CrossMark

Annual Review of Applied Linguistics (2014), **34**, 102–122.
© Cambridge University Press, 2014, 0267-1905/14 $16.00
doi: 10.1017/S0267190514000075

Testing Hypotheses about Language Learning Using Structural Equation Modeling

Paula Winke

In this chapter, I review a statistical method for hypothesis or theory testing called *structural equation modeling* (SEM). First, I describe what a model of second language acquisition (SLA) is. I do this so anyone, even those new to the field of applied linguistics, can understand the basic concepts underlying SEM; that is, SEM researchers first articulate a model of SLA, then get empirical data from the real world that operationalize the variables in the model. Researchers use an SEM program to test the model on the data (to see if the model fits the data; if the model is plausible in relation to the learning context of the people from whom the data were collected). After explaining the basics of SEM, I provide a review of 39 applied linguistics studies that have been published in the last five years (between 2008 and 2013) and that present at least one SEM analysis as part of the results. I discuss four problematic areas related to the use of SEM that I believe these 39 studies highlighted: (a) sample size, (b) model presentation, (c) reliability, and (d) the number of Likert-scale points. I conclude with possible solutions for the four problem areas and outline future directions.

TESTING HYPOTHESES ABOUT LANGUAGE LEARNING USING STRUCTURAL EQUATION MODELING

For more than 15 years, researchers in second language acquisition (SLA) have written that structural equation modeling (SEM) is an opportune research tool for investigating the likelihood and strengths of the field's theories and hypotheses (In'nami & Koizumi, 2011; Isemonger, 2007; Kunnan, 1998, 1999, 2010). SLA researchers have proposed that SEM is good for investigating fairness in language testing (Kunnan, 2010), for measuring the effects of different instructional types on learning (In'nami & Koizumi, 2011), for investigating what influences bilingual development (Lewis & Vladeanu, 2006), and for measuring the effects of knowledge structures (such as metalinguistic knowledge) that impact second language (L2) performance (Isemonger, 2007). Prior to 2008, SLA researchers used SEM and demonstrated its great potential for L2-testing research (In'nami, 2006; Purpura, 1998; Schoonen, 2005; Turner, 1989), individual differences (IDs) research (Csizér & Dörnyei, 2005; Gardner, Tremblay, & Masgoret, 1997;

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

Matsumura, 2003; Sasaki, 1993; Tremblay & Gardner, 1995; Woodrow, 2006; Yashima, 2002), investigations into the effects of knowledge structures (Schoonen et al., 2003), and questions about bilingual development (e.g., Verhoeven, 2007). Within the last five years, the pace at which SEM is being applied to SLA data has increased. SEM is becoming a more frequently used tool, appearing in a wider variety of SLA journals, and it is used with more varied contexts of language learning, as I will demonstrate.

Using SEM for data analysis is becoming more popular in SLA most likely because SLA theories are coming of age. They are more sophisticated and complex, with proposed associations among multiple, complex variables, necessitating the need for robust model testing. Two requirements for SEM are more often being achieved. The first is reliable data collection tools. Many of the field's data collection instruments have been tested over time and have been proven to be reliable and valid, making the use of SEM possible: SEM takes measurement error into account when a model is tested (Kline, 2012; Schumacker & Lomax, 2010), so little measurement error (high reliability) in the tests or surveys used to operationalize the traits in the model (often demonstrated via high Cronbach's alpha coefficients) is essential for model testing with SEM. The second requirement is large participant numbers. How large is debated, especially for the field of SLA (In'nami & Koizumi, 2011; Winke, 2013), but in general, SEM "is a large-sample technique" (Kline, 2012, p. 11), and the more complex the model, the more participants the testing of that model requires. Further augmenting the use of SEM, as noted by Schumacker and Lomax (2010), is that the software packages for SEM have become more user friendly. After a class on SEM or an in-depth workshop, even statistically adverse researchers can employ SEM by using statistical packages such as IBM's SPSS Amos (http://www-03.ibm.com/software/products/en/spss-amos). Thus it is time to take a step back and look at this statistical technique as it is being used in SLA. What are the overarching patterns of use, and how can researchers continue to use it successfully?

In this article, I first briefly define SEM in a way that novice SLA researchers can understand the basic concepts underlying SEM. Then I summarize how SEM has been used over the last five years (between 2008 and 2013) in the field of SLA, and I review what I see as SLA researchers' main problems in using SEM. In connection with these problems, I propose some possible solutions and methodological directions for future researchers who use this data-analysis technique.

