Tabaran Institute of Higher Education ISSN 2476-5880
International Journal of Language Testing
Special Issue, 2023: *Advanced Psychometric Methods in Language Testing*

Structural Equation Modeling in L2 Research: A Systematic Review

Hessameddin Ghanbar[1] & Reza Rezvani[2]*

Abstract
Structural equation modeling (SEM), as a flexible and versatile multivariate statistical technique, has been growingly used since its introduction in the 1970s. This article presents a methodological synthesis of the characteristics of the use of SEM in L2 research by examining the reporting practices in light of the current SEM literature to eventually provide some empirically grounded recommendations for future research. A total of 722 instances of SEM found in 145 empirical reports published in 16 leading L2 journals across two periods of 1981-2008 and 2009-2020 were systematically reviewed. Each study was coded for a wide range of analytic and reporting practices. The results indicate that despite the growing popularity of SEM in L2 research, there was a wide variation and inconsistency in its uses and reports within and across the two periods in regard to the underlying assumptions, variables and models, model specification and estimation, and fit statistics. Drawing on the current SEM literature, we will discuss the findings and research implications for future use and reporting of SEM in L2 research.

Keywords: L2 journals, L2 research, Multivariate data analysis, Structural equation modeling; Systematic review

1. Introduction

Structural equation modeling (SEM) is a family of multivariate statistical techniques that seeks to explore and explain correlations and covariances of variables. This highly flexible procedure can be employed when working with a wide variety of variables whether observed (IQ score, GPA, proficiency tests' scores) or latent (motivation, anxiety, success, self-esteem), categorical or continuous (Tabachnick & Fidell, 2013). The flexibility inherent in SEM has contributed to its vast use in recent years in a variety of domains including second-language (L2) research (Khany & Tazik, 2019; Winke, 2013). Hence, it is growingly becoming common to see articles reporting SEM results in all leading journals of L2 research.

On the other hand, many L2 researchers are still unfamiliar with this versatile yet complicated procedure, relative to other, more frequently practiced analyses such as ANOVA or correlation (Loewen et al., 2014; Plonsky, 2013). This is also compounded by the numerous analogous terminologies and monikers used for SEM in the literature such a causal analysis, causal modeling, simultaneous equation modeling, analysis of covariance structure, path analysis, and confirmatory factor analysis.

[1] Department of Language and Linguistics, Islamic Azad University, Fereshtegaan International Branch, Tehran, Iran; +98 9380743500, hessam.ghanbar@gmail.com
[2]* Department of English Language, Humanities, Yasouj University, Yasouj, Iran; +98 09177038620, rezvanireza@gmail.com

Given the defying complication of SEM and variation in using and reporting it, the present study aims to present a systematic review of its current use in L2 research. It is ultimately intended to provide empirically grounded recommendations of more legitimate and consistent use of SEM in future L2 research through building on the growing methodological syntheses (e.g., Derrick, 2016; In'nami & Koizumi, 2011; Plonsky & Ghanbar, 2018). In what follows, a brief introduction is provided for the general readership. For more technical details of SEM, interested readers are referred to SEM guidebooks and relevant software manuals (for a review of SEM software programs see Narayanan, 2012; and for a review of SEM books see Verkuilen, 2011).

2. Introduction to SEM

As research questions become more complex addressing more sophisticated and multifaceted research problems (see Ghanbar & Rezvani, 2023) in social sciences, more robust and vigorous psychometric methods and statistical tools have also been developed. Parallel to the advancement in computers and soon after the Bentlers' (1980) call to use SEM to deal with latent variables, SEM has become a routine statistical analysis approach in a wide range of disciplines (Baumgartner & Homburg, 1996). The main asset of SEM lies in its ability to perform multiple analyses like multiple regression and factor analysis simultaneously to address various questions (See Tabachnick & Fidel, 2013 for a list of research questions or scenarios that can be answered by SEM). In exploring theoretical specifications and validating models it estimates manifest variables along with associated measurement errors, and also, it tests multiple direct and indirect causal relationships among latent and indicator (manifest) variables in postulated models (Kline, 2016).

In order to address diverse research questions, SEM has been used in various ways. The way it is usually utilized can be taken more simply to involve two general steps or more elaborately several specific steps. These two views of how SEM is used are outlined in the following. An attempt is also made to give an overview of the common SEM concepts and terminologies. The paper will then proceed to some critical issues in conducting SEM and reporting the results.

Regarding stages for conducting a SEM study, in a typical two-step procedure, SEM application consists of a *measurement* or *factor* model and a *structural* model (Anderson & Gerbing, 1988; Barati, et al., 2013; James et al., 1982). The measurement model involves confirmatory factor analysis (CFA) that examines the relationships between a construct (depicted by ovals) and indicators or manifest variables (represented by rectangles), as shown in Figures 1 and 2. The specification of the relationships among manifest and latent variables should be informed by the relevant literature or a theory as sketched in the two figures (see Bollen & Bauldry, 2011 and aslo Diamantopoulos & Winklhofer, 2001). In Figure 1, in a standard CFA, the underlying factors cause the *effect* or *reflective* indicators. The same variables might be hypothesized to have a reverse relationship by postulating that the factors are affected by the cause or formative indicators (Figure 2 as a formative model). As depicted in Figure 1, in a reflective model, for each measured variable there is an error term labeled "e" which is viewed as variance in observed variables that is not explained by the factor or construct.

Figure 1

Path Diagram of a Reflective Model in SEM

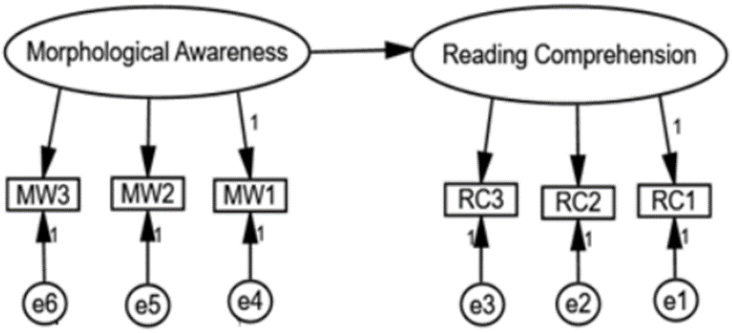
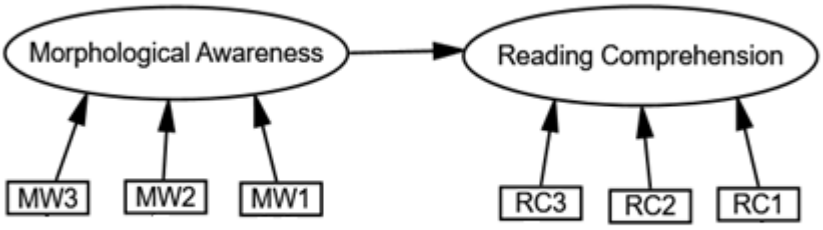


Figure 2

Path Diagram of a Formative Model in SEM



Testing the structural model depends logically on how well the hypothesized model fits the data. Unlike many statistical approaches like ANOVA determining model fit and interpretation is complicated and does not rely on a single powerful index like *F statistic* (Schumacker & Lomax, 2016). There is an increasing number of model fit indexes such as the comparative fit index (CFI), chi-squared statistic, the adjusted goodness-of-fit index (AGFI), the root mean square error of approximation (RMSEA), and the root mean square residual index (RMR) (Kline, 2016) with specific cutoff criteria (for details see Hu & Bentler 1999). Therefore, it is recommended that at least two overall fit indexes be used and reported (Hu & Bentler 1999). The measurement model also provides a test of convergent and discriminant validity (Anderson & Gerbing, 1988) essential in model development and testing. The underlying rational is that variables presumed to measure the same construct show high intercorrelations supporting convergent validity, and those which are supposed to reflect different constructs do not show high intercorrelations as indicative of discriminant validity (Campbell & Fiske, 1959).

The second step in SEM analysis involves hypothesizing and testing a structural model or latent variable path analysis (LVPA) (Mueller & Hancock, 2019) manifesting the postulated causal relationships among the constructs as in the one-headed arrow between morphological awareness (latent *exogenous* variable, considered as an independent variable [IV] and reading comprehension (latent *endogenous* variable, viewed as a dependent variable [DV]) in Figure 1 and Figure 2. It should be mentioned that a unique feature of SEM is its ability to assess several multiple regression

equations and dependency paths simultaneously (Hoyle, 1995) rather than examining them separately as it is the case in other multivariate analyses.

Some SEM experts extended the two-step approach to procedures more sensitive to misspecification errors in measurement models (Kline, 2016). For example, Hoyle (2012) proposed a five-step model (for a four-step approach see Hayduk & Glaser, 2000; Mulaik & Millsap, 2000) comprising (a) specification, (b) estimation, (c) evaluation of fit, (d) respecification, and (e) interpretation and reporting. First, in the specification stage, all the hypothesized relationships among the variables are theoretically demarcated based on theoretical and empirical evidence, and are visually sketched using formal SEM notations (see Ho et al., 2012) for a discussion of SEM graphical representation). Important in this stage is the identification of the specified model. As Kenny and Millan (2012) emphasized, in this stage, researchers should check whether they have enough known information (measured variables) for estimating the unknown information (parameters) including variances, covariances, structural coefficients, and factor loadings.

Second, after specification and identification, the model is estimated in terms of calculating the parameters of the hypothesized model in order to minimize the discrepancy between the model-implied covariance matrix (the model) and the observed covariance matrix (population). A variety of estimation methods can be utilized in SEM model estimation. The default technique in most SEM software packages is the maximum likelihood (ML). Some alternatives are: Robust ML, full-information maximum likelihood (FIML), unweighted least squares (ULS), generalized least squares (GLS), weighted least squares (WLS), and asymptotically distribution-free (ADF). The choice of estimation methods depends on several statistical criteria such as sample size, outliers, and distribution of the variables (Schumacker & Lomax, 2016; Ullman, 2007).

The third step is evaluating the model fit to the data by examining different fit indices generated by SEM packages. This assessment of fit or discrepancy between the model and the data is a critical issue both in SEM literature and use. It is recommended that various goodness-of-fit indexes and residuals (discrepancy between model-implied and observed covariance matrices) (Sawaki, 2013), along with substantive interpretability (Kline, 2016) should be taken into account in model evaluation and modification as the next step. In the following fourth stage, the researchers aim to find the sources of misfit in the model and try to improve the model fit. Any model posited is taken, at best, to be an approximation of the reality, and statistical tests informed by substantive meaningfulness help researchers to exploratorily respecify the model until the most acceptable one is developed (Mueller & Hancock, 2008). The procedure ends, as a final step, in a report of various different statistics of the model and an interpretation of the results through elaborating on the substantive meaning of the paths and model tenability in relation to the observed data.

