The Journal of English Language Pedagogy and Practice
Vol. 16, No.32, Spring and Summer 2023(101-117)
DOI: 10.30495/jal.2023.1988106.1495

Research Article

Applying Structural Equation Modeling to Second-language (L2) Research: Key Concepts and Fundamental Reconsiderations

Hessameddin Ghanbar
Department of Languages and Linguistics, Fereshtegaan International Branch, Islamic Azad University, Tehran, Iran
Email: Hessameddin.ghanbar@iau.ac.ir
(Received: 2023/06/06; Accepted: 2023/11/19)

Online publication: 2023/12/09

Abstract

As conceptual models of language learning, use, and processing mature, it is both natural and necessary for the statistical models we apply to follow suit. One statistical approach with great potential in the field of Applied Linguistics and L2 studies is structural equation modeling (SEM). SEM is introduced in this review paper as a powerful and highly flexible family of analyses. In doing so, the paper outlines (a) the types of variables and possible modeled relationships that SEM is equipped to address and (b) statistical considerations for applying SEM in L2 research, and (c) a number of additional and key considerations for those interested in delving deeper into SEM (e.g., goodness of fit indices, model modification procedures, etc.). This paper also describes the potential of SEM to contribute to construct validation (e.g., convergent and discriminant validity). Throughout the paper, a plethora of examples pertaining to applications of SEM in L2 research are provided.

***Keywords*:** structural equation modeling, multivariate data analysis, L2 research, quantitative research methods, advanced statistics

Introduction

The statistical repertoire in second-language (L2) research has, without a doubt, expanded in recent years (see, e.g., Gass, Loewen and Plonsky, 2021; Khany and Tazik, 2019; Author & Other, 2023). However, as theoretical models mature, offering potentially greater insight into the strength and nature of many different relationships of interest, it is only natural that we require correspondingly more advanced and complex statistical models. Structural equation modeling (SEM) provides a powerful approach to doing so. Despite of the versatility of this approach, two most recent and comprehensive systematic reviews conducted on SEM practices (In'nami & Koizumi, 2011; Ghanbar & Rezvani, 2023) in the field pointed to several methodological shortcomings and gaps. For example, Ghanbar and Rezvani (2023) revealed that more than 53 % of SEM practices which were examined (383 out of the total of 722 SEM practices investigated) did not provide information about normality of data submitted to SEM. The situation was worse when it came to checking other statistical consideration such as missing data, outliers and linearity. There were also other questionable practices reported in Ghanbar and Rezvani (2023) pertaining to the type of relationships in models (e.g., overreliance merely on one type of model set-up), sample size, model specification issues (e.g., sticking to merely one type of model specification), model estimation methods (e.g., utilizing solely one type of method despite of having many other versatile model estimation techniques) and also reporting practices relating to different model parameters and goodness-of-fit (goodness of models). Hence, the main motivation behind this method note article is to create a snapshot of SEM through outlining and synthesizing basic conceptualizations and issues relating to different steps in this technique, from model setup to estimation and evaluation of models, which can play a pivotal role in boosting the methodological rigor and transparency of SEM practices in the future works. In each of the following sections, one SEM issue was sketched wherein important relevant notions and points were recapitulated, in accompanying with several useful and updated sources for L2 researchers who aim to use SEM in future studies.

SEM: The Basics Concepts and Modeling Issues

SEM is perhaps best conceived of as a family of statistical techniques used to investigate a wide range of both causal and correlational links. SEM is also quite flexible, allowing researchers to examine many such links simultaneously, taking in and modeling relationships among different types of variables. Similar to other analyses within the general linear model (e.g., correlation, ANOVA), SEM can, of course, handle (a) observed variables such as measures of speech decoding and reading comprehension. In such

cases, SEM can be considered an extension of multiple regression which is called path analysis (see Janssen, Segers, McQueen, and Verhoeven, 2016). However, SEM can also be used to analyze (b) latent variables (i.e., variables that are not directly observable or measurable with a single score), such as second-language (L2) proficiency and aptitude and hence should be measured by their observed variables; these modeled relationships are called latent variable path analysis or structural model which is explained later in the paper. In addition, SEM can simultaneously address both (c) observed and latent variables relationships which is called in SEM literature 'measurement model'. And finally, one can and will often apply all these options (mixing [b] and [c] which is called a full SEM), as generally in bilingualism research have been exploited in tandem. Put differently, a full SEM can be viewed as a combination of a measurement model and a structural model. Before going further, these terms will be unpacked a bit.