A SIMPLE INTRODUCTION TO SEM

One does not have to be an applied linguist to have an opinion about what it takes to learn an L2. Ask anyone on the street what is important, and he or she might say, for example, that three things are needed: (a) One has to be good at learning languages (i.e., have an aptitude for language learning), (b) be dedicated to learning the language (i.e., put time and effort into learning), and (c) spend time abroad in a country where the language is spoken. So in a way, this person on the street has articulated an imagined model of what it takes to learn a new language.

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

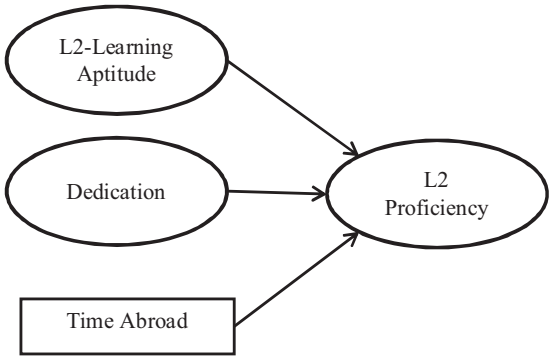


Fig. 1. A simple hypothesized model or path of L2 learning with both latent and manifest variables.

In other words, he or she specified a model that can be estimated and tested, which is the first step in an application of SEM (Hoyle, 2012b). This particular hypothesized model, if drawn out, might look like what is presented in Figure 1. Applied linguists familiar with the conventions of modeling (e.g., Byrne, 2010; Ho, Stark, & Chernyshenko, 2012; Kline, 2011; Schumacker & Lomax, 2010) will recognize the variables in the ovals are latent traits, variables that exist but that cannot be measured directly (such as L2-learning aptitude and dedication, which are as of yet undefined in this model and therefore hidden), while the variables in rectangles are indicator or manifest variables, things that are directly measured (e.g., the amount of time abroad can simply be measured in months; see Bollen & Hoyle, 2012, for more on latent variables.) The applied linguist wanting to test this model (see if the model fits with data) is looking to see how the variables on the left (the independent variables) affect or influence the variable on the right (the dependent variable). The cause and effect relationships (direct effects; regression coefficients) are symbolized with single-headed arrows, and these effects are what are equated through SEM (see Pearl, 2012, for a discussion of the cause and effect foundations in SEM).

The model in Figure 1 is simple. As researchers ask more questions about the model and further define its variables and the relationships among them, the model becomes more complete (testable), but also more complicated, as shown in Figure 2. L2 proficiency, it might be decided, will be measured through two performance tests (speaking and writing) and one grammaticality judgment test. The applied linguist might suppose a relationship between two of the independent variables: Learners who are more dedicated will spend time abroad and vice versa. Such a relationship is a hypothesized correlation, symbolized with a curved, double-headed arrow (Schumacker & Lomax, 2010, see chapter 3 on correlations). L2-learning aptitude might be defined through three composite measures from the Computer-Based Modern Language Aptitude Test (CB-MLAT; see the Language Learning and Testing Foundation website for information on the test at http://lltf.net/): the Number Learning test (which measures, theoretically, phonetic coding ability); the Words in Sentences test (which assesses

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

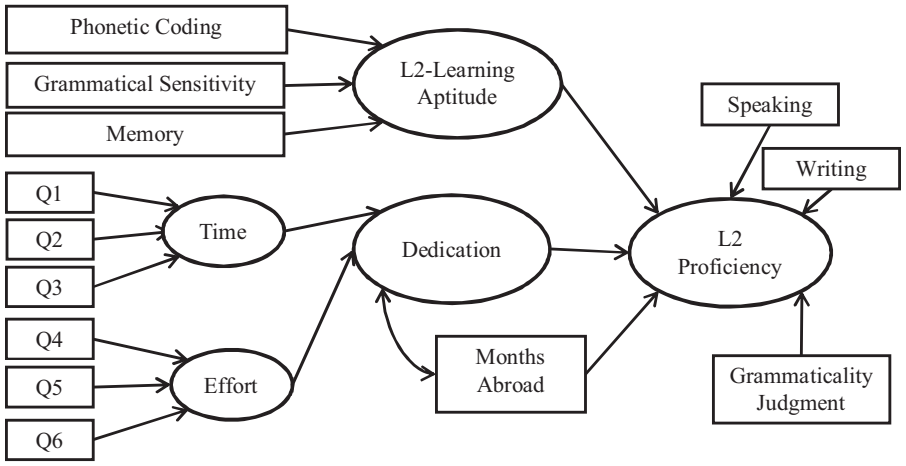


Fig. 2. A hypothesized model of L2 learning. This is a full latent variable (structural) model that can be tested with SEM.