3. Considerations in SEM Application and Report

SEM, like any other statistical techniques, is predicated on a set of assumptions about data and the model posited and evaluated (see Jöreskog & Sörbom, 1996; Ullman, 2007). Yet, there has been much variation in the guidelines on how to use SEM and even more on how to report the results (Sawaki, 2013). This, at times, led to serious problems in SEM applications and reports (Kline, 2016). Since this systematic review is intended to capture the current SEM use to offer

useful guidelines on how to report SEM results, the most important considerations in this regard are briefly presented in what follows to contextualize the issues reviewed and synthesized. Researchers desiring even more advanced and technical details can refer to existing resources (e.g., Hancock & Mueller, 2006; McDonald & Ho, 2002; Meyers et al., 2013; Pituch & Stevens, 2016; Schoonen, 2015; Ullman, 2007)

The first consideration in SEM application concerns the sample size. SEM, as a covariance-based statistical technique is bound to relatively large sample sizes and small samples can risk the normality of the distribution. There is no generally agreed-upon rule for SEM data. There have been some rules of the thumb recommendations like Kline (2016), for example, suggesting that sample sizes below 100, between 100 and 200, and more than 200 are considered small, medium, and large, respectively.

Large data sets for SEM are likely to include some outliers and miss some data. Consequently, both cases affect SEM analysis and raise concerns for researchers. Because of the complications involved, there have been many resources on how to handle missing data in data sets (see for example Enders, 2010; Graham & Coffman, 2012; Peters & Enders, 2002). In general, larger sample sizes are recommended when there are missing data in the analysis in order to make up for the missing information. It is also recommended that SEM data be screened for the existence of both univariate and multivariate extremes (see Ho & Naugher, 2000, Nicklin & Plonsky, 2020) or outliers. As with outliers, there is no single definition for outliers but atypical data above two standard deviations can be viewed as outliers (Kline, 2016).

Regarding the underlying statistical assumptions, the first issue is the normality of the data. In the case of SEM, the tenability of both univariate and multivariate normality assumptions is necessary as most model estimation methods assume these two conditions (Ulman, 2007). Further, because of the important role of linear correlations among variables in SEM as an essentially correlational research method, linearity needs to be detected in SEM analysis (Schumacker & Lomax, 2016) through examining screeplots (Pituch & Stevens, 2016). By carefully examining linearity, researchers can check and confirm the absence of singularity (perfect correlation) and multicollinearity (correlation of 0.90 and above) (Tabachnick & Fidell, 2013).

4. The Present Study

The present study seeks to build on the growing body of methodological syntheses in L2 research (see Marsden & Plonsky, 2018). Studies of this nature apply synthetic or meta-analytic techniques, coding a set of research and reporting practices across a representative if not exhaustive sample in a given domain. Some methodological syntheses have sought to describe and evaluate research and reporting practices within a given substantive domain such as interaction (Plonsky & Gass, 2011) and written feedback (Liu & Brown, 2016). Others, more in line with the present study, have been concerned with individual research techniques such as multiple regression (Plonsky & Ghanbar, 2018) and qualitative data coding and analysis (Riazi et al., 2023). In'nami and Koizumi (2011) also conducted a methodological synthesis of the use of SEM in language testing and learning research that is comparable to the present study. They reviewed the characteristics of a total of 50 SEM articles in these two areas published from 1981 to 2008. The present study seeks to build on and update their review after more than a decade by systematically reviewing and evaluating SEM-based studies from 2009 to 2020. Ultimately, we also seek to make

some empirically-grounded recommendations for future SEM use and reports. In line with these objectives, the present study addresses the following two research questions:

1. What are the characteristics of the use of structural equation modeling in L2 research?
2. To what extent, and in what ways, did SEM-related research and reporting practices change over time, following the period under investigation in In'nami & Koizumi (2011) (i.e., 1981-2008)?

5. Method

5.1. Study Retrieval

Following Plonsky and Oswald's (2015, see also Chong & Plonsky, 2023; Riazi et al., 2023) guidelines, we first searched 16 top-tier L2 research journals (see Figure 5) assumed to be representative of empirical research publications in the field in order to arrive at a more transparent (see Plonsky & Oswald, 2015) and relatively representative, if not exhaustive, a sample of instances of SEM use in L2 research. We also searched multiple online databases including Academic Search Premier, Education Source, ERIC, EBSCO, Google, Google Scholar, JSTOR, IRIS, SSCI (Social Sciences Citation Index), and ScienceDirect, Cambridge, Oxford, Wiley, Taylor and Francis and SAGE databases to complement the first search results. The search was repeatedly conducted at different occasions to include all relevant studies with the last search conducted in February 2020.

The pool of empirical research we came across, comprised 10,223 articles. The corpus included abstracts, titles, keywords, body texts, acknowledgments, references, and supplementary materials. All these sections in the corpus were searched using AntConc (Anthony, 2019) for key phrases such as "structural equation modeling" and "_SEM_". The contexts of all hits were then examined, and the original PDFs were then collected whenever the articles were found to refer to the use of structural equation modeling in an empirical study. In total, 145 articles were found to employ SEM with 722 unique models.

5.2. Data coding and analysis

The next step in this synthetic study of SEM applications in L2 research was developing a coding scheme to tap the key issue in SEM. The coding scheme for this study was designed by drawing on (a) the different checklists and recommendations in statistical guides (Brown, 2015; Byrne, 2016; Hair, Hult, Ringle, & Sarstedt, 2017; Hoyle, 2012; Keith, 2019; Mulaik, 2010; Pituch & Stevens, 2016; Stevens, 2009; Tabachnick & Fidell, 2013), (b) general (non-SEM-specific) recommendations and findings presented in recently published reviews of L2 research (e.g., Larson-Hall & Plonsky, 2015; Ockey & Choi, 2015), and (c) previous methodological syntheses of closely related statistical techniques (e.g., Plonsky & Gonulal 2015; In'nami & Koizumi, 2011; Plonsky & Ghanbar, 2018).

After several rounds of development and reflections, the final coding scheme included several sections and items to elicit (1) article information, (2) model specification (model complexity), (3) measurement model-related issues, (4) structural model-related issues, (5) sample size related issues, (6) model estimation method, (7) measurement model estimation results, (8) structural model estimation results, (9) fit indices, and (10) software. The full version of the coding

scheme will be available upon publication on the IRIS database (Marsden et al., 2016) as well as the request.

To ensure consistency and accuracy of data coding, two rounds of interrater reliability checks were performed. Initially, a sample of 7 studies that included a total of 22 models were coded by both researchers resulting in relatively high inter-coder reliability at the scheme and item levels. Specifically, among a total of 1892 cells, there were only a total of 28 disagreements (1.48%) with no more than 1 disagreement on any individual item. The disagreements were resolved through discussion leading also to a number of refinements of the coding scheme. In the second round, an additional 10 randomly chosen studies (including 24 models) were re-coded. Agreement was perfect on all but 9 items (99.2%). Resolving this small discrepancy and refining the coding scheme, both researchers conducted the data coding. A number of discrepancies led to further refinements in the coding scheme as well as the code book.

Similar to other methodological syntheses (e.g., Plonsky & Gonulal, 2015; In'nami & Koizumi, 2011), the data analysis of this study was implemented using descriptive statistics. More specifically, the research question concerned with the use and reporting of SEM in L2 research, was addressed through frequencies and percentages of categorical variables such as type of endogenous and exogenous variables (DVs and IVs) and the examination of underlying statistical assumptions. Continuous variables such as model (measurement or structural) estimation results and sample size were summarized using means, standard deviations (SDs), or medians depending on their distributions. With the exception of the initial results describing the sample, models (not reports) constituted the unit of analysis. In other words, unless otherwise indicated, the percentages reported below refer to the portion of models that contained different techniques, reporting practices, and so forth. For example, In'nami (2006) reported on two models (see Figures 2 and 3 of their paper). When summarizing the initial results describing the sample (for Figures 3 and 4 of the current paper), we counted In'nami only once as a single data set. In all other cases, when reporting the use of different techniques (for all other tables and figures of the current paper), we focused on each model (not only the final model as in In'nami & Koizumi, 2011). The results are presented both overall to address RQ1 and, to address RQ2, across two time periods: 1981-2008 and 2009-2020.

6. Results

6.1. Distribution of SEM L2 studies

As a first step in understanding the use of SEM in L2 research, we present basic bibliographic findings concerning the studies that employed it. As regards Period 1 (see Figure 3) it is indicated that SEM use was on a steady rise in recent decades marking a peak in 2005, despite a slight decrease from that year to 2008.

For Period 2, as can be seen in Figure 4, there was also a rising trend in the use of SEM with its ups and downs hitting two peaks in 2013 and 2017.

Figure 3

Frequency Distribution of SEM L2 Studies in Period 1 by Year (51 Articles)

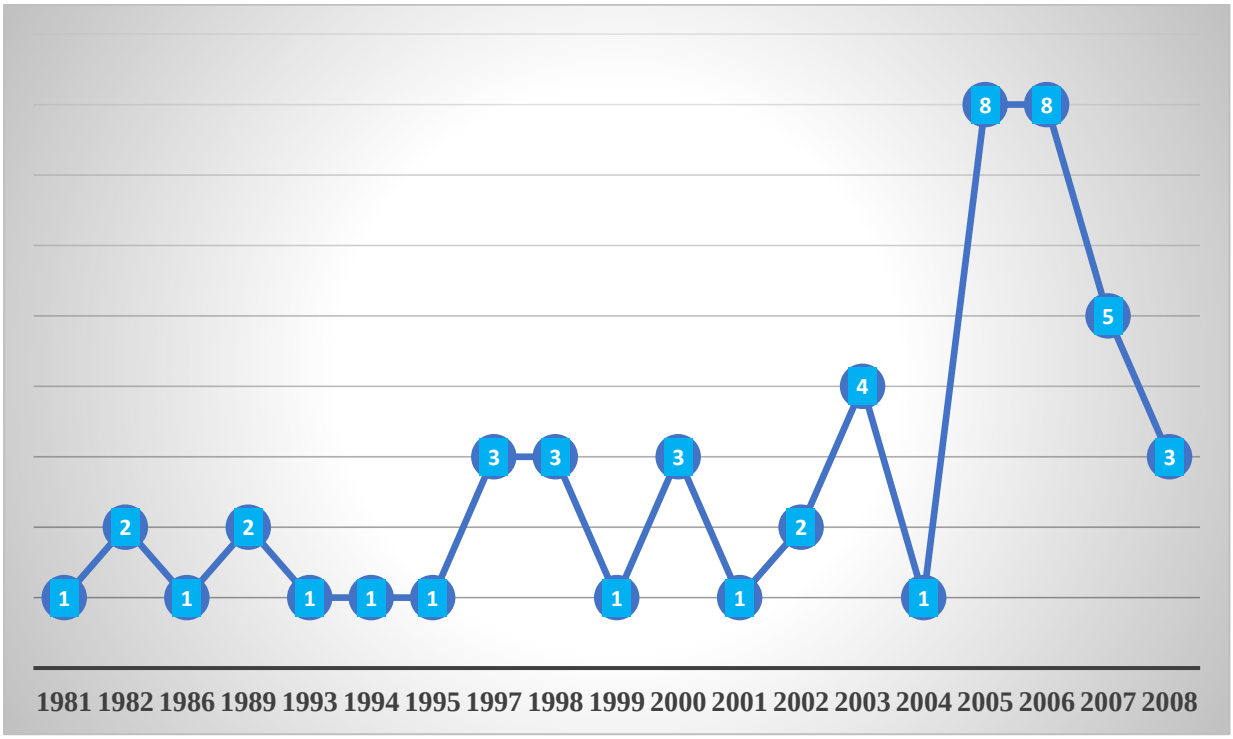


Figure 4

Frequency Distribution of SEM L2 Studies in Period 2 by Year (94 Articles)