A measurement model, which can be conceived of as a confirmatory factor analysis (CFA), can be defined as a statistical model of relationships between a construct (circles in Figure 1, which cannot be directly measured) and its related measured variables (See Tabachnick & Fidell, 2019 for elaborated discussions on types of factor analytic methods including CFA). It is through measured variables (also called 'observed variables or 'indicators', such as scores on a test of working memory [WM], depicted as squares in Figure 1) that we come to understand latent variables.

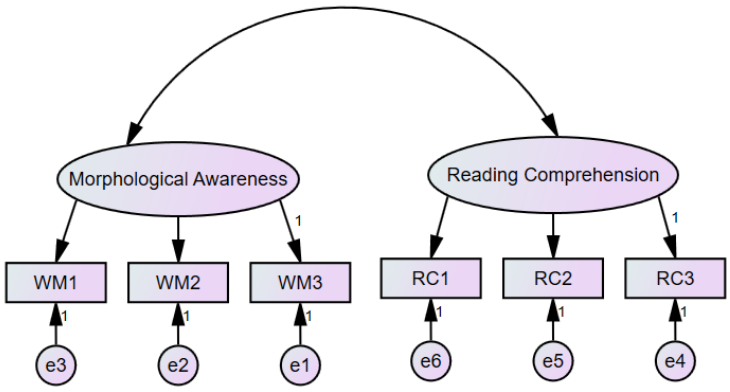

Figure1: *Path Diagram of a Reflective Measurement Model in SEM*

In Figure 1, for example, one can examine the relationship between indicators (e.g., WM1, WM2, and WM3) and morphological awareness (construct). This type of model is referred to as a CFA. The structural portion of the model (Figure 2) also represents a causal relationship between two

latent variables, morphological awareness and reading comprehension, as indicated by a one-headed arrow and in SEM these casual relationships would be statistically tested.

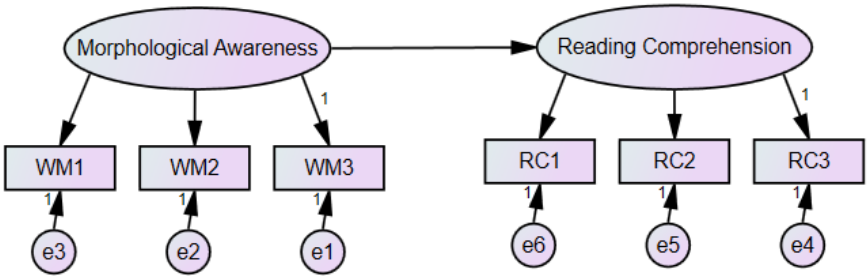

Figure 2: *Path Diagram of Structural Model in SEM*

For example, de Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012) presents a two-phase modeling (for analyzing a structural model two-phase modeling can be used [Kline, 2016], in that first the structural model first is re-specified as a CFA measurement model and if the data fit this model well, the second phase, structural phase begins), comprising measurement and structural models to investigate a componential view of L2 speaking proficiency that consists of language knowledge and language-processing components.

The relationship between observed variables and latent variables in SEM can be conceived of in two forms: (a) reflective and (b) formative. The reflective model (Figure 1) has its root in classical test theory (CTT) in that all the indicators are representing the effect of their corresponding latent variable (Hair, Hult, Ringle, & Sarstedt, 2017). In that model specification, a construct is viewed as the common cause of its indicators. In a reflective model, all the indicators of a construct should be highly correlated since they are considered a sample of potential indictors of that construct (i.e., each item can be removed without any major loss of meaning in the pertinent construct).