grammatical sensitivity); and the Paired Associates test (a test of memory; Skehan, 2002). Dedication might be hypothesized as made up of two separate concepts (or factors). The first might be time spent on average per week doing three things in the L2 (movie watching, book reading, and talking; measured through three questionnaire questions). The second factor might be self-estimated levels of effort in studying vocabulary, grammar, and writing conventions (measured through three additional questions given on the same questionnaire). This part of the measurement model will be subjected to *confirmatory factor analysis* or CFA (see Henry & Cliffordson, 2013; Kline, 2011; Turner, 1989) because the model suggests that the six items subsumed under *dedication* actually break down into two independent components or factors.

As plotted in Figure 2, the applied linguist has now created a complex structure of variables and their hypothesized relationships. Theoretically, the model can be broken down into two parts: First, the observed variables (in rectangles) that indicate (through arrows) the latent traits (in ovals) are what constitute the measurement model, and the relationships among the latent traits (i.e., the ovals and the arrows connecting the ovals) are the path model (see McDonald & Ho, 2002). The means of and the variance within the factors (or variables) in the model can be compared and tested with SEM. In this particular full latent trait (structural) model, the analysis through SEM combines CFA, correlation, and regression analyses into one package so that the entire model can be tested at once (Byrne, 2010).[1] Researchers testing this model with SEM will find out, for example, how much of the variation in L2-learning aptitude accounts for (or predicts) variation in L2 proficiency (revealed through a regression coefficient). In fact, every arrow (each parameter) in the model will be equated, that is, each will have an effect size (single-headed arrows) or correlation (double-headed arrows) estimated for it. Each of these coefficients, in turn, will be tested for significance. The procedure for generating these estimates is by associating

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

indicator-variable data from a specific population with the model structure. The biggest question of all is this: From whom shall the data come?

The data set is of utmost importance because there are, as any applied linguist knows, many different models of L2 learning, and any given model may work with one population, but not another. This is why models tested with SEM are described as plausible with a given data set or not. As stated earlier, with SEM, researchers can investigate a single proposed set of relationships among one or more independent variables and one or more dependent variables. A model that is found to be consistent with the data through SEM has nothing to say about whether a better model might be possible with the existing set of data or with other data. Thus, the results stemming from SEM are grounded within the context of the particular data set (i.e., within the context of the particular L2 being learned, and with the particular people from whom the data were collected). Because of this, a benefit of SEM is that researchers can use it to compare models (using chi-square difference tests, as explained in Kline, 2011). A model can be tested over time with the same population (see Barkaoui, this volume; Li, 2013; Little, 2013). A model can be tested across different data sets (e.g., Andringa, Olsthoorn, van Beunigen, Schoonen, & Hulstijn, 2012; Dale, Harlaar, & Plomin, 2012; Ghonsooly, Khajavy, & Asadpour, 2012; Guo & Roehrig, 2011; Henry & Cliffordson, 2013; Kieffer & Lesaux, 2012), or different models can be tested and compared (e.g., Cochran, McCallum, & Mee Bell, 2010; MacIntyre, MacMaster, & Baker, 2001; Song, 2008, 2012; W.-T. Tseng & Schmitt, 2008). Best of all, with SEM, the structural relations can be plotted graphically, as in Figure 2, which enables a very clear visualization of the theory under study (Byrne, 2010). These benefits are what attract applied linguists to this method of data analysis.