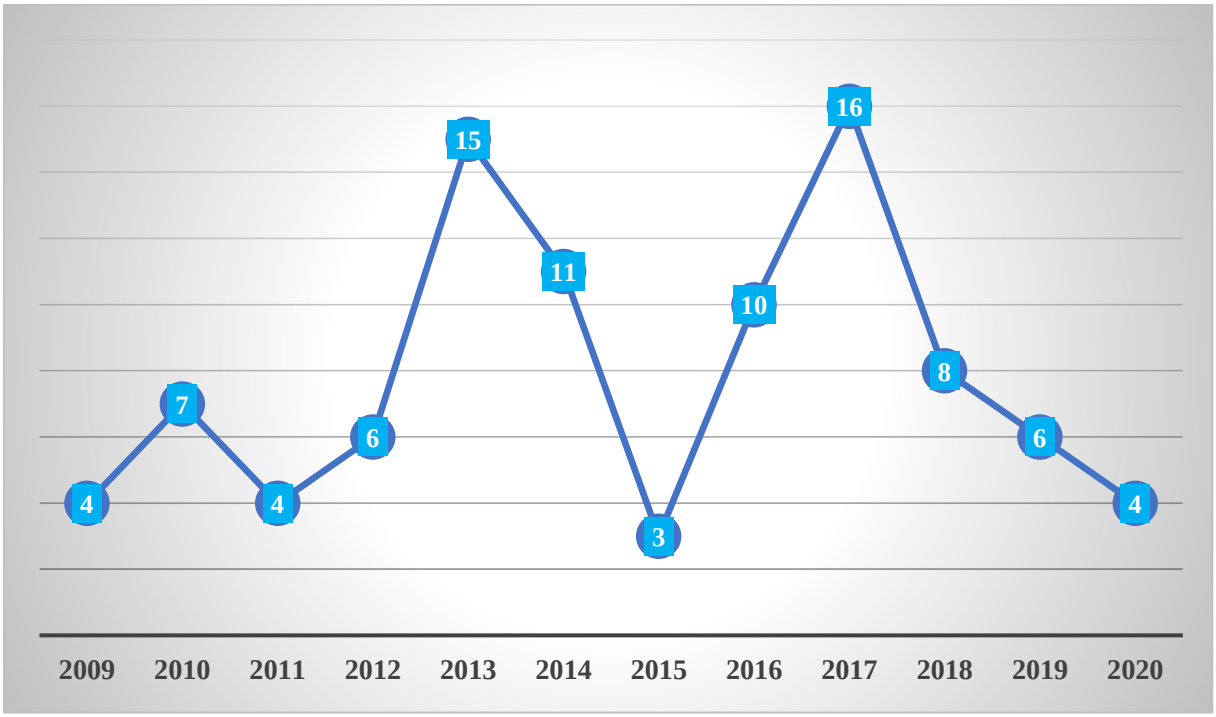
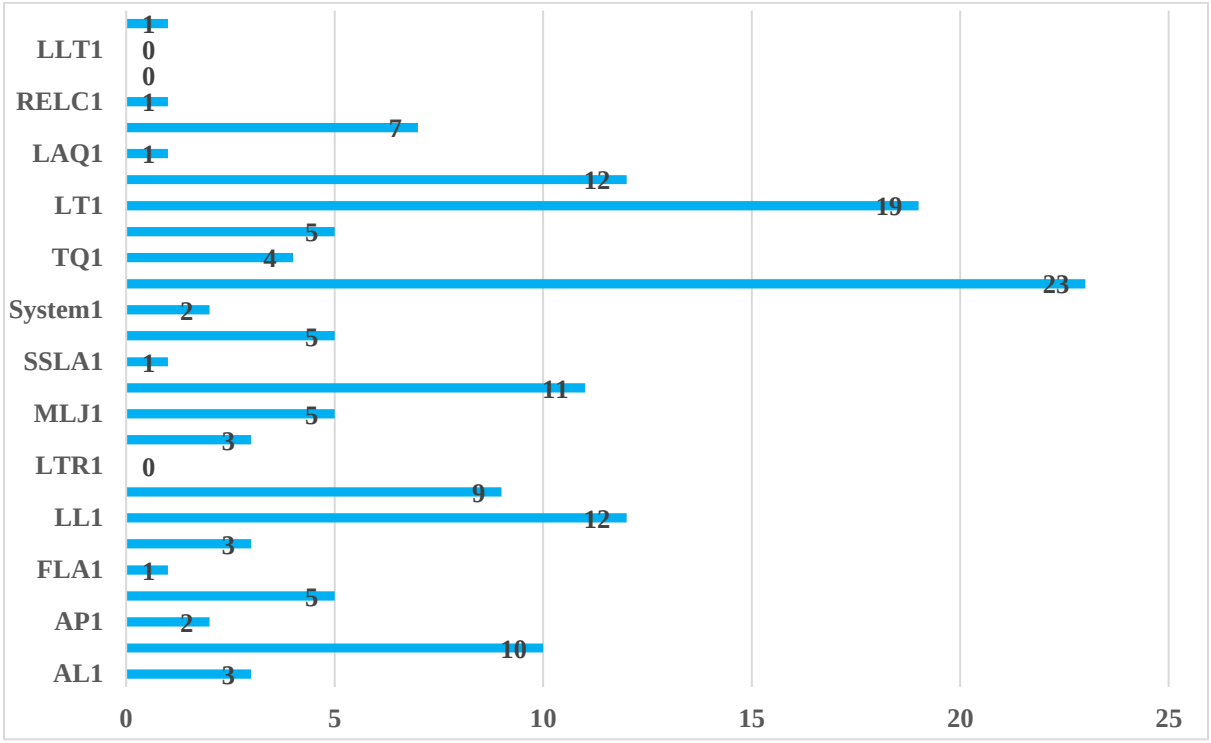


Figure 5 shows that studies employing SEM can be found in all applied linguistics journals ranging from more generalist (e.g., *Language Learning*) to more specialized journals (e.g., *Language Testing*). With the exception of *Language Learning*, *Language Testing and RELC*, it was also noted that articles involving SEM were on the rise in both periods.

Figure 5

Frequency Distribution of SEM L2 Studies by Journal across two Periods (51 Articles in Period 1 and 94 Articles in Period 2)



Note. LTR = *Language Teaching Research*; RELC = RELC Journal; SSLA = *Studies in Second Language Acquisition*; FLA = *Foreign Language Annals*; LAQ = *Language Assessment Quarterly*; AP = *Applied Psycholinguistics*; TQ = *TESOL Quarterly*; AL = *Applied Linguistics*; MLJ = *Modern Language Journal*; LL = *Language Learning*; LT = *Language Testing*; LLT = *Language Learning & Technology*

6.2. Statistical Assumptions and Issues

SEM use rests on a number of basic assumptions. As it is shown in Table 1, in Period 1, around half of the studies (53%) reported normality tests most of which univariate, 31% discussed missing data, 17% evaluated linearity, 25% identified outliers, and 11% assessed multicollinearity. Only about 6% of the studies in the sample for this period were reported to have checked for normality of residuals. The sample sizes ranged from 47 to 8,593 and on average (n = 182) it was adequate according to Kline (2016). A large share of the sampled studies (85%) provided a diagram to illustrate the relationships among variables. Concerning measured variables, 72% of the studies provided at least one reliability estimates.

Table 1
Statistical Assumptions and Related Considerations

	Period 1 (1981-2008)[a]	Period 2 (2009-2020)[b]
Normality check	53%	52%
Missing data	31%	23%
Linearity	17%	5%
Outliers	25%	6%
Multicollinearity	11%	11%
Residuals	6%	0.70%
Median Sample size	182	325
Others		
Diagram	85%	82%
Reliability	72%	80%

Notes. [a & b] In all the tables the total numbers of models in the first and second periods are 302 and 420 models from 51 articles and 94 articles respectively.

As shown in Table 1, the studies in the second period resembled those of the first in assessing normality (52%), multicollinearity (11%), and in providing diagrams (82%). However, less often linearity tests (5%), missing data (23%), outliers (6%), and residual examinations (0.70%) were reported and discussed. In addition, the median sample size in Period 2 was large enough (*N*=325) according to Kline (2016).

6.3. Variables and Models

As pointed out in the literature review, as a highly flexible procedure, SEM allows for a wide range of variables and relationships to be modeled. Following Plonsky and Ghanbar's (2018) review of multiple regression analyses, we coded the studies in the sample for both the number and the nature (e.g., linguistic, non-linguistic; categorical, continuous) of independent (exogenous) and dependent (endogenous) variables. Despite its analytical flexibility, SEM is generally based on a relatively small set of continuous variables. The median and the modal exogenous (continuous independent) variables in the SEM analyses were 3 and 2, respectively, in Period 1. By contrast, the vast majority of the SEM analyses (97%) did not include a single categorical independent variable and vary in nature. More specifically, 69% of them involved only linguistic independent variables, 13% were entirely non-linguistic (e.g., cognitive, demographic), and 18% included a combination of linguistic and non-linguistic IVs.

The results for Period 1 were quite similar to those in Period 2 in that both the median and the modal values for exogenous (continuous independent) variables were 2 and 2, respectively. Similarly, only 6% included a single categorical independent variable. In terms of the nature of variables, 45% of the analyses involved only linguistically independent variables, 35% were solely non-linguistic (e.g., cognitive, demographic), and 20% combined linguistic and non-linguistic IVs.