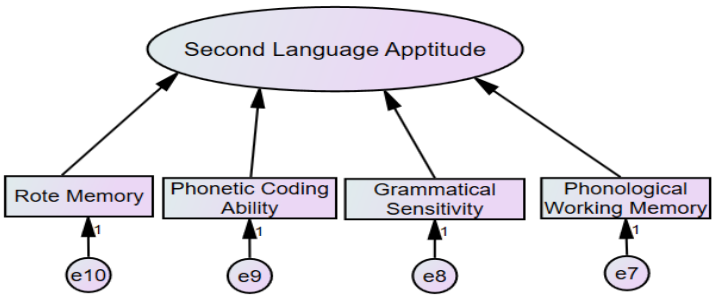

Figure 3: *A Formative Construct with its Indicators*

In formative model, shown in Figure 3, however, the direction of the relationship is reversed, given that, here, indicators cause (form) the construct (it is also referred to as an 'emergent factor'). Furthermore, in a formative model each indicator represents a unique aspect of the construct as in Figure 3, each indicator taps one aspect of self-perception of language skills (Diamantopoulos & Siguaw, 2006, Hair, Hult, Ringle, & Sarstedt, 2017). Contrary to a reflective model, the indicators in a formative model are neither interchangeable nor can be deleted without any loss of meaning in the construct (Diamantopoulos, Riefler, & Roth, 2008) and this meaning loss varies on the basis of the number of indicators as well (in case of subtests, each of them also needs to tap one different aspect of the construct and bilingualism researchers should first determine this point). For example, in our example in Figure 3 (adapted from Winke, 2013) cause indicators tap different aspects of aptitude, and hence removing one of them resulting losing information regarding that aspect of the construct. Critically, conceptualizing these two relationships between a construct and its indicators (reflective or formative) should be grounded in theory or some statistical considerations like collinearity (See Hair, Hult, Ringle, & Sarstedt, 2017 for specific guidelines to this conceptualization as the formative constructs have been rarely seen in L2 research, despite the fact that it can have potentials for better representing L2 constructs).

Contrary to the widely held belief that SEM can only be used for confirming a model or theory, this technique is very robust in exploring or even developing theory. More specifically, if the aim of a study is to test/confirm a hypothesis, a reflective model is utilized. However, if the goal is to find the best indicators (aspects) of a construct (i.e., exploring a theory), the formative model is considered more appropriate.

Another distinction pertains to latent variables in the structural portion of SEM. As can be seen in Figure 1, morphological awareness affects reading comprehension, and therefore the former is considered exogenous and the latter endogenous (as indicated by the arrow pointing toward it). Exogenous variables are analogous to predictors in multiple regression (see Plonsky & Ghanbar for more information about multiple regression) in that they impact endogenous variables, which, in turn, are considered dependent or criterion variables. However, it should be emphasized that exogeneity is based on this assumption that an exogenous variable is not affected by any unobserved confounding variables. This is a very strong assumption and overlooking it might result in biased estimated parameters, so L2 researchers should carefully consider confounds in their studies and elaborate on them in limitations of their studies.

Statistical Considerations and Assumptions in SEM

Several statistical requirements should be met for SEM to yield accurate and credible results. First, sample size is of prime importance in SEM, as it is a multidimensional issue, necessitating consideration of such factors as estimation methods (techniques) for calculating regression coefficients in a model's statistical power, model complexity (having many constructs and their accompanying indicators in a model would necessitate a larger sample size), reliability of indicators, and expected R^2 values. Further, it should be highlighted that a priori decisions about sample size are very difficult as often-cited rules of thumb fail to capture the complex nature of any given SEM (see Byrne, 2016; Kline, 2016; Mueller & Hancock, 2019; Raykov & Marcoulides, 2006, for a more detailed discussion on sample size and other statistical issues in SEM).

There are several other important statistical assumptions and concerns related to data structure such as normality (univariate and multivariate), linearity, absence of singularity and multicollinearity, and finally missing values as well as outliers (univariate and multivariate). Overlooking normality, for example, can adversely impact both generated fit indices and estimated model parameters alongside their standard errors (see Ockey & Choi, 2015 and Lei & Wu, 2012 for some recommendations and techniques in non-normality situations), threatening the accuracy of the final results (Byrne, 2016; Pituch & Stevens, 2016).

From Model Specification to Model Estimation

The main aim of this section is to complement what was expounded upon in the previous section through discussing the issues relating to the number of indicators in each construct of a model and the type of structural relationships between constructs in SEM. Furthermore, the main agendas and fundamental considerations in other two successive stages of SEM, that is, model identification and estimation stages, which were followed after model specification stage, were explicated in what follows.