THE LAST FIVE YEARS OF SEM IN SLA RESEARCH

The last five years has seen tremendous growth in the use of SEM by SLA researchers. By searching for SEM studies in the *Linguistics and Language Behavior Abstracts* (LLBA), and by restricting this search to peer-reviewed articles within foreign or SLA, bilingualism, L2 learning, or L2 testing, I found 39 articles using SEM have been published since 2008. I found two SEM overview articles (In'nami & Koizumi, 2011; Kunnan, 2010). These are excellent, as are recent textbooks on SEM (i.e., Byrne, 2010; Hoyle, 2012a; Kline, 2011; Little, 2013; Schumacker & Lomax, 2010), and I refer readers to them to further understand the technical details of SEM and how SEM compares to non-model-testing (descriptive) statistics, such as exploratory factor analysis, correlation, and simple and multiple regression. I review only peer-reviewed research articles from the last five years to continue from where In'nami and Koizumi left off (their review of SEM studies was on those published up until 2008) and to give more of an overview of how researchers are using this data analysis technique most recently. The field of psychology also has review articles on how SEM has been applied since 1985 to psychological research, and I refer readers to those articles to understand more fully SEM's various applications (Hershberger, 2003; MacCallum & Austin, 2000; Tremblay & Gardner, 1996).

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

For the purpose of this review, I first made a spreadsheet summary of the 39 SEM research articles to categorize the articles by author, year, journal, number of participants, their independent and dependent variables, and number of parameter estimates. I also noted which type of SEM the researchers used. Notably, 37 out of the 39 studies presented basic applications of SEM (Kline, 2011): CFA, path analysis, or full latent trait (structural) model testing.[2] The two others presented latent growth curve modeling (Kieffer, 2011; Mancilla-Martinez & Lesaux, 2010), which is but one of the many advanced applications of SEM (see parts IV and V in Hoyle, 2012a, for an overview of the basic and advanced applications). I also noted whether language learning or development was part of the measurement model (it was for 26 out of the 39). I recorded if there was a comparison of models (yes for 12). I recorded which software was used for the analysis; most used Amos, but several used EQS, LISREL, or Mplus (see Byrne, 2012, for information on SEM software types). I also recorded how many points the researchers had on their Likert scales and the average reliability of the researchers' instruments. I will discuss some of these details after I first review the major topics the researchers of the 39 studies investigated.

Topics Investigated Using SEM

Over the last five years, research in SEM has included eight broad topic areas.

1. Linguistic knowledge structures.

Eleven studies centered on how knowledge structures, as predictor (independent) variables, affect language development. The knowledge structures consisted of, for example, metacognitive knowledge, working memory, phonological awareness, or syntactic knowledge, and the dependent variables were measures of, for example, discourse comprehension, reading comprehension, reading fluency, listening comprehension, writing proficiency, or vocabulary development. The authors tested their knowledge-structure models on the language development of adults older than 18 years of age (Andringa et al., 2012; Guo & Roehrig, 2011; Pulido & Hambrick, 2008; Zhang, 2012), teenagers (Marx & Roick, 2012; Schoonen, van Gelderen, Stoel, Hulstijn, & de Glopper, 2011), or children (Dale et al., 2012; Kieffer, 2011; Kieffer & Lesaux, 2012; Yaghoub Zadeh, Farnia, & Geva, 2012), with three additional studies specifically focusing on children's Spanish/English (Mancilla-Martinez & Lesaux, 2010; Proctor, August, Snow, & Barr, 2010) or Chinese/English (Pasquarella, Chen, Lam, Luo, & Ramirez, 2011) bilingual development. Three of these seven child studies were extremely large in scale in comparison to most others within this strand (and also when compared to all of the 39 studies' sample sizes), with 952, 1,632, and 9,189 participants, respectively. (Excluding these three, the remaining nine studies investigating knowledge structures had an average of 232 participants.) The one with 1,632 children was developed out of well-funded data-collection efforts in Britain (Dale et al., 2012), and the other two were based on U.S. census data (Kieffer, 2011, with 9,189 children; Kieffer & Lesaux, 2012, with 952 children). Two (Dale et al., 2012; Kieffer & Lesaux, 2012) of these three large-scale studies are extremely unique and important because they tracked children's L2 development from

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

toddlerhood to the young teen years and included socioeconomic status as one of the predictor variables. Meanwhile, Kieffer (2011) revealed that socioeconomic status (in particular, poverty) is a factor that shapes future language learning development over and beyond initial language levels in kindergarten (around age 5 years). More studies on knowledge structures are certainly needed, and large-scale ones that investigate factors that contribute to language learning (and bilingual development) during early childhood are clearly warranted as they can help define theories about child-language education in general.