Regarding the structural parts of the models (k = 62) in Period 1, 41% analysis included only one DV. And a smaller number of models comprised two (13%), three (27%), four (17%),

five (0%), and six (2%) DVs. As with IVs, DVs were more often linguistic in nature (52%) than non-linguistic (11%) or both (38%).

In Period 2, the structural models (k = 152) were mainly composed of one (39%), two (21%), and three (22%) DVs. Likewise, fewer models were observed with four (9%), five (6%), and six (3%) DVs. The variables were linguistic (53%), non-linguistic (29%) or both (18%).

6.4. Types of Models

With respect to the different types of relationships being modeled in Period 1, the studies exhibited a stronger preference towards measurement models (the CFA part) (79%) than structural models (21%) in the first period. Twenty articles (38%) analyses involved two-phase modeling whereas 32 articles (63%) encompassed only CFAs. Of all the measurement models, 94% used reflective measurement models, none had any causal indicators (formative model) or path analysis, and 6% could not be specified as they did not have a schematic representation. Furthermore, 24 % of the 62 studies comprising structural models involved mediation analyses. It was also noted that a small portion of the analyses (15%) specified higher-order factors.

Table 2
Type of Models

	Period 1 (1981-2008)	Period 2 (2009-2020)
Type of relationships in models		
Measurement models (CFA)	79%	61%
Structural models (LVPA[a])	21%	36%
Path analysis	0%	3%
Model specification issues		
Reflective	94%	72%
Formative	0%	1%
unspecified	6%	27%
Mediation analysis	24%	25%
Higher-order factor model	15%	14%

Note. [a] Latent variable path analysis which is commonly called a structural model.

In period 2, there was a larger number of structural models including 61% and 36% measurement (CFA) and structural models, respectively, though 56 articles (60%) presented both types of models in tandem (two-phase modeling), whereas 38 articles (40%) just contained CFAs and/or path analysis. We observed that L2 researchers preferred to use two-phase modeling to a larger extent in the second period. Of note in this period is the existence of 11 (3%) of models identified as path analysis SEM testing only the relationship between observed variables. There appeared that fewer studies tended to use formative indicators (72%) in this period. However, the number of SEM studies involving mediation and second-order analysis were found to be similar to that in the first period.

6.5. Model Estimation

According to Table 3, in Period 1, as expected the most frequent model estimation technique, by far, was ML, which was found in 57% of the sample. ML is the most popular and default estimation method in most SEM software packages (Mueller & Hancock, 2008). The other techniques observed were bootstrapping (3%) and robust ML (5%). No studies reported applying ordinary least squares (OLS), WLS, GLS, or ADF. Despite the recommendation to report the estimation method and the rational for its use, even if it is a default one in the SEM software package (see for example, Kline 2016), for 35% of the sample, model estimation techniques were not reported at all. The results in the second period generally resembled those in the first period except that ML use reduced from 57% to 45%. Given the recency of this period and the growing SEM literacy, we expected to see this difference to have gone under the other techniques. However, it was just because more studies ignored reporting the estimation method.

Table 3
Model Estimation

	Period 1 (1981-2008)	Period 2 (2009-2020)
Model estimation		
Not reported	35%	50%
OLS (PLS)	0%	2%
ML (Covariance-based)	57%	45%
WLS	0%	1%
GLS	0%	0%
ADF	0%	0%
Bootstrapping	3%	1%
Robust ML	5%	2%

6.6. Reporting Practices

The raw input to SEM analyses yields several correlations, variances and covariances for both measurement and structural models. There has been much variation in what SEM researchers opted to report and justify. Figures 6 and 7 summarize the statistics reported in periods 1 and 2. Notably, none of the estimation results for measurement or structural models surpassed 50% except for squared multiple correlations, and most were actually found in single-digit percentages. The results were slightly different in Period 2. As shown in Figures 6 and 7, there was a rise in reporting standardized regression weights both in measurement and structural models. Of note in this period was a dramatic increase in reporting descriptive statistics of measured variables and correlation matrix in measurement models and a decrease in presenting squared multiple correlations in structural models.

Figure 6

Reported Statistics Associated with Measurement Model Estimation across two Time Periods (in %)

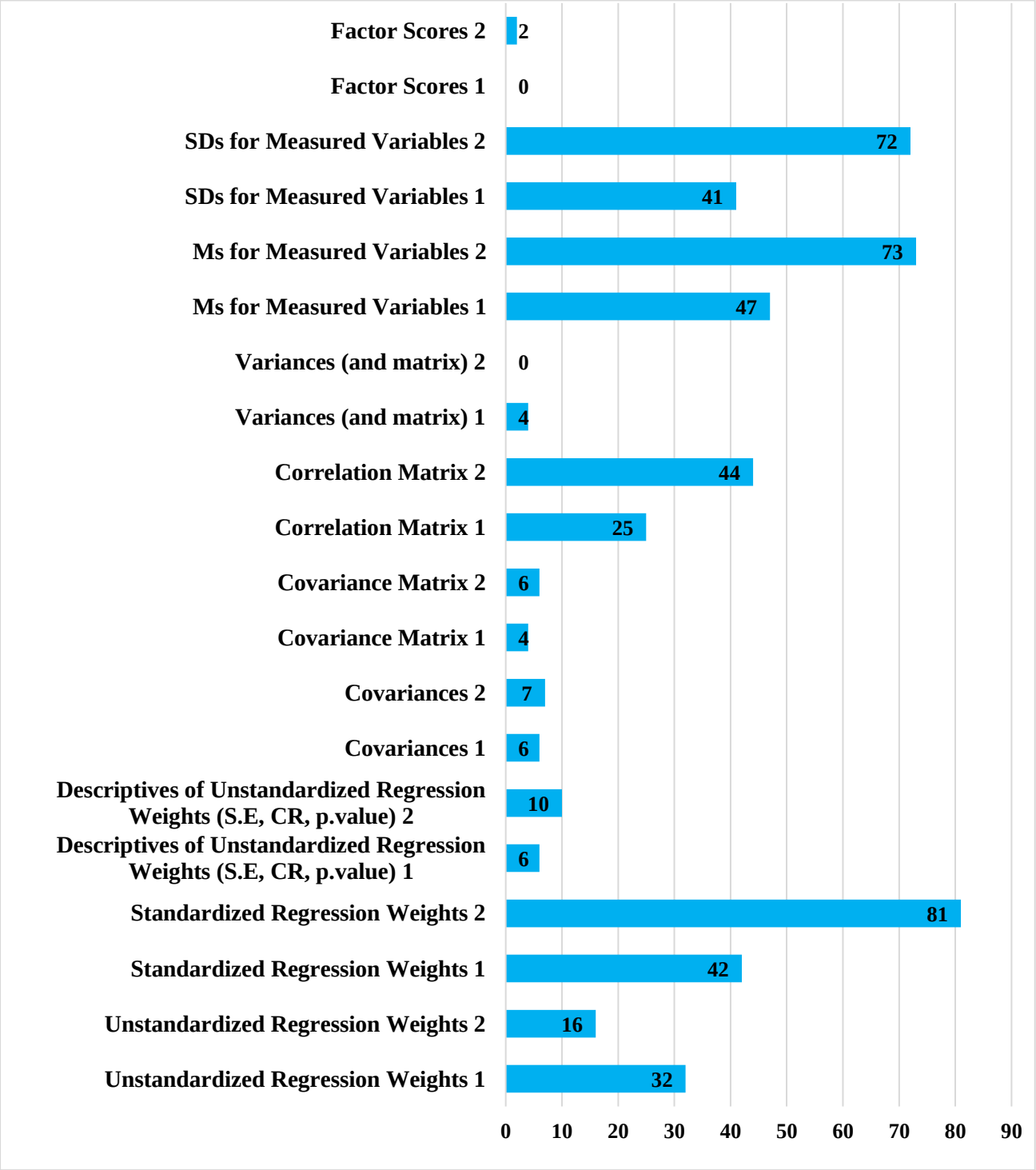
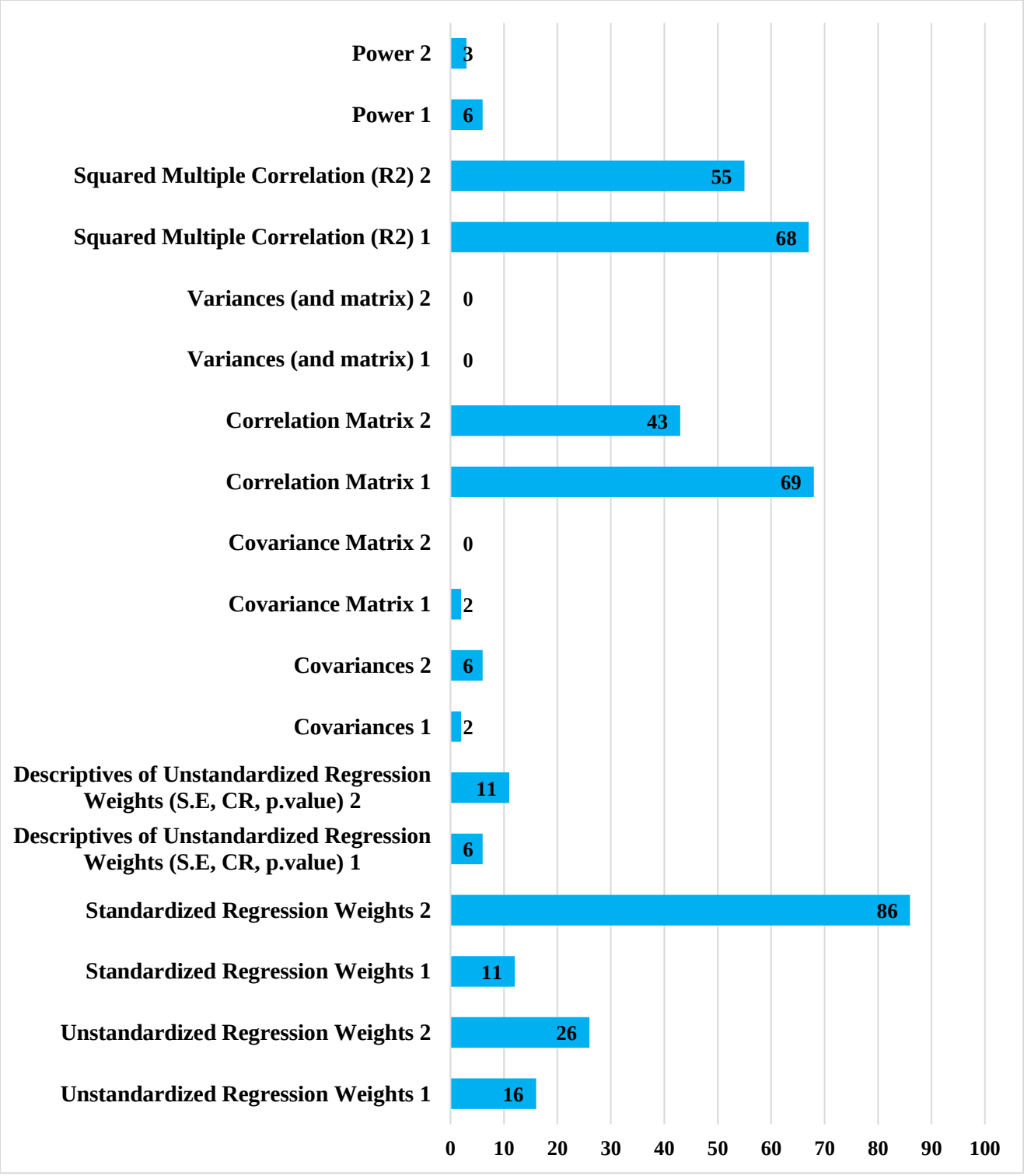


Figure 7

Reported Statistics Associated with Structural Model Estimation across two Time Periods (in%).