As described above, SEM yields both a measurement model and a structural model. Simply specifying and setting up these two models, however, is not sufficient. Of paramount significance is specifying the relationships between indicators and constructs in the measurement portion of a SEM on the one hand (see the previous section), and constructs with other constructs in its structural counterpart on the other. Regarding the measurement model, as discussed before, choosing between reflective indictors and formative ones should always be based on the literature and solid theory. One crucial consideration in many scale validation studies is the number of indicators (items) per construct (Kline, 2016).

Kline (2016) mentioned that the minimum number of indicators per factor in CFA models with more than two factors is two but he also mentioned that this minimum number might result in technical problems in data analysis so he proposed three to five indicator per factor criterion. Some scholars have argued that adding more indicators might hinder model estimation (e.g., Hair, Hult, Ringle & Sarstedt, 2017; Kline, 2016). However, additional indicators might not impede specification even in relatively smaller samples (Mueller and Hancock, 2019). For instance, as Mueller and Hancock (2019) mentioned, having four to six indicators with their standardized loadings (standardized parameters [standardized coefficient] in a model) larger than 0.6 or 0.7 is an ideal situation. Furthermore, in the structural portion, literature and theory should again inform both the types of constructs (i.e., exogenous or endogenous), direct effects (i.e., when two constructs are linked with a single arrow between them, see Figure 4 for the direct effect of self-confidence on attitude toward L2 speakers), and also indirect effects (i.e., a sequence of relationships in the structural model with at least one intervening construct in the model, see Figure 5 for both the direct effect of self-confidence on attitude toward L2 speakers and also the indirect effect of self-confidence on attitude toward L2 speakers via cultural interest, also see Kline, 2016 for a detailed discussion on effects in structural models), contributing to the model's *nomological validity* (nomological validity determines whether your modeled relationships are based on theory or prior research), the essence of construct validity (Cronbach & Meehl, 1955). As a more general principle, all model setup decisions in the specification phase can have a major impact on model identification and estimation phases.

Figure 4*: Direct Effect of Self-confidence on Attitude Towards L2 Speakers*

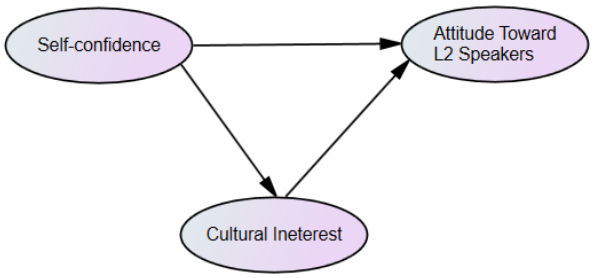

Figure 5: *Direct and Indirect Effect of Self-confidence on Attitude towards L2 Speakers*

Identification and estimation stages, implementing after model specification stage, are two interlocking steps in SEM, as under-identified models—models without enough measured variables (known information or data points) in proportion to unknown parameters (parameters to be estimated) such as variances, covariances, structural coefficients and factor loadings—cannot be estimated. For example, in a model with 14 observed variables, we will have k (k+1) data points, so we have 210 such elements. Hence, this model, with 215 parameters to be estimated and 210 data points (calculated above), is under-identified implying that we do not have information to estimate the unknown parameters (see Kenny & Milan, 2012 for a detailed discussion on model specification and identification issues). Accordingly, as emphasized before, a theoretically-grounded specification stage helps to ensure model identification and estimation.

Pertinent to the model estimation step, in which all the unknown parameters are estimated from known ones, there exists a wide array of methods, such as maximum likelihood (ML) and full-information maximum likelihood (FIML), with each being well-suited to specific conditions (see Lei & Wu, 2012 for an elaborated discussion on different estimation methods and their working conditions). For example, as missing data points impede model estimation in SEM, FIMIL can be used by L2 researchers in this situation. To close this section, it should be said that choosing among estimation methods has to be grounded upon several statistical criteria such as sample size, metrics, and distribution of the variables, all of which should be made clear in the manuscript (see Ghanbar & Rezvani, 2023 for more information about features and functionalities of each estimation method). For example, ML requires normal distribution, and hence, not very large sample is needed to yield accurate parameter estimates, however, asymptotic distribution free (ADF), another estimation method, which is used for non-normal distribution data requires larger samples (Mueller & Hancock, 2019)

Quality Criteria and Goodness-of-Fit in SEM

The evaluation of SEM results can be undertaken utilizing three different types of criteria: (a) theoretical (a model should be based on a theory, without including irrelevant variables or omitting important ones), (b) statistical (reasonableness of parameters, that is, not having any negative variances, out of/beyond range correlation coefficients, and high standard errors), and (c) goodness-of-fit (GoF).