2. Language testing.

Seven studies were in the area of language testing. All of these language-assessment studies were conducted with college or university English language learners, an obviously abundant type of test taker. The researchers of these studies looked at what influences test preparation (Xie & Andrews, 2013), how test-taking strategies influence writing test performance (Yang, 2012), and how note-taking affects listening test performance (Song, 2012). Four articles were construct-validation studies (Aryadoust, 2010; Phakiti, 2008a, 2008b; Song, 2008), three of which used CFA to investigate the subcomponents of larger constructs (Aryadoust, 2010; Phakiti, 2008a, 2008b). The SEM studies in language testing have continued to demonstrate the great promise of SEM to aid in the validation of language assessments. Measurement researchers have paved the way for language-testing researchers who want to use SEM; in particular, SEM specialists have described SEM as a way to validate tests and estimate the reliability of tests (Raykov, 2012; Raykov, Dimitrov, & Asparouhov, 2010; Raykov & Marcoulides, 2013). However, language-testing-SEM studies need to be diversified through theory testing with data from more varied populations (i.e., with data from test takers being assessed in languages other than English).

3. Motivation.

Six studies were squarely on motivation. Of these, however, only one was on motivation's effects on learning; in particular, it was on how vocabulary-learning motivation affected vocabulary knowledge (measured in size and depth; W.-T. Tseng & Schmitt, 2008). Of these six, one was on German as an L2 (with young teens) in Germany and Switzerland (Wolfgramm, Rau, Zander-Musić, Neuhaus, & Hannover, 2010). One was on German and English as a foreign language in Hungary (Csizér & Kormos, 2009). The other four were on English as a foreign language in Iran (Papi, 2010), Japan (Hiromori, 2009), Sweden (Henry & Cliffordson, 2013), and (the vocabulary motivation study) Taiwan (W.-T. Tseng & Schmitt, 2008). Interestingly, none of these SEM studies on motivation investigated whether a model of motivation necessarily changes over time for any given set of learners. This most likely needs to be done because motivation theorists have posited that motivation is an ever-shifting variable that must be sustained over time (Dörnyei, 2009; Ushioda & Dörnyei, 2012). Elements that contribute to the sustainment need to be investigated, and cross-sectional or longitudinal multimodel testing within this area may be fruitful.

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

4. Willingness to communicate.

Five studies were on language learners' willingness to communicate (WTC) in the target language. One of these studies investigated how WTC affects language development (speaking, listening, reading, and writing), but the researchers of that study did not provide information on how they assessed the language (Aidinlou & Ghobadi, 2012). Two studies had clearly defined dependent variables: perceived stress (Gallagher, 2013) or WTC during group work (Fushino, 2010). The researchers of the other two studies investigated the components of WTC in general and how they were interrelated (Ghonsooly et al., 2012; Peng & Woodrow, 2010). All five studies involved university students learning English. Four of the five were on English as a foreign language (Aidinlou & Ghobadi, 2012; Fushino, 2010; Ghonsooly et al., 2012; Peng & Woodrow, 2010) and one was on English as an L2 in Britain (Gallagher, 2013). As with motivation, research on WTC should expand into other language-learning contexts so that the theory (that WTC is an important factor in L2 development; MacIntyre, Dörnyei, Clément, & Noels, 1998) can be more robustly investigated.

5. L2-learning aptitude.

Three SEM studies were on L2-learning aptitude,[3] and all three used components of the MLAT. Cochran et al. (2010) explored how aptitude, native-language reading levels, attributions, and attitudes affected foreign language learning. They tested their model on 648 learners from three different L2s: Spanish, French, and German. The learners were all in 100-level language classes, and L2 learning was measured via the students' final exam scores. Winke (2013) looked at how aptitude and working memory (WM), mediated by motivation and strategy use, affected Chinese learning. Li (2013) also investigated aptitude and WM for learning Chinese, but he additionally investigated how feedback type (implicit or explicit), in connection with aptitude, affects L2 learning. Two of these studies were plagued with low participant numbers (78 in Li; 96 in Winke) because the researchers tried to ensure that the participants comprised a homogeneous group with each person having received the same quality, type, and length of language instruction. Participant numbers can be larger (as in Cochran et al., 2010) when instructional variance is considered to have a relatively small effect. There may be trade-offs and benefits to both approaches: If a researcher controls tightly for instruction, he or she will have fewer participants (reducing the researcher's power to reject the null hypothesis, perhaps to a dangerously low level). If he or she allows for variance in instruction, obtaining higher participant numbers (and more power), he or she will concomitantly have reduced assurance that any observed aptitude effect is the sole cause of learning differences. Perhaps in the field of SLA, both approaches are warranted because modeling the components of aptitude and relating those components with language-learning outcomes is needed on both theoretical and practical grounds. L2 aptitude researchers wrestle with whether aptitude is a static or malleable trait, a unitary or multifaceted construct, one that includes WM or inductive language learning, and whether L2 aptitude is potentially mediated by other IDs. Thus more aptitude researchers should take up SEM as a data analysis tool.