6.7. Fit Indices

Fundamental to presenting the results of structural equation models are a number of different types of fit indices including basic fit indices, incremental or comparative fit indices,

model parsimony fit indices, discrepancy fit indices, residual fit indices, and predictive fit indices. Table 4 summarizes the array of reported indices as well as their values that were aggregated (meta-analyzed) here as a point of possible comparison for future studies. We will here point to the most noteworthy indices or changes.

As with basic fit indices, in both periods, the three common indices of Chi-square (80% vs. 71%), Delta Chi-square (47% vs. 60%), and *p*-value (56% vs. 71%) were often reported. As indicated, the report of Delta Chi-square and *p*-value (56% vs. 71%) increased noticeably. Two comparative fit statistics commonly reported and also recommended in the SEM literature, that is, CFI and TLI were also presented in the sampled studies and the most commonly reported absolute fit statistics in the SEM literature is GFI. GFI was also most frequent in this category though it reduced from 31% to 21% in the second period. Interestingly, no discrepancy fit index was reported in the whole sample. Pertaining to residuals, it can be said that it was the category in which all the statistics were reported more frequently in the second period. In this category SRMR, RMSE, and RMSEA with confidence interval increased substantially across time. Finally, among the predictive fit indices, AIC, which is the best known (Kline, 2016), saw a dramatic increase (19% vs. 47%) in the second second-period reports.

Table 4
Distribution of Reported Fit Indices in L2 Research

	Period 1 (1981-2008)		Period 2 (2009-2020)	
Fit Type	%	M (SD)	%	M (SD)
Basic Fit Indices				
Chi-square (χ^2)	80	312.26 (920.66)	71	493.76 (1472.58)
p-value (F)	56	N/A	58	N/A
Degree of freedom (*df*)	37	N/A	26	N/A
χ^2/df	30	3.32 (2.75)	26	3.71 (12.26)
Delta Chi-square	47[a]	N/A	60[a]	N/A
Incremental or Comparative Fit Indices				
Comparative Fit Index (CFI)	54	.91 (.12)	72	.94 (.11)
Differences in CFI	3[a]	N/A	18[a]	N/A
Incremental Fit Index (IFI)	4	.90 (.10)	5	.96 (.03)
Normed Fit Index (NFI)	8	.87 (.09)	15	.94 (.04)
Parsimony Comparative Fit Index (PCFI)	1	.79 (.19)	3	.65 (.15)
Parsimony Normed Fit Index (PNFI)	1	.73 (.06)	2	.66 (.09)
TLI	4	.83 (.28)	27	.86 (.20)
Relative Fit Index (RFI)	1	.73 (.17)	0.2	.92[b] (N/A)
P Ratio	0.3	.88 (N/A)	0	N/A
Absolute Fit Indices				
Goodness-of-Fit Index (GFI)	31	.92 (.08)	21	.95 (.06)
Adjusted Goodness-of-Fit Index (AGFI)	11	.84 (.12)	11	.90 (.13)
Parsimony Goodness-of-Fit Index (PGFI)	0	N/A	0.2	.66[b] (N/A)
Discrepancy Fit Indices				
Noncentrality parameter (NCP)	0	N/A	0	N/A

FIMIN	0	N/A	0	N/A
Residual Fit Indices				
Root Mean Square Residual (RMR)	2	.05[c] (.01)	4	.03[c] (.18)
Standardized Root Mean Square Residual (SRMR)	3	.03[c] (.13)	32	.05[c] (.10)
Root Mean Square Error of Approximation (RMSEA)	50	.07[c] (.11)	66	.05[c] (.10)
RMSEA with confidence interval	8	N/A	32	N/A
P-value for Test of Close Fit (PCLOSE)	0	N/A	3	N/A
Predictive Fit Indices				
Akaike Information Criterion (AIC)	18[c]	163.21[c]	47[a]	1026[c]
Consistent Akaike Information Criterion (CAIC)	18[c]	223.23[c]	16[a]	289[c]
Browne-Cudeck Criterion (BCC)	13[a]	196.20[c]	5[a]	41866[c]
Bayesian Information Criterion (BIC)	0	N/A	19[a]	33586[c]
Expected Cross Validation Index (ECVI)	26[a]	2.91[a]	0	N/A
Modified Expected Cross Validation Index (MECVI)	0	N/A	0	N/A

Notes. [a] Normalized to include only those studies that use an alternative model comparison. [b] Reported in just one study. [c] Median reported due to very high variance in fit indices.

6.8. Model Modification and Acceptance

Two additional sets of practices associated with SEM were also examined. In Period 1, although no studies reported having examined standardized residuals (0%), almost one-third (27%) considered potential model misspecification using modification indices. Further, of the two main approaches for determining when model fitting should cease, 2% reported relying on substantive theory and 98% applied statistical fit criteria. The results were similar in Period 2. Few studies reported having examined standardized residuals (0.01%) and almost one-third (32%) considered potential model misspecification using modification indices. Similarly, only 2% of the studies reported relying on substantive theory to justify that the modifications make sense and 98% resorted to statistical criteria in model modification.

7. Discussion and Recommendations for Future SEM Applications

This study set out to look into the use of SEM in L2 research and to bring to light the rigor and transparency of SEM analyses across two time periods, that is, from 1980 to 2008 (the period examined in In'nami & Koizumi, 2011) and from 2009 to 2020. In total, we analyzed 722 models found in 145 articles (302 models from 51 articles in the first period and 420 models from 94 articles in the second period). In what follows, we expound upon major issues emerging from our review of SEM use across the two periods. It is certain that this useful and versatile method will continue to burgeon. Thus, we also seek to provide L2 researchers with recommendations to enhance the quality of SEM utilization in future studies.

The first major theme identified in the findings was the growing trend in the utilization of SEM as also noted earlier by In'nami and Koizumi (2011). The growing popularity lies in a number of reasons. L2 is deemed to have attained theoretical and methodological maturity (see Gass et al.,

2021; Plonsky, 2014) abounding with many constructs and variables. L2 researchers have long been interested in looking at such variables in tandem. This is exactly where SEM comes in. SEM can assist L2 researchers to model and assess more complex relationships of multiple variables, manifest or latent, in a single study. SEM use gained in popularity by the development of a computer program to examine linear structural relationships (LISREL) by Jöreskog (1970). It grew, even further, by a variety of software programs for SEM, modeling more sophisticated relationships (Schumacker & Lomax, 2016) and with yet simpler syntax and user interface (Kline, 1998).

Although SEM is quite flexible, several statistical assumptions and considerations must be examined. A basic statistical assumption in some model estimation methods like ML (Byrne, 2016, Pituch & Stevens, 2016) is normality (univariate and multivariate), overlooked in approximately 50% of both samples. The other half of the studies also examined and reported for the most part the univariate normality. The absence of normality in the SEM data or negligence in examining it can jeopardize the accuracy of the SEM analysis and more specifically fit indices and estimated error and model parameters (Byrne, 2006). Hence, for univariate normality, careful scrutiny of skewness and kurtosis (e.g., within the Z statistics) is recommended. For multivariate normality, Mardia's normalized coefficient of multivariate kurtosis should be checked and reported (Bentler, 2005).

Another important preliminary step in performing SEM analysis involves checking for missing and outliers, not reported to have been checked in the majority of studies in both samples, although some improvement was seen in the second period. Missing values can have an adverse effect on modification indices or model estimation results (Enders, 2010). As Mueller and Hancock (2019) pointed out, traditional techniques for missing data treatment (e.g., listwise or pairwise deletion, or mean replacement) are now considered inadequate, unless missing values comprise less than 5% of data in each variable (Hair et al., 2017). We, thus, strongly recommend that L2 researchers use multiple methods such as full-information maximum likelihood (FIML) estimation and multiple imputations (MI), and report the proportion of missing cases corresponding to each variable (for more on missing data, see Graham & Coffman, 2012).

Only one-quarter of the studies in the first period (25%) and a negligible number of them in the second period (6%) examined and reported atypical or extreme data. Similar to normality, the presence of outliers should also be detected from both univariate and multivariate perspectives (see Kline, 2016; Tabachnick & Fidell, 2013), and when statistically justified, be removed from the data set and reported in the study. When they are not justifiably deleted, they should be accommodated or explained in the report (Schumacker & Lomax, 2016).

Another key statistical assumption in SEM is multicollinearity, which was largely overlooked in both samples. In SEM, as a correlation-based technique, covariance matrixes are used for estimating the model parameters. Fortunately, many SEM programs like …… now terminate model estimation if the covariance matrix is singular (variables are perfectly correlated with each other). We, accordingly, recommend inspecting covariance/correlation matrixes available in program outputs to examine multicollinearity or singularity.

One of the most controversial, but critical issues in SEM, is the sample size. The SEM literature is laden with a rich variety of suggestions and recommendations germane to sample size. For example, as discussed before, Kline (2016) proposed a classification for different sizes of

sample for a typical SEM study, and Raykov and Marcoulides (2006), taking into account the model complexity, proposed a 10-times rule, that is, an appropriate sample size would be 10 times the number of free parameters of a model. This idea is motivated by the 10-times rule of Barclay et al. (1995), which suggested that the minimum sample size should be 10 times the maximum number of arrows pointing at a latent variable in a model. Nonetheless, the issue of sample size in SEM is not that straightforward and depends also on a number of other considerations. For instance, Mueller and Hancock (2019) recommended authors to consider both adequacy for correct parameter estimation and desired level of statistical power (rarely reported in sampled studies) when they aim to determine a desirable sample size. Power analysis in SEM can be conducted using the Mplus software (see In'nami & Koizumi, 2013; Muthén & Muthén, 2002) or the simulation-based Shiny app pwrSEM (Wang & Rhemtulla, 2021). Another issue to be taken into account when determining the sample size is the estimation method to be used in the analysis. ML, for instance, requires a minimum of five cases for each model parameter in comparison to WLS, which entails larger sample sizes (see Lei & Wu, 2012).

