightforward to compare the fit of two models when one model is nested within the other model because the difference between the chi-square statistics of two nested models follows the chi-square distribution with the degrees of freedom equal to the difference between the two model degrees of freedom. This procedure is often called the chi-square difference test. A series of nested models can be compared using the chi-square difference test for each paired comparison. Research which includes nested model comparisons should report the chi-square statistics with degrees of freedom and p-values of all compared models, along with the chi-square and degrees of freedom differences and p-values for each comparison.

When the Satorra-Bentler (1988, 1994) chi-square statistic is used in evaluating the fit of each of compared models, the difference between two Satorra-Bentler chi-square statistics does not follow the chi-square distribution, and therefore an additional procedure is needed to conduct the chi-square difference test properly (Satorra & Bentler, 2001, 2010). Along with reporting the

statistics needed to report a chi-square difference test, researchers should report the use of these correction procedures to make it clear that the appropriate technique has been used.

Reporting the comparison of non-nested models

Research which compares the relative fit of two non-nested models, to determine which model most closely represents the data, generally uses information indices. The most commonly encountered indices for comparing the fit of two non-nested models are the AIC (Akaike, 1987) and its consistent version (CAIC; Bozdogan, 1987). Less known to the field of language testing is the Bayesian information criteria (BIC; Raftery, 1995; Schwarz, 1978). We recommend that either AIC or CAIC be reported. BIC may also be reported in addition to one of the others. In addition, each of the compared models should be accompanied by the package of fit indices recommended above when reporting model fit.

Reporting the test of multi-group invariance

Language assessment research commonly investigates the extent to which a set of items assesses the same construct in different groups. Multiple-sample confirmatory factor analysis, which compares the fit of two nested models, has been used for this purpose (Llosa, 2005; Purpura, 1998; Shin, 2005). Fit comparison for determining multi-group invariance generally employs two types of fit indices, the model chi-square difference test and a comparative fit index such as the CFI. The former provides a strict significance test, whereas the latter provides an indication of the practical importance of the difference. Chi-square values along with the difference between the values should be accompanied by p-values and degrees of freedom for the chi-square difference test, and CFI values for each model accompanied by differences in CFI values for a the practical importance comparison, should be reported. When researchers choose

to use values other than the traditional .05 level for indicating a significant difference in models for the Chi-square difference test and a change in CFI of .01 or greater for indicating practical importance (Cheung & Rensvold, 2002), they should provide justification for these alternative criteria.

Model interpretation

To be able to validly interpret a model, a number of estimates need to be reported. Standard errors should be reported along with an indication of the extent to which the results can be trusted. When some standard errors are disproportionately larger than others, reports should caution against accepting these estimates as meaningful. In addition, improper solutions, such as ones with Heywood cases, parameter estimates with negative variances, should not be interpreted, and their presence should be clearly stated in the report.

We recommend reporting both standardized and unstandardized estimates when space allows. Standardized estimates, by virtue of enforcing the same variance to all parameter estimates, are usually easier to interpret. However, because deriving standard errors for the standardized estimates is not straightforward, meaning unstandardized estimates are needed to make statistical inferences, we recommend that both standardized and unstandardized estimates be reported. We agree with McDonald and Ho (2002) that the best practice is to report one on the path diagram of the final model while the other is summarized in a table. In most cases, we recommend having standardized estimates on a path diagram since both the path diagram and standardized estimates are designed to facilitate the interpretation of a model. Statistical test results have commonly been reported using a table, and therefore unstandardized estimates and the corresponding standard errors can be reported in the same manner.

While factor loadings and structural regression coefficients are often the parameters of the most interest, all free parameter estimates should be reported. When multiple correlated latent variables are present, their covariance matrix should be reported and interpreted. Residual variances of endogenous variables should also be reported since they are important to understand the quality of a model. When models involve a large number of variables, we recommend that a table of residual variances as an appendix or as supplemental material be provided. We recommend reporting the results of a large-sample approximate z-test, which can be obtained in most commonly used SEM software packages, based on unstandardized estimates and the corresponding standard errors to make it possible to gauge the significance of a parameter.

We recommend that both factor loadings and structural coefficients are reported. When latent variables are correlated, the lack of a direct path from a latent variable to an observed variable does not necessarily mean that they are not related (Bentler & Yuan, 2000). Their relationship is indicated by the correlation between latent variables. The direct relationship between a latent variable and its indicator is reflected in the corresponding factor loading, while the indirect relationship between a latent variable and an indicator of another factor is represented as the corresponding structure coefficient. Because ignoring structure coefficients can lead to interpretation errors, especially when cross-loadings are present (Graham, Guthrie, & Thompson, 2003), we recommend that both factor loadings and structure coefficients be reported and discussed.