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

6. International posture.

Three studies were on international posture or general dispositions toward language learning in relation to the targeted language's perceived status. One was on how international posture affects motivated learning behavior (Csizér & Kormos, 2008); a second was on how international posture affects attitudes toward L2 learning (Rivers, 2010); and the third was on how various dispositions toward English (as a lingua franca, important for professional development) affect individuals' beliefs about English and English-learning goals (Csizér & Kontra, 2012). Csizér and Kormos (2008) asked German-language–learning children (ages 13 and 14 years) in Hungary about their linguistic self-confidence, their foreign media use, and their attitudes about English. They investigated whether those and other variables predicted children's self-reported motivated learning behaviors. Likewise, Rivers investigated learners' attitudes toward language learning, this time with university students learning English in Japan. Rivers surveyed the students as to their internationalism, patriotism, and their view on the intercultural appeal of speakers of English. Csizér and Kontra investigated English learning by university students in Hungary and found that students' English-learning goals were strongly shaped by the students' desires to become nativelike in the language, which they noted is needed for communication with almost any non-Hungarian (for international communication in general). These studies are promising because they help applied linguists understand how cross-cultural communication, global politics, and national feelings about peoples and cultures (including potentially xenophobic inclinations) affect learning outcomes, as well as learning and teaching trends.

7. Computer-assisted language learning.

One study explored individuals' attitudes toward computer-assisted language learning (CALL). Vandewaetere, Desmet, and Leuven (2010) investigated how 240 individuals' CALL experience, combined with their attitudes toward foreign language learning in general, affected their general attitude toward CALL. Given the plethora of research on how individual differences and different learning or teaching environments impact the effectiveness of CALL (i.e., Dooly, 2007, 2011; Razagifard & Rahimpour, 2010; Stracke, 2007; Y.-C. Tseng & Liou, 2006), I was surprised to find this one sole article within CALL that uses SEM as an analysis tool. This is likely a subject within SLA that may well see more SEM studies in the future. For example, Godwin-Jones (2011) discussed how CALL promotes autonomous language learning because CALL helps students develop learning strategies that will most likely change as learning changes. Someone investigating such a hypothesis (or proposed cause and effect relationship between CALL and language learning, mediated by strategies) may find SEM to be a good statistical tool for doing so.

8. Acculturation.

One SEM study directly examined acculturation.[4] The authors, Miglietta and Tartaglia (2009), investigated how length of stay, linguistic competence, and media exposure, mediated by actual in-country use of the target language (Italian, in

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

this case), affected Italian-as-a-second-language-speakers' sense of emotional belonging. Using SEM, the authors found support of their model, which was based on the theory that greater competence in Italian would increase acculturation (adaption to life in Italy) and that higher language ability also would influence emotional attachment to the country both directly and indirectly. Increases in mass media consumption also aligned with greater emotional attachment to the country. This study may break new ground for study abroad researchers interested in seeing how their research may benefit from SEM. Most directly, it exemplifies SEM's potential for research that involves acculturation. Miglietta and Tartaglia conducted their study on 576 adult immigrants in Italy, but the questions are similar to those that study abroad researchers (e.g., those like Du, 2013; Mererdith, 2010; Segalowitz & Freed, 2004; Yu, 2010) often investigate: What conditions or factors contribute most to success when people learn languages outside of their home country? Study abroad is technically defined as a temporary stay in a foreign country with the purpose of language acquisition (Kinginger, 2009, p. 18), and this is not to be confused with the permanent type of stay dealt with in Miglietta and Tartaglia (2009). However, if study abroad researchers investigating aspects of acculturation or study abroad L2-learning success (or both) can obtain large enough sample sizes, they could consider using SEM to test their theories.