As indicated in the results of the study, L2 researchers largely included sufficiently large samples based on the Kline's recommendation. What complements the accuracy of the findings is the transparency of the research methodologies by communicating how the decisions were made about determining the sample size or even failing to include an adequate sample. To recap, we strongly urge L2 researchers to consider several factors such as data characteristics, data structure, estimation method, statistical power, model complexity, reliability of indicators, and expected R^2 values, when they aim to select a sample size, as a priori decisions about sample size are very difficult and often-cited rules of thumb fail to capture the complex nature of individual models. Although L2 researchers in both periods rarely used Hoelter's Critical *N*, which can be used to determine whether the selected sample size was large enough or not, we cannot say anything here about the quality of their decisions with regard to the sample size, given that it is contingent upon many factors. However, considering Kline's (2016) a priori classification, researchers in our field have generally selected a sufficiently large sample size for their studies in both periods, with samples found to be somewhat larger in the second period.

Model specification is another point worthy of being discussed here. Recent SEM software packages provide graphical user interface platforms for practitioners and researchers. Although many studies in both periods included a visual representation of models, we noted that many studies which involved two-phase modeling, did not provide specific visualization for each phase and tended to present the integrative representations that are usually complicated and difficult to follow readily. We recommend that L2 researchers take advantage of the graphical presentation interface of SEM software programs and visually provide the hypotheses and models as they are developed and modified for the interest of their clarity and to help readers to keep track of the steps and changes.

Further, there was a paucity of studies incorporating categorical variables like gender or age in the studies in both periods. We reckon that the complexity of techniques in estimating models involving such variables multigroup invariance technique (for a discussion see Byrne, 2004, 2016) is defying. Nevertheless, comparing estimated parameters across different groups or populations is one of the most valuable options offered by SEM, which is not available in EFA. We strongly recommend L2 researchers to make use of this technique (factorial equivalence) rather

than utilizing the mean of indicators in questionnaires to test for potential differences across groups of participants. Also, pertaining to the number of latent variables in the model which was not very high in this review, it should be said this cannot be judged as a categorically 'good' or 'bad' finding, because L2 researchers are always forced to strike a balance between model parsimony and substantive meaningfulness of model specification. In other words, adding more latent variables can boost model fit and represent more variance, but it can also pose a threat to model parsimony, indicating the delicate nature of the relationship between model parsimony and model fit in SEM.

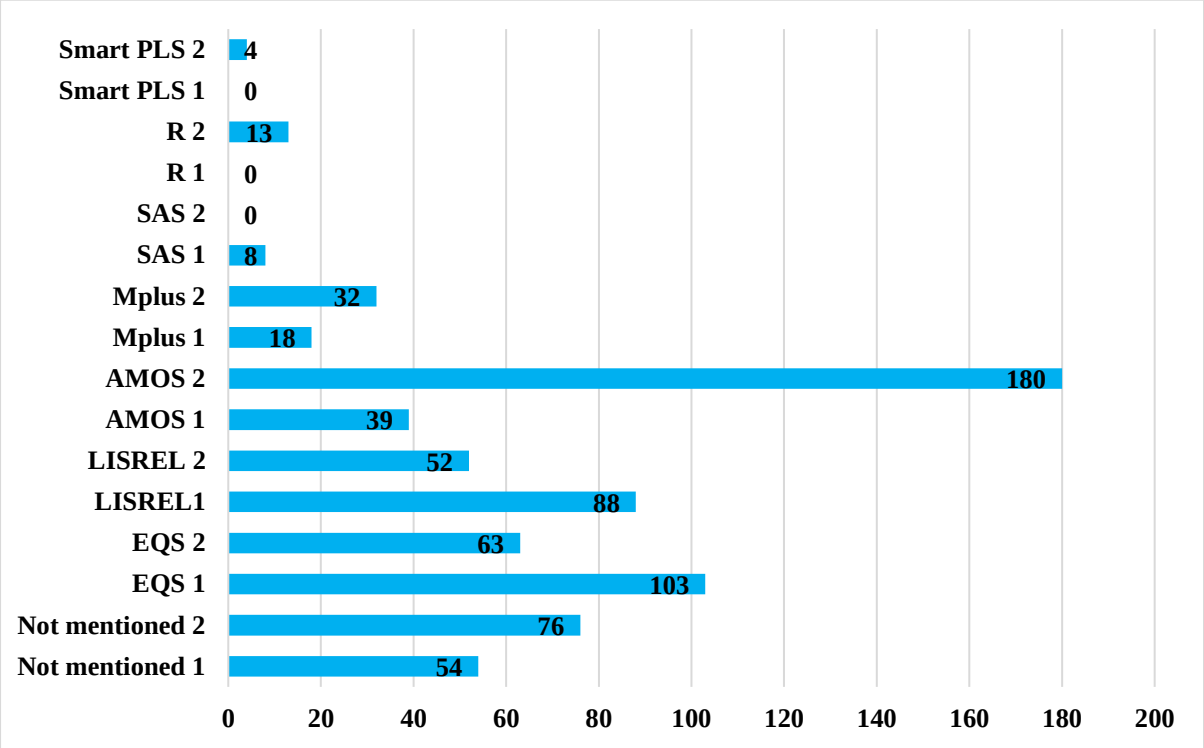
The next component of SEM reports that we examined is the model estimation method. We discuss the application of these methods here in conjunction with an analysis of the SEM software programs used in the sample, as there is a close relationship between the estimation method and the type of program used. We found that ML was the most often-used estimation method in both samples, with a decrease in its popularity in the second period in favor of other methods such as PLS (it is used in smart PLS programs) or WLS (see Figure 8). However, its robust counterpart, robust ML, was rarely utilized in either period. This finding was expected as ML is the default estimator in all four of the most frequently used programs (see Figure 8 for a view on the usage of SEM programs in L2 research and also see Byrne, 2001 for a detailed review of SEM software). Nonetheless, robust ML, which is a regular ML along with robust standard errors and scaled model X^2 of Satorra and Bentler (1994), is only available in LISREL, EQS, and Mplus. It should be said that robust ML is a very good choice when continuous data lacks normality. It should also be noted that EQS, which decreased its popularity in the second period and became the second most widely used SEM program, provides a wide range of residual-based X^2 tests which are very versatile in studies based on smaller samples (Yuan & Bentler, 2000).

Also intriguing from this set of findings was that ADF (WLS) was seldom used in either period, despite the fact that it is available in four programs and can be used for ordered categorical variables (Browne, 1984). The ADF estimator can also be a choice when the data is not normal, notwithstanding that it requires a large sample size (>1,000), which may explain its infrequent use among L2 researchers. EQS, by providing Yuan–Bentler corrected arbitrary distribution generalized least squares (AGLS) test statistic (Yuan & Bentler, 1997), Yuan–Bentler AGLS *F*-statistics (Yuan & Bentler, 1999), and corrected standard error estimates for small samples can be a good option when non-normality is present. As multivariate normality of variables required by ML is very difficult to become tenable, and under small sample sizes ML produces biased estimates (Ferron & Hess, 2007; Pituch & Stevens, 2016), we strongly recommend that future SEM applications use alternative estimation methods such as Satorra–Bentler scaled X^2, Yuan–Bentler residual-based X^2 (produced in EQS), robust WLS (Flora & Curran, 2004), PLS, and Bayesian methods. For instance, PLS, used in just one study in the second period can be utilized by L2 researchers when they seek to develop a theory and explain variance for the purpose of prediction of constructs, as it is a variance-based approach to SEM and functions very effectively under conditions of small sample size, complex models, and erratic data structures (see Hair et al., 2011, for a full discussion on options provided by PLS and also Hair et al., 2017 for a discussion on capabilities of SmartPLS program). To close this section, we urge prospective users of SEM to consider such factors as sample size, distributional assumptions and structure of data, model complexity, and empirical findings regarding the results of various types of estimators in diverse

practical conditions to have a more judicious choice of estimation methods, as it significantly affects both estimated parameters and fit indices.

Figure 8

Use of Software Packages in SEM L2 Research across two Periods (in %)



Comprehensiveness and transparency are two critical elements in any statistical analysis. SEM, as a complex statistical technique, yields different types of estimation results, some of which are worthy of being reported. One output is regression weights, standardized and unstandardized, which are of prime importance in comparing estimations across different populations or constructs. Nevertheless, half of the sample failed to report them for either the measurement (CFA) or structural models in the first period. Some improvement was observed in the second period, suggesting growing statistical and SEM-specific literacy in recent years. It was of note that descriptive statistics of unstandardized regression coefficients were not reported routinely in either period despite a modest increase in the second period. Normally, the exact magnitude of estimates, their standard errors (S.E.), critical ratio (C.R.), and the exact *p* values should have also been reported both in the measurement and structural portions of the model. Standard errors, for example, are very informative, as they show the level of accuracy with which a parameter is estimated. Additionally, as Byrne (2016) pointed out, standard errors can signify poor model fit given very large or small standard errors can be considered as a red flag, although no universal cut-offs have been proposed for the standard errors. Further indications can also be obtained from the critical ratio (a parameter estimate divided by its standard error), as it displays the statistical significance of each parameter estimate, with values more than +1.96 signifying a significant contribution of an indicator at the .05 level.

Of critical import here, also, is examining and reporting the covariance matrix and its standardized correlation matrix. However, L2 researchers seldom provided such matrices in either period, despite the noticeable increase in reporting correlation matrix in the second period. Since SEM is a correlation-based technique, these matrices provide valuable information regarding the modeled constructs. A relevant issue here is multicollinearity and singularity, two important statistical assumptions of SEM. As recommended by Meyers, Gamst, and Guarino (2013), researchers would do better to investigate the correlation matrix to find constructs that are not independent from each other. Along the same lines, the covariance residual matrix (matrix of the differences between the observed covariance matrix and the model-implied covariance matrix), which was not seen in the sample in either period, can also be used as an indicator of model adequacy, as it shows the extent to which a model represents the reality of data under analysis, facilitating the identification of potential model misspecification.