Nontechnically, GoF indicates the extent to which the hypothesized model approximates/fits the observed data. In an instrument validation study, GoF can be conceived of as the extent to which the instrument accurately and

thoroughly represents the participants' responses). Several groups of GoF indices have been proposed to take into account different aspects of it (see Byrne, 2016; Mueller & Hancock, 2019; Pituch & Stevens, 2016; and Ockey & Choi, 2015 for a list and full discussion on the functions of different sets of fit indices).

The first recommended type of fit indices, basic fit indices, shed light on the overall fit of the theoretical model (e.g., X^2 and X^2/df). This important point should be added that X^2/df like X^2 is dependent on sample size and also it penalizes for model complexity, in that it gets worse when more parameters are added to a model (West, Tylor, & Wu, 2012).The second, incremental or comparative fit indices, measure the proportionate improvement in a proposed model relative to an independence or baseline model (i.e., a model in which no correlations are assumed among variables as SEM, most often, is based on variance-covariance modeling and thus usually models correlations of variables). Absolute fit indices, the third category, investigate the extent to which a model is successful in reproducing the reality, resembling an R^2 value in multiple regression. Finally, residual-based fit indices examine the average differences between what a model estimates and what is happening in the data (predictive fit indices can also be considered in this section but they are used only to compare GoF of competing non-nested models). The bottom line is that a combination of these fit indices should be reported given that each considers a unique aspect of model-fit. (See further related discussion in Fan, Thompson, & Wang, 1999, and see Hu & Bentler, 1999, for guidelines on interpreting GoF indices; also refer to Byrne, 2016; Kline, 2016; Ockey & Choi, 2015 for a list and a detailed discussion on each of the aforementioned fit indices). Also, this precaution would be given to L2 researchers that the cut-off values proposed for fit indices should be used with a caution, as these values vary on the basis of different conditions such as types of misspecifications and number of latent variables as well as indicators (see Heene, Hilbert, Freudenthaler, & Bühner, 2012 for a discussion on the sensitivity of commonly used cutoff values of global-model-fit indexes, with regard to different degrees of violations of the assumption of uncorrelated error).

If a measurement model fits the data well, the researcher moves on to the structural phase of modeling. If, however, the model does not fit the data, internal specification errors (i.e., inclusion of unimportant paths or exclusion of important paths) might be corrected utilizing three common methods. First, chi-square difference test can be used to compare nested models' fit (this is not the mere function of this test and see Byrne, 2016 for a discussions on its further functions in SEM), that is, models in which a simpler model's parameters are a subset of a more complex model, as sometimes L2

researchers aim to compare a priori alternative models with the target model, yielding support for its plausibility (see Henry & Cliffordson, 2013 for an application of chi-square difference test in second language motivation research). Second, Lagrange multiplier tests or modification indices are used to assess fit improvement, if several parameters are added, which is similar to forward stepwise regression. And third, Wald tests evaluate fit improvement, if several paths are deleted, which is similar to backward deletion stepwise regression (see Others and author, 2018 for discussion on these and other types of regression analysis and also see Stevens, 2009 for a detailed discussion on Lagrange multiplier and Wald tests).

Construct Validation of Scales Using SEM in L2 Research

Instrument (scale) validity can be conceived of as how thoroughly (content validity), and accurately (construct validity) (Bachman, 1990; Hair, Hult, Ringle, & Sarstedt, 2017) a given tool taps a psychological, linguistic, or psycholinguistic construct (e.g., listening comprehension, reading comprehension, phonological memory, phonemic awareness), as well as how efficient it is in predicting/explaining other important constructs (criterion validity). Given that the evaluation of criterion validity is undertaken using bivariate statistical procedures (see Teng, Sun, & Xu, 2018 as an example of its use in a validation study using SEM), ways in which content and construct validity as well as reliability of constructs in different scales (or survey scales) can be assessed via SEM will be discussed in what follows. Indeed, the focus here on validating reflective constructs which are more prevalent than formative constructs in L2 research. In what follows different types of construct validity are defined. Then, explanations are provided on how L2 researchers can investigate construct validity of scales via multivariate statistical techniques such as exploratory factor analysis (EFA) and SEM.