SLA RESEARCHERS' MAIN PROBLEMS IN USING SEM AND POSSIBLE SOLUTIONS

By overviewing the 39 empirical research articles that have used SEM over the last five years, I can see that the authors of the studies are doing most things right, which is not surprising given that the journals in which the studies appeared have rigorous reviewing processes. The authors are following recommendations to report multiple fit indices (Hu & Bentler, 1999; McDonald & Ho, 2002). They mostly have adequate sample sizes (as I will explain), and they generally report the reliabilities of their instruments. But this is not always the case. I found that a few of the issues (including sample size and reliability) were previously discussed by In'nami and Koizumi (2011), but they are still slightly problematic. In the following section I will review four problems (sample size, model presentation, reliability, and Likert-scale points) that I think can be addressed well in this article. I will discuss the potential sources of the four problems and also provide some possible solutions. These four issues are important to discuss because novices embarking on their own research with SEM need to be aware of these up front before they finalize their research designs.

Sample Size

The 39 studies mostly have adequate sample sizes, except for, perhaps, Andringa et al. (2012), who tested a complex model with more than 50 parameters on data from as few as 113 participants. As noted by In'nami and Koizumi (2011), who were citing Raykov and Marcoulides (2006), researchers who want to apply SEM

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

to their data should have a minimum of 10 participants per parameter to be estimated. With 50 parameters, the minimum number would be 500 subjects. In fact, with such numbers (50 parameters, 500 subjects), the power to reject the model is quite close to 1 (see table 2 in MacCallum, Browne, & Sugawara, 1996). A statistically desired power is 0.8 (an 80% probability that the testing of the model would not make a type II error, that is, fail to reject the model when it is actually not plausible). But of course in SLA, unless census data is tapped into (like in Kieffer, 2011, a study with 9,189 participants), or unless the model is notably noncomplex (with few parameters to estimate, as in Li, 2013), obtaining a high ratio of parameters to participants is notably difficult. Within the field of SLA, SEM-research reviewers may be wise to accept a minimum of 10 subjects per parameter, lest the field suffer, perhaps, from a lack of model testing. The field could compensate for low power in its SEM studies by encouraging replications and model testing across various subject groups and learning contexts. Another solution researchers can apply is to collapse theoretically related indicator variables into larger composites, which renders latent traits into indicator variables, but at the same time reduces the number of parameters to be estimated. While this may be controversial, this was done in Winke (2013), where I created a composite score out of items on a survey (the Strategy Inventory for Language Learning or SILL; Oxford, 1990) that theoretically represented five factors or subsets of items. I used the one score (rather than five subscores) to indicate what normally is considered a latent variable, effectively stripping out a submodel that would require both CFA and more participants. I did this to make the model less complex (with fewer parameters) and thus easier to identify, although it weakened the theoretical model (rendered the model a less accurate representation of the underlying strategy construct). While there are potential solutions when participant numbers are low, SLA researchers should strive for larger participant numbers when possible, and perhaps this can be achieved through data mining (as done by Kieffer, 2011) rather than data collecting.

Model Presentation

When reviewing the SEM articles, I sometimes saw the authors split the presentation of the measurement model from the path model. While this is fine, it can be confusing for those new to SEM. For example, Guo and Roehrig (2011) suppressed the indicator variables from their latent trait (structural) model (see figure 2, p. 56 in their article)—they presented the path model only in figure 2. They left the indicator variables off for simplicity in presenting the model. The measurement portion of the model was presented earlier in figure 1 (p. 55 of their article). Thus, readers wanting to visualize the full model must combine the two parts. Similarly, path-only (no-indicator) models were presented by Hiromori (2009), Pulido and Hambrick (2008), Csizér and Kormos (2008), and Csizér and Kontra (2012). In Csizér and Kormos, the full model (the combined measurement and path model) initially tested appeared earlier in the article, along with all of its indicators. Thus interpreting Csizér and Kormos's model in its entirety is easy, and the full model is clear. Pulido and Hambrick also presented the measurement portion separate from the path model, as done in Guo