Another SEM output in regard to latent variable path analysis is the magnitude of direct, indirect, and total effects of latent endogenous variables on each other. These statistics were not frequently reported in either time period. The reporting of the standardized direct and indirect effects is of paramount importance in mediation analysis in SEM (see Meyers et al., 2013 and Jose, 2019 for a discussion on mediation analysis). We recommend that future L2 researchers report the squared multiple correlations as a statistic of basic importance and indicative of the extent to which the variance in DVs (endogenous variables) is accounted for by other latent variables.

The final note here is related to the reliability of constructs. Many studies reported reliability, a reporting practice that improved in Period 2 mostly using Cronbach's alpha. This index is somewhat problematic in that it assumes, unrealistically, that all indicators of a construct are equally reliable, and in that, it is very sensitive to the number of items in a scale. We, thus, recommend reporting composite reliability (CR), which is calculated based on different magnitudes of regression weights of a construct (values between .6 to .9 are considered acceptable, see Nunnally & Bernstein, 1994). Further, we recommend maximal reliability estimates such as Coefficient H (Raykov et al., 2015), as it reflects the correlation that the factor is predicted to have with itself over repeated measurements, suggesting the stability of a construct, with values more than .7 considered generally acceptable (Mueller & Hancock, 2019).

The last part of our review and analysis concerned model modification and goodness of fit statistics in SEM. Once the model is specified and estimated, it is imperative to examine the extent to which the hypothesized model fits the observed data. Given that fit of a model can be examined from different points of view, which is a complicated, multidimensional issue, various fit indices have been proposed to provide information on specific aspects of model fit. Different software programs offer various indices. Amos, for example, provides 25 indices (see Byrne, 2016, for a comprehensive list of fit indices). The first type of fit indices, basic fit indices, shed light on the overall fit of the theoretical model to the data, which was reported by the majority of studies in the sample, but one of them X^2/df was not reported as frequently in either period. It should be noted that the Chi-Square test of model fit is very sensitive to sample size (Hu & Bentler 1999) and is likely to yield a significant result when a sample size is large suggesting the population covariance matrix is different from the reproduced implied matrix (it shows a lack of fit). Hence, our first recommendation is that L2 researchers report X^2/df to address that drawback, and values between 2 and 5 can be considered to have indicated an acceptable model fit (see Meyers, Gamst, &

Guarino, 2013). The results illustrated that studies in both periods often reported values within that range.

The next set of fit indices complementing the Chi-square test includes fit indices which measure the proportionate improvement in a proposed model relative to an independence or baseline model (a model in which no correlations are assumed among variables). In both samples, CFI and TLI were reported rather frequently with an increase in the second period) and Tucker-Lewis Index (TLI), not reported frequently in the first period, showed a dramatic increase in the second period. However, other indices such as the Normed Fit Index (NFI or its parsimony version, PNFI), Incremental Fit Index (IFI), PCFI (or ΔCFI for competing models), and RFI were not reported frequently and in conjunction. Our recommendation, in general, is that L2 researchers present several fit indices and avoid selective reporting of more supportive indices. The more relevant fit indices are reported, the more vivid picture is provided for the interest readers and future researchers. For example, IFI, which was not routinely reported in either time period, was developed by Bollen (1989) because of NFI's problem of underestimating the fit when a sample size is small and a model is complex. Accordingly, we recommend reporting CFI, or PCFI (ΔCFI is used for comparing competing models, which was rarely reported in the first period when alternative models were compared but we observed a promising increase in the second period), NFI (or its parsimony version, PNFI), NNFI (TLI), RFI, and IFI to have a full, lucid picture regarding the extent to which the proposed model is a better fit than a baseline model. According to Hu and Bentler (1999), values of approximately .95 represent a well-fitting model (for PCFI and PNFI values more than .5 are considered acceptable [results showed that reviewed studies in both periods reported acceptable values for PCFI and PNFI], and they are optional when CFI and NFI are reported in conjunction, see Williams & Holahan, 1994). Overall, considering Hu and Bentler's (1999) recommendation, it was found that incremental fit indices in the sample were marginally acceptable.

Another set of results in this study was concerned with absolute fit indices examining the extent to which a model is successful in reproducing an observed correlation/covariance matrix. The pattern of reporting practices of these indices was unsatisfactory in both periods. For example, Goodness-of-Fit (GFI) as one of the most commonly used index, showing the extent to which variance in the observed correlation/covariance matrix is accounted for by the imposed model (Kline, 2016) (this is similar to the R^2 value in regression analysis), was not reported frequently in the whole sample of this study. Similarly, it should be said that not only was GFI not reported frequently in the first time period, but its reporting actually declined in the second period. Adjusted Goodness-of-Fit (AGFI), rarely reported in the two samples, is similar to GFI, but it is only adjusted for the number of degrees of freedom in a model (i.e., as the number of parameters increases, AGFI decreases). Parsimony Goodness-of-Fit (PGFI), seldom used in any studies in either time period, has a useful function of taking into account model complexity (i.e., the number of estimated parameters) in evaluating model fit. We, thus, suggest reporting all three absolute fit indices, as they offer valuable and precise information regarding the amount of variance accounted for by a model, with values of .95 (for GFI and AGFI), and .50 (for PGFI) or higher signifying a good fit (Byrne, 2016; Meyers, Gamst, & Guarino, 2013). Overall, results revealed that absolute fit indices in both samples were just marginally acceptable.

Another crucially important set of fit indices is residual-based fit indices, representing the average differences between the observed correlations/covariances and estimated, model implied ones for the population. Several indices can be reported in this part, with Root Mean Square Error of Approximation (RMSEA) being the most frequently reported one (we saw a decrease in reporting RMSEA in the second period), in spite of the fact that its accompanying confidence interval (reporting confidence interval for RMSEA improved in the second period which showed a good consciousness in this regard), range, and a test for the closeness of fit (PCLOSE) were seldom included in SEM reports. Regarding interpretations of such values, the lower bound of the 90% CI for the RMSEA should ideally approach zero, while the upper bound should be below .08 (Byrne, 2016). These pieces of information are of prime significance and they should be reported, as they show that we can be 90% confident that the real RMSEA value in the population will be in a specific range, and that PCLOSE suggests the significance of closeness of fit (*p* values lower than .05 are preferred). Of relevance to this part is reporting Standardized Root Mean Square Residual (SRMR), recommended by Hu and Bentler (1999), which was not identified as a norm in either period, with a slight decrease in the second period. This is the standard version of RMR, and because the values of RMR vary according to the magnitude of variances and covariances (Byrne, 2016), it is necessary for L2 researchers to report SRMR, which is standardized and easier to interpret. As Amos is the most frequently used program in the sample and this program does not directly produce SRMR (a plugin must be installed for so doing), the lack of reporting of this index was somewhat expected, as noted by In'nami & Koizumi (2011). Nevertheless, we recommend that L2 researchers report residual-based fit indices including RMSEA (accompanying associate CI and *p*-value, with RMSEA value of .06 to .08 indicating an adequate fit, as recommended by Hu and Bentler, 1999) and SRMR of .10 or less (Hu & Bentler, 1999), given that we saw that L2 researchers were very inconsistent in reporting this set of indices, especially regarding CI and *p* values of RMSEA and SRMR values. However, we acknowledge that the reported residual-based fit indices were in the acceptable range, considering Hu and Bentler (1999). Overall, as can be seen in Table 4, the reporting of this set of indices has shown an improvement over two periods.

The last set of fit indices that we examined was predictive fit indices. These indices were not often considered by L2 researchers in either period despite the fact that they offer very valuable information regarding the fit of a model when non-nested models are being compared. For example, Akaike Information Criterion (AIC) and Consistent Akaike Information Criterion (CAIC) are very useful for the cross-validation of models. The only difference between them is that the former is very sensitive to degrees of freedom (the more estimated parameters, the lower the fit), and the latter takes sample size into account in assessing the fit (see Byrne, 2016 for a full list and explanations on their functions; Bandalos, 1993). We strongly recommend using AIC, CAIC, and ECVI or BIC, BCC, and ECVI, when L2 researchers compare several models. When doing so, a model with lower predictive fit indices has greater potential for being replicated across different samples. To close this section about fit indices, we would simply remind the field that these indices can produce inconsistent results regarding model fit. Consequently, reviewers should expect to see several indices from different categories to better judge the goodness of fit of a model.

8. Conclusion

In conclusion, we acknowledge that SEM use entails expert statistical literacy and its use currently produces numerous sophisticated outputs. The SEM analysis results are often inconsistently reported and this might impede clear interpretation and subsequent use. This is what motivated this study. Reflecting on the current practice of SEM use and reporting in L2 research we sought to come up with a set of more clear recommendations as to what to report and discuss in studies involving SEM. There are also several classical books and SEM software manuals. They can also be referred to and consulted for further details. The problem, however, with these resources is that, unlike this review, they are not based on empirical data and, at times, present too many details compounding the confusion. There is no specific resource on the application of SEM in L2 research.

The price of versatility and diversity of hypotheses and research questions which can be addressed by SEM is that there are many requirements to be met and assumptions to be checked for its proper application and accurate and transparent reporting. What we did in this study was a systematic review of a wide range of SEM-related research and reporting practices in a sample of 722 SEMs found in published L2 research and then we discussed the trends in light of the SEM literature for future L2 SEM studies. In closing this study, we tend to wrap up the recommendations and suggestions made above in further clarity. They can be used as a reminder or a checklist guiding future L2 researchers in using and reporting SEM. We tend to, in the first place, argue that despite the SEM sophistication and several statistical guideline and recommendations given, researchers are not, and should not be, captive to statistical significance. Any SEM use, interpretation, and report should be guided by reason and substantive theories, not pure statistics. Given the abundance of SEM outputs and indices and space limitations for publication, L2 researchers are to be selective in its applications and reports. Too much or irrelevant information might create a mess hindering the readers to follow and evaluate the SEM use. Overall, we recommend that L2 researchers articulate the decisive steps taken and choices made in the SEM application transparently and justify them. They are also urged not to be selective in the report of fit indices in particular in favor of the supporting ones. Besides the SEM requirements, L2 researchers should include, in general, whatsoever that permits other researchers to be able to replicate their study and verify the results.

Declaration of Conflicting Interests

The authors declare that they have no conflicts of interest in the current research.

Funding

The authors received no financial support for the current research.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423. https://doi.org/10.1037/0033-2909.103.3.411

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Bandalos, D. L. (1993). Factors influencing cross-validation of confirmatory factor analysis models. *Multivariate Behavioral Research, 28*, 351–374. https://doi.org/10.1207/s15327906mbr2803_3

Barati, H., Ravand, H., & Ghasemi, V. (2013). Investigating Relationships among Test Takers' Characteristics and Response Formats in a Reading Comprehension Test: A Structural Equation Modeling Approach. *International Journal of Language Testing*, *3*(2), 38-59.