The first type of validity, content validity, is conceptually defined as the extent to which different aspects of a construct are represented by the items of an instrument (here in this study we merely focus of scales) (Bachman, 1990). Thus, in order to enhance the accuracy of construct's measurement, several items are generally used for different constructs in EFA or SEM. Scheele, Leseman and Mayo (2010) examined the relationships among socioeconomic status (SES) as well as L1 and L2 abilities. Rather than including a single measure of L1 input, the authors included several indicators (measured variables) in the model including time spent reading and storytelling in the L1.

Many studies evaluate content validity through expert opinion. However, such an approach is often not sufficient. This paper recommends that L2

researchers use procedures such as EFA to investigate content validity (Bryant, 2000). This is particularly necessary when little or no information is available on issues such as the number of underlying constructs in an instrument (the preliminary factorial structure), the number of indicators constituting each construct and how strongly they represent it, and the magnitude of correlations among underlying constructs. To conclude, when a construct and its newly developed items/indicators are at hand, it is best to start with EFA, which is more flexible, since there are a variety of extraction and rotation methods available in EFA. It needs to be mentioned that Using CFA through SEM might result in high degree of misfit, that is, low GoF indices. Then, if a clear, interpretable factor structure emerges, SEM with a new, independent sample can be carried out. Using EFA and CFA successively in the same sample, a common methodological flaw in L2 research (Bandalos & Finney, 2019), however, is not recommended. As Bandalos and Finney (2019) recommended, this practice capitalizes on chance and relies too heavily on sample-specific idiosyncrasies. Nonetheless, when GoF indices are low, using EFA after conducting CFA is suggested (see Hu, 2005, for an L2-specific example of an EFA of a newly developed scale and its consecutive CFA applied through SEM, without using an independent sample; see also Polat & Cepik, 2016, for an appropriate use of EFA to examine the factorial structure of the widely used sheltered instruction observation protocol [SIOP]).

In contrast to content validity, evaluating construct validity is complex and multifaceted, presenting a critical yet often overlooked challenge to applied linguistics researchers. At stake in this phase of a study is establishing the conceptual accuracy of measurement, and SEM provides a robust framework for doing so.

Construct validity can be operationally defined as the extent to which a construct (e.g., phonemic awareness) is meaningfully (conceptual adequacy) and accurately measured by its measures (indicators) (Bachman, 1990; Mueller & Knapp, 2019). As noted before and to emphasize here, initially, preoperational explication (*nomological validity*), which is considered a consequential prerequisite of construct validity should be presented (Bryant, 2000; Fulcher & Davidson, 2007). This is of the utmost significance since adding irrelevant variables to and removing important variables from a model result in external specification errors, which, in turn, threaten the accuracy of results. After investigating nomological validity, construct validity evaluation, via information provided in a SEM program's output, is conducted in two phases: (a) convergent validity, and (b) discriminant validity.

Convergent validity is the extent to which indicators of a construct (e.g., measures of phonemic recognition, phonemic blending, phonemic judgment and phonemic counting as indicators of phonemic awareness) converge (concur) in a measurement model of SEM, that is, share a high amount of common variance (Bryant, 2000; Hair, Hult, Ringle, & Sarstedt, 2017). More particularly, each item loading (i.e., standardized regression weights in SEM output) should be higher than .708. The rationale behind setting this value is that, in order to guarantee convergent validity, average variance extracted (AVE), defined as the grand mean value of squared items' loadings of a construct, should be .5 or higher (see Hair, Black, Babin, & Anderson, 2010). Items with loadings less than .4 should be deleted and those with loadings between .4 and .7 should be deleted if their removal boosts AVE and composite reliability estimate (CR). CR is an indicator of how well the observed variables (indicators) associated with a latent construct or factor are related to each other (see Ghanbar & Rezvani, 2023 and Hair, Hult, Ringle, & Sarstedt, 2017 for a more technical discussion). In other words, it measures the extent to which the observed variables collectively reflect the underlying latent construct. It should be noted that a higher composite reliability score indicates a stronger and more consistent relationship among the observed variables, which, in turn, suggests better reliability of the latent construct.