Guidelines which determine whether or not two correlated factors should be considered distinct factors are important in many language assessment research studies. It is, therefore, important that the guidelines that one uses for this purpose be provided and justified. One way to test the factor distinctness is to rely on a nested model comparison using the chi-square

difference test. General recommendations about what constitutes distinct factors are available, such as correlations above .85 indicate the variables are too similar to be considered distinct (Brown, 2006). However, because such rules of thumb should be considered in light of substantive theories and previous findings and that there is a lack of agreement on what constitutes a correlation that can no longer be considered distinct, we recommend that the specific criteria used to judge distinctness be provided along with a defensible argument for the criteria.

Alternative models

Reports should recognize possible alternative models that have not been ruled out either statistically or theoretically. In SEM research, there can be multiple models with different structures that are mathematically equivalent. Thus, although a researcher can provide convincing support for a model, there may be other models that are equally defensible. One of the most well-known examples of such equivalent models, commonly encountered in language assessment research, involves a second-order factor model with three first-order factors and a correlated three-factor model. Other things being equal, the two models are mathematically equivalent, and therefore yield the same fit indices. When such equivalent models exist, reports should at a minimum mention these equivalent models and indicate that they were not investigated and may also provide plausible explanations for the data. We therefore recommend that researchers recognize alternative models in their reports and provide theoretical justification for not investigating them. When alternative models can conceivably provide substantively meaningful interpretations, they should be discussed as reasonable alternatives.

Conclusion

SEM is a technique with great potential for aiding the language assessment community. However, research which does not conform to acceptable reporting guidelines can be at best difficult to follow and at worst misleading to the field. We recognize that space limitations imposed by publishing venues may not always allow as much detail about an analysis as we suggest. In these situations, we recommend that consideration be given to providing some of the details about the analysis at an online location where readers can be directed. When publishing companies cannot provide a site for such information, it may be possible for researchers to include such details on their own web pages.

Given that SEM is a developing field in which new ways of conducting and reporting analyses are emerging, it will undoubtedly become necessary to revisit these guidelines from time to time. Therefore, we recommend that these guidelines be reviewed and revised as the field progresses. We also note that it is not reasonable to judge SEM research conducted prior to the publication of this paper by the guidelines provided here. We end with a plea to the field of language testing to take steps to adopt these guidelines for best reporting practices in SEM research. This will make it easier for readers, writers, and reviewers of manuscripts which employ an SEM approach, and it will help to ensure that SEM research leads to better understanding of issues relevant to the field of language assessment.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*(3), 317-332.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 545-557.

Arbuckle, J. L. (2006). Amos (Version 7.0) [Computer Program]. Chicago: SPSS.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M. (2008). *EQS program manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 586-606.

Bentler, P. M., & Yuan, K. (2000). On adding a mean structure to a covariance structure model. *Educational and Psychological Measurement, 60*(3), 326-339.

Bollen, K. A. (1989a). A new incremental fit index for general structural equation models. *Sociological Methods and Research, 17*(3), 303-316.

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

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

Bozdogan, H. (1987). Model selection and Akaike's information criteria (AIC). The general theory and its analytical extensions. *Psychometrika, 52*(3), 345-370.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research.* New York: Guilford.

Browne, M. W. (1984). Asymptotic distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*(1), 62-83.

Carr, N. T. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing, 23*(2), 1-21.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255.

Graham, J., Guthrie, A., & Thompson, B. (2003). Consequences of not interpreting structure coefficients in published CFA research: A reminder. *Structural Equation Modeling, 10*(1), 142-153.

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.

Hoyle, R. H., & Isherwood, J. C. (2013). Reporting results from structural equation modeling analyses in Archives of Scientific Psychology. *Archives of Scientific Psychology, 1*(1), 14-22.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage Publications.

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. (1981). Listrel V: Analysis of linear structural relationships by the method of maximum likelihood. Chicago: National Education Resources.

Jöreskog. K. G., & Sörbom. D. (2006). Listrel 8.8 for Windows [Computer Software]. Skokie, IL: Scientific Software International, Inc.

Kline, R. (2011). *Principles and practice of structural equation modeling, (2nd ed.)*. New York: The Guilford Press.