Barclay, D. W., Higgins, C. A., & Thompson, R. (1995). The partial least squares approach to causal modeling: Personal computer adoption and use as illustration. *Technology Studies, 2*, 285–309.

Baumgartner, H. & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, *13*(2), 139-161. https://doi.org/10.1016/0167-8116(95)00038-0

Bentler, P. M. (1980). Multivariate analysis with latent variables: causal modeling. *Annual Review of Psychology*, *31*(1), 419–456.

Bentler, P. M. (2005). *EQS 6 Structural equations program manual*. Encino, CA: Multivariate Software.

Bollen, K. A. (1989). A new incremental fit index for general structural models. *Sociological Methods & Research*, *17*, 303–316. http://dx.doi.org/10.1177/0049124189017003004

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods, 16*, 265–284. https://doi.org/10.1037/a0024448

Brown, J. D. (2015). Why bother learning advanced quantitative methods in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 9-20). New York: Routledge.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed). London: Guilford Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83. 10.1111/j.2044-8317.1984.tb00789.x

Byrne, B. M. (2001). Structural Equation Modeling with AMOS, EQS, and LISREL: Comparative Approaches to Testing for the Factorial Validity of a Measuring Instrument. *International Journal of Testing*, *1*(1), 55–86. https://doi.org/10.1207/S15327574IJT0101_4

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling, 11*(2), 272–300. https://doi.org/10.1207/s15328007sem1102_8

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

Byrne, B. M. (2016). *Structural equation modeling with AMOS*: *Basic concepts, applications, and programming* (3rd ed). New York, NY: Taylor & Francis.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105. https://doi.org/10.1037/h0046016

Chong, S. & Plonsky, L. (2023). A typology of secondary research in Applied Linguistics. *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2022-0189

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly, 50,* 132-153. https://doi.org/10.1002/tesq.217

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38,* 269–277. https://doi.org/10.1509/jmkr.38.2.269.188

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Ferron, J. M., & Hess, M. R. (2007). Estimation in SEM: A concrete example. *Journal of Educational and Behavioral Statistics, 32*(1), 110–120. https://doi.org/10.3102/1076998606298025

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466–491.

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: The past, present, and future of quantitative SLA research. *Language Teaching*. doi:10.1017/S0261444819000430

Ghanbar, H. & Rezvani, R. (2023). Research Questions in Applied Linguistics Research: A Microscopic Analysis of their Distributional and Syntactical Aspects. *Journal of Research in Applied Linguistics,14*(1), 156-167. https://doi.org/10.22055/RALS.2023.18074

Graham, J. W., & Coffman, D. L. (2012). Structural equation modeling with missing data. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 277-295). New York, NY: The Guilford Press.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice, 19,* 139–151. https://doi.org/10.2753/MTP1069-6679190202

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling* (PLS-SEM) (2nd ed). London: SAGE Publication.

Hancock, G. R., & Mueller, R. O. (Eds.). (2006). *Structural equation modeling: A second course.* Greenwich, CT: Information Age Publishing.

Hayduk, L. A., & Glaser, D. N. (2000). Jiving the four-step, waltzing around factor analysis, and other serious fun. *Structural Equation Modeling, 7,* 1–35. https://doi.org/10.1207/S15328007SEM0701_01

Ho, K., & Naugher, J. R. (2000). Outliers lie: an illustrative example of identifying outliers and applying robust methods. *Multiple Linear Regression Viewpoints, 26*(2), 2–6.

Ho, R. M., Stark, S., & Chernyshenko, O. (2012). Graphical representation of structural equation models using path diagrams. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 43-55). New York, NY: The Guilford Press.

Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oak: Sage.

Hoyle, H. R. (2012). Introduction and overview. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 3-16). New York, NY: The Guilford Press.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118

In'nami, Y. (2006), The effects of test anxiety on listening test performance, *System*, 34(3), 317-340. https://doi.org/10.1016/j.system.2006.04.005

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly, 8*(3), 250–276. https://doi.org/10.1080/15434303.2011.582203

In'nami, Y., & Koizumi, R. (2013). Review of sample size for structural equation models in second language testing and learning research: A Monte Carlo approach. *International Journal of Testing, 13*, 329–353. https://doi.org/10.1080/15305058.2013.806925

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Los Angeles, CA: Sage.

Jose, P. E. (2019). Mediation and moderation. In G. R. Hancock, R. O. Mueller, & L. M. Stapleton (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 248-259). New York, NY: Routledge.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika, 57*(2), 239–251. https://doi.org/10.1093/biomet/57.2.239

Joreskog, K. G., & Sorbom, D. (1996). *LISREL8: User's reference guide*. Mooresville: Scientific Software.

Keith, T. Z. (2019). *Multiple regression and beyond* (2nd ed.). New York, NY: Routledge.

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

Khany, R., & Tazik, K. (2019). Levels of statistical use in applied linguistics research articles: From 1986-2015. *Journal of Quantitative Linguistics, 26*, 48-65. https://doi.org/10.1080/09296174.2017.1421498

Kline, R. B. (1998). Software review: Software programs for structural equation modeling: Amos, EQS, and LISREL. *Journal of Psychoeducational Assessment, 16*(4), 343–364. https://doi.org/10.1177/073428299801600407

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: The Guilford Press.

Larson-Hall, J. & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning, 65*(S1), 127-159. https://doi.org/10.1111/lang.12115

Lei, P., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164-180). New York, NY: The Guilford Press.

Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing, 30*, 66-81. https://doi.org/10.1016/j.jslw.2015.08.011

Loewen, S., Lavolette, E., Spino L., Papi, M., Schmidtke, J. Sterling, S. & Wolf, D. (2014). Statistical literacy among applied linguists and second language acquisition researchers. *TESOL Quarterly*, *48*, 360–88. https://doi.org/10.1002/tesq.128

Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS repository: Advancing research practice and methodology. In Mackey, A. & Marsden E. (Eds.) Instruments for Research into Second Languages: Empirical studies advancing methodology (pp. 1-21). New York: Routledge.

Marsden, E., & Plonsky, L. (2018). Data, open science, and methodological reform in second language acquisition research. In A. Gudmestad, & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 219-228). Philadelphia, PA: John Benjamins.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82. 10.1037/1082-989x.7.1.64

Meyers, L. S., Gamst, G. C., & Guarino, A. J. (2013). *Performing data analysis using IBM SPSS*. Hoboken, NJ: Wiley.

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

Mueller, R. O., & Hancock, G. R. (2019). Structural equation modeling. In G. R. Hancock, R. O. Mueller, & L. M. Stapleton (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 445-456). New York, NY: Routledge.

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed). London: CRC Press.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Narayanan, A. (2012). A Review of Eight Software Packages for Structural Equation Modeling. *The American Statistician*, *66* (2), 129-138. doi: 10.1080/00031305.2012.708641

Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research: A synthesis and data re-analysis from self-paced reading. *Annual Review of Applied Linguistics*, *40*, 26-55. doi:10.1017/S0267190520000057

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Ockey, G. (2011). Self-consciousness and assertiveness as explanatory variables of L2 oral ability: A latent variable approach. *Language Learning*, *61*, 968–989. doi:10.1111/j.1467-9922.2010.00625.x

Ockey, G. J. (2014). Exploratory factor analysis and structural equation modeling. In A. J. Kunnan (Ed.), *The companion to language assessment: Evaluation, methodology, and interdisciplinary themes* (pp.1224–1244). Malden, MA: John Wiley & Sons.

Ockey, G. J., & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Language Assessment Quarterly*, *12*(3), 305–319. https://doi.org/10.1080/15434303.2015.1050101

Peters, C. L. O., & Enders, C. (2002). A primer for the estimation of structural equation models in the presence of missing data. *Journal of Targeting, Measurement and Analysis for Marketing*, *11*, 81–95. doi: 10.1057/palgrave.jt.5740069

Pituch, K. A., & Stevens, J. P. (2016). *Applied multivariate statistics for social sciences* (6th ed.). New York, NY: Routledge.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition, 35,* 655-687. doi:10.1017/S0272263113000399

Plonsky, L. (2014). Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform. *Modern Language Journal, 98,* 450-470. https://doi.org/10.1111/j.1540-4781.2014.12058.x

Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning, 61,* 325-366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, *102*(4), 713–731. https://doi.org/10.1111/modl.12509

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106-128). New York, NY: Routledge.

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

Raykov, T., Gabler, S., & Dimitrov. D. M. (2016) Maximal reliability and composite reliability: Examining their difference for multicomponent measuring instruments using latent variable modeling. *Structural Equation Modeling, 23*(3), 384–391. https://doi.org/10.1080/10705511.2014.966369

Riazi, A., Ghanbar, H., & Rezvani, R. (2023). Qualitative Data Coding and Analysis: A Systematic Review of the Papers Published in the Journal of Second Language Writing. *Iranian Journal of Language Teaching Research.* doi: 10.30466/ijltr.2023.121271

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 variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Sawaki, Y. (2013). Structural equation modeling in language assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 5422–5427). Malden, CA: Blackwell Publishing.

Schoonen, R. (2015). Structural equation modeling in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 213-242). New York, NY: Routledge.

Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling* (2nd ed.). New York, NY: Routledge.

Stevens. J. (2009). *Applied multivariate statistics for social sciences* (5th ed). London: Routledge.

Tabachnik, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Allyn and Bacon.

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

Verkuilen J. (2011). A Comparative Review of Four Structural Equation Modeling Books. *Journal of Educational and Behavioral Statistics*, *36*(6), 832-834. doi:10.3102/1076998611420440

Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*(1), 1-17. https://doi.org/10.1177/2515245920918253

Williams, L. J., & Holahan, P. J. (1994). Parsimony-based fit indices for multiple indicator models: Do they work? *Structural Equation Modeling*, *1*, 161–189. https://doi.org/10.1080/10705519409539970

Winke, P. (2013). An investigation into second language aptitude for advanced Chinese language learning. *Modern Language Journal*, *97*, 109-130. https://doi.org/10.1111/j.1540-4781.2013.01428.x

Yuan, K. H., & Bentler, P. M. (1997). Improving parameter tests in covariance structure analysis. *Computational Statistics and Data Analysis*, *26*, 177–198.

Yuan, K. H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics*, *24*, 225–243. https://doi.org/10.3102/107699860240032

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. E. Sobel & M. P. Becker (Eds.), *Sociological methodology 2000* (pp. 165–200). Washington, DC: American Sociological Association.