Discriminant (divergent) validity, on the other hand, suggests the extent to which a construct is made distinct from other constructs by its items/indicators (Hair, Hult, Ringle, & Sarstedt, 2017). This can be assessed in two ways: (a) cross-loadings of items (i.e., standardized regression weights of a construct's items should be higher than their weights on other constructs), and (b) *Fornell-Larcker criterion*, which stipulates that the square root of each construct's AVE should be higher than its correlation with other constructs (as Hair et al., 2017 proposed, discriminant validity can also be evaluated in another way, namely, that the AVE of a construct should be higher than both its maximum shared variance [MSV] and average shared variance [ASV]). All these measures are provided for L2 researchers in the output of SEM software such as AMOS and MPlus (they can also be generated via Excel macros). L2 researchers should be cautious that substantial overlaps between items of different constructs in an instrument would threaten discriminant validity, requiring meticulous attention to both nomological validity and wording of items (see Teng et al. 2018, and also Hiver & Al-Hoorie, 2020 for the application aforementioned construct validation techniques in a recent SEM study in L2 research).

There are a number of more advanced issues pertaining to construct validation which are largely outside the scope of this brief conceptual

introduction but which bear mentioning. The first is the use of more complex techniques, such as multitrait-multimethod modeling (MTMM; Campbell and Fiske, 1956) to investigate the construct validity of an instrument. (See Byrne, 2016, for a full chapter on conducting MTMM, and also Llosa, 2007, for an example of MTMM design in the context of L2 research). The second issue concerns the capacity of SEM for examining construct validity across different populations (e.g., comparing item loadings across different genders, for example, or between heritage and L2 learners). This can be done using special type of SEM, called multigroup invariance or multigroup analysis (SEM-MGA), and is more robust than comparing items' means across groups of interests. See Shiotsu and Weir (2007) as an example of SEM-MGA to examine the relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance in a heterogeneous population.

The final note is about construct reliability. Cronbach alpha is reported as an index of reliability in the vast majority of SEM studies both within and outside of the language sciences. Nevertheless, this index assumes tau equivalence (McNeish, 2018), that is, they all have the same regression coefficients on a construct, which is not a realistic presumption. Alpha is also very sensitive to the number of items in an instrument. Consequently, this paper recommends CR which is based on different magnitudes of regression weights of a construct (values between .6 and .9 are considered acceptable, see Nunally and Bernstein, 1994). Additionally, this article recommends maximal reliability estimates such as *Coefficient H* (Raykov, Gabler, & Dimitrov, 2016), as it reflects the correlation that a factor is predicted to have with itself over repeated measurements, showing the stability of a construct, with values more than .7 considered acceptable (Mueller & Hancock, 2019).

Conclusion

In writing this research note, I set out to accomplish two goals. The first was to discuss and expound upon the *potentials* of SEM regarding the community of L2 researchers. In order to accomplish this goal, I discussed a number of different types of variables, relationships, and models that can be explored using SEM in accompanying with running example studies conducted in the field. This paper also aimed to make known some of the many considerations that come into play when applying SEM, complemented by providing recommendations and suggestions to enhance methodological rigor and transparency of SEM. To be sure, here this conceptual paper just has only scratched the surface. Nevertheless, this paper strived to make L2

researchers cognizant of the potentials of SEM to inform and advance their future practices of this technique.

Declaration of interest: none

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bandalos, D. L., Finney, S. J. (2019). Factor analysis: Exploratory and confirmatory. In G. R. Hancock, R. O. Mueller, & L. M. Stapleton (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 98-122). New York, NY: Routledge.

Bryant, F. B. (2000). Assessing the validity of measurement. In L. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 99-146). Washington, DC, US: American Psychological Association.

Byrne, B. M. (2016). *Structural equation modeling with AMOS*: *Basic concepts, applications, and programming* (3th 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*(2), 81-105. https://doi.org/10.1037/h0046016

Cronbach, L. J., & Meehl, P. C. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative vs. reflective indicators in measure development: Does the choice of indicators matter? *British Journal of Management*, 13, 263-282. https://doi.org /10.1111 /j.1467-8551.2006.00500.x

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*, 1203-1218. https://doi.org/10.1016/j.jbusres.2008.01.009

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5-34. https://doi.org/10.1017/S0272263111000489

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 56-83. https://doi.org/10.1080/10705519909540119.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York, NY: Routledge.