Kunnan, A. J. (1994). Modeling relationships among some test-taker characteristics and tests of English as a Foreign Language. *Language Testing, 11*(3), 225-252.

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

Lee, T., & Cai, L. (2012). Alternative multiple imputation inference for mean and covariance structure modeling. *Journal of Educational and Behavioral Statistics, 37*(6), 675-702.

Llosa, L. (2005). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency. Unpublished PhD dissertation, University of California, Los Angeles.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing*, *24*(4), 489–515.

Long, J. S. (1983). *Covariance structure models: An introduction to LISREL*. Beverly Hills, CA: Sage.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100*(1), 107-120.

Maccallum, R. C., Browne, M. W., & Sugawara, H. M. (1996), Power analysis and determination of sample size for covariance structure modeling, *Psychological Methods 1*(2), 130–149.

MacCallum, R. C., Roznowski, M., &Necowitz, L. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.

MacCallum, R. C., Widaman, K. F., Zhang, S., Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530.

McDonald, R., & Ho, M. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64-82.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Ockey, G. J. (2011). Assertiveness and self-consciousness as explanatory variables of L2 oral ability: A latent variable approach. *Language Learning, 61*(3), 968-989.

Ockey, G. J. (2013). Exploratory factor analysis and structural equation modeling. In A. J. Kunnan, *The companion to language assessment*.

Phakiti, A. (2008). Strategic competence as a fourth order factor model: A structural equation modeling approach. *Language Assessment Quarterly, 5*(1), 20-42.

Purpura, J. E. (1997). An analysis of the relationships between test takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning, 47*(2), pp. 289-294.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: a structural equation modelling approach. *Language Testing, 15*(3), 333-379.

Purpura, J. (1999). Strategy use and second language test performance: A structural equation modeling approach. Cambridge: Cambridge University Press.

Purpura, J. E. (2004). Validating Questionnaires to examine personal factors in L2 test performance. In M. Milanovich & C. Weir (Eds.), Eur*opean Language Testing in a*

Global Context. Proceedings of the Association of Language Testers of Europe (ALTE) Conference in Barcelona (pp. 93-115). Cambridge: Cambridge University Press.

Raftery, A. E. (1995). Bayesian model selection in social research. In A. E. Raferty (Ed.), *Sociological Methodology* (pp. 111-164). Oxford: UK: Blackwell.

Raykov, T., & Marcoulides, G. (2006). A first course in structural equation modeling, Second edition. Mawah, New Jersey: Lawrence Erlbaum Associates.

Raykov, T., Tomer, A., & Nesselroade, J. R. (1991). Reporting structural equation modeling results in psychology and aging: Some proposed guidelines. *Psychology and Aging, 6*(4), 499-503.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Satorra, A., & Bentler, P. M. (1988). Scaling correction for chi-square statistics in covariance structure analysis. *Proceedings of the American Statistical Association*, 308-313.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507-514.

Satorra, A, & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference Chi-square test. *Psychometrika, 75*(2), 243-248.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*(1), 83-90.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, *24*(3), 355–390.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*(1), 5–30.

Schumacker, R., & Lomax, R. (2010). A beginner's guide to structural equation modeling, third edition. Mawah, New Jersey: Lawrence Erlbaum Associates.

Shin, S. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing*, *22*(1), 31–57.

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

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461-464.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1-10.

Ullman, J. (2001). Structural equation modeling. In B. Tabachnick & L. Fidell, *Using multivariate statistics, (4th ed.) (pp. 653-771).* Boston: Allyn and Bacon.

[i] It is not clear how many of these articles were found in the language assessment literature since the researchers did not distinguish between language learning and language assessment. Moreover, the researchers did not include books in this review, and it is in books that we would expect the space for detailed information about the procedures used.

[ii] These are informal and heavily simplified conditions. Bollen (1989b) and Kline (2011) provide a comprehensive yet accessible discussion of model identification in SEM.

[iii] We suggest interested readers consult Allison (2003) who provides a clear and accessible review of missing data techniques for SEM and Rubin (1976) for the definition of missing data mechanism.

[iv] Although multiple imputation is a general strategy that is applicable to a number of analysis situations, its use in SEM contexts is not without complications and is still an active area of methodological research (Allison, 2003; Lee & Cai, 2012).

[v] This classification system is not consistent in the SEM literature. For instance, sometimes absolute and corrected for parsimony indices are collapsed into one category. Moreover, the three approaches are not always exclusive.