Gass, S., Loewen, S., & Plonsky, L. (2021). Coming of age: the past, present, and future of quantitative SLA research. *Language Teaching, 54(2),* 245-258. https://doi.org/10.1017/S0261444819000430

Ghanbar, H., & Rezvani, R. (2023). Structural Equation Modeling in L2 Research: A Systematic Review. *International Journal of Language Testing, 13*(Special Issue), 79-108. 10.22034/ijlt.2023.381619.1224

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

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.

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling, 19*(1), 36-50. https://doi.org/10.1080 /10705511 .2012.634710

Henry, A., & Cliffordson, C. (2013), Motivation, Gender, and Possible Selves. *Language Learning, 63*, 271-295. https://doi.org /10.1111 /lang. 12009

Hiver, P., & Al-Hoorie, A. H. (2019). Reexamining the role of vision in second language motivation: A preregistered conceptual replication of You, Dörnyei, and Csizér (2016). *Language Learning*. https://doi.org /10.1111/lang.12371

Hu, G. (2005). Contextual influences on instructional practices: A Chinese case for an ecological Approach to ELT. *TESOL Quarterly, 39*(4), 635-660. https://doi.org/10.2307/3588525

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

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

Janssen, C., Segers, E., McQueen, J. M., & Verhoeven, L. (2017). Transfer from implicit to explicit phonological abilities in first and second language learners. *Bilingualism: Language and Cognition, 20*(4), 795-812. https://doi.org/10.1017/S1366728916000523

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

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, US: The Guilford Press.

Llosa, L. (2007). Validating a standards-based classroom assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24*(4), 489-515. https://doi.org/10.1177/0265532207080770

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*, 412-433. doi:10.1037/met0000144

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.

Mueller, R. O., & Knapp, T. R. (2019). Reliability and validity. In G. R. Hancock, R. O. Mueller, & L. M. Stapleton (Eds.), The reviewer's guide to quantitative methods in the social sciences (pp. 397-401). New York, NY: Routledge.

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

Ockey, G. J., & Choi, I. (2015). Structural equation modeling reporting practices for language assessment. Language Assessment Quarterly, 12(3), 305-319.

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

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

Polat, N., & Cepik, S. (2016). An exploratory factor analysis of the sheltered instruction observation protocol as an evaluation tool to measure teaching effectiveness. *TESOL Quarterly, 50*(4), 817-843. https://doi.org/ 10.1002 /tesq.248

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

Scheele, A. F., Leseman, P. P. M., & Mayo, A. Y. (2010). The home language environment of monolingual and bilingual children and their language proficiency. *Applied Psycholinguistics, 31*, 117-140. https://doi.org /10.1017/S0142716409990191

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing, 24*(1), 99-128. https://doi.org /10.1177/0265532207071513

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

Tabachnik, B. G., & Fidell, L. S. (2019). Using multivariate Statistics (7th ed.). Boston: Allyn and Bacon.

Teng, L. S., Sun, P. P., & Xu, L. (2018). Conceptualizing writing self-Efficacy in English as a foreign language context: Scale validation through structural equation modeling. *TESOL Quarterly, 52*(4), 911-942. https://doi.org/10.1002/tesq.432

West, S.G., Taylor, A.B., &Wu, W. (2012). Model fit and model selection in structural equation modeling. In Hoyle RH (ed.) *Handbook of structural equation modeling* (pp. 209-231). New York, NY, US: The Guilford Press.

Winke, P. (2013). An investigation into second language aptitude for advanced Chinese language learning. *The Modern Language Journal* 97(1): 109-130. https://doi.org/10.1111/j.1540-4781.2013.01428.x

Biodata

Hessameddin Ghanbar is an Assistant Professor of Applied Linguistics at Islamic Azad University, Fereshtegaan International branch, Tehran, Iran. His areas of interest include meta-analysis and research synthesis in L2 research. His recent research syntheses appeared in *Modern Language Journal* in 2018, *Journal of English for Academic Purposes* in 2020, *Language Learning* in 2021, *Studies in Second Language Learning and Teaching* in 2022, and *Research Methods in Applied Linguistics* in 2023.