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

and Roehrig (2011). But in Hiromori (2009) and in Csizér and Kontra (2012) the full model was never presented, nor did they present the measurement portions of their models graphically. In Hiromori (2009), the latent traits should have been transformed into indicator or manifest variables (represented by rectangles, not ovals) because each trait was directly measured (observed) by a single score. Hiromi's model is a path analysis, not a full latent trait (structural) model analysis. (See Proctor et al., 2010, as an example of a path analysis, the sole one in the set of 39 reviewed for this chapter.) This is a small quibble, but underscores that model representation and following modeling conventions are important. Readers should be able to clearly understand the intricacies of the analysis that was conducted by viewing the models. Likewise, if only the structural portion of the model is illustrated, the authors should report, as Csizér and Kormos (2008) did, from whence the data for the latent traits came. It might be possible to provide the full model, even if it is complicated and large, in a supplemental file. (Some are already adding large SEM files as journal supplements: e.g., Henry & Cliffordson, 2013, included a large SEM correlation table as a supplement.) Readers could download the full models (with all indicator variables, latent traits, and parameter estimates) from the publisher's website.

Reliability

The third problematic area relates to reliability. The authors of the 39 studies almost always reported the reliabilities of their research instruments, but not always. As I mentioned in the first part of this article, SEM takes measurement error into account when a model is tested (Kline, 2012; Schumacker & Lomax, 2010), so high reliability is essential for model testing with SEM. As reported by Kline (2012, p. 69), "if international consistency [an instrument's Cronbach's alpha] is low, then the content of the items may be so heterogeneous that the total score is not the best possible unit of analysis for the measure." Kline further noted that if reliability is less than 0.50, "most of the observed score variance is due to random error, an unacceptable amount of imprecision in most research" (p. 70). Thus, seeing some instruments with reliability estimates as low as 0.42 (Andringa et al., 2012) and 0.18 (Pulido & Hambrick, 2008) was disheartening. If an instrument's reliability is lower than 0.50, the score (and the variable it represents) should not be included for testing in the model.

Another issue with the reporting of reliability surfaced when I tried to note the reliabilities of all instruments in the 39 studies. While a few researchers did not report reliability (Csizér & Kontra, 2012; Dale et al., 2012; Li, 2013), others reported reliability for some instruments, but not others (e.g., Andringa et al., 2012; Cochran et al., 2010; Kieffer & Lesaux, 2012; Perfetti & Lesgold, 1977; Yaghoub Zadeh et al., 2012). However, there were often good, solid reasons for the omissions. Many of the researchers administered to their participants standardized, off-the-shelf or online instruments that generated total (composite) scores, but not scores for the individual items on the measures. Alternatively, the researchers received composite scores on measures administered by other institutions or organizations. In these cases, data on reliability for the sample population is not available. Kieffer and Lesaux (2012) addressed this issue well

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

by clearly stating the Cronbach's alpha when they could calculate it themselves and by clearly reporting on the test-publisher's reliability estimate when they used a commercial test that had a published reliability estimate. Yaghoub Zedeh et al. (2012) also indicated when a reliability estimate was from their sample or provided by the testing company.

Kline (2012) noted that wherever possible, SEM researchers should report the reliability coefficients that are based on the sample at hand because the equation modeling will be based on the sample's measurement error, not some hypothetical reliability estimate. The reality is, however, SLA researchers often borrow measures or obtain test scores from other sources. In such cases, Kline noted that SEM researchers should report the published reliability coefficients and cite from where those estimates came. This, he stated, is "reliability induction" (p. 70), not reliability estimation, and warned that such coefficients may not generalize to the study's population. Thus, authors need to be very clear if the reliability reported is real (based on the sample population) or induced; and if induced, the researchers should go a step further according to Kline (p. 70), and describe whether the induced reliability estimate can be generalized to the sample population (based on whether the there are similarities between the participants in the SEM study and the test takers whose data were used to generate the instrument's published reliability estimate).

Likert-Scale Points

The fourth and final issue relates to the Likert scales used in the studies. Out of 39 studies, at least 29 used Likert-scale questions to obtain indicator-variable scores. It struck me that most of the SEM researchers used Likert scales with 5 points. A handful used up to 6 points (Papi, 2010; Phakiti, 2008a, 2008b; Rivers, 2010; Robinson, Rivers, & Brecht, 2006; Vandewaetere et al., 2010; Xie & Andrews, 2013). A few used up to 7 (Gallagher, 2013; Hiromori, 2009; W.-T. Tseng & Schmitt, 2008). A couple used up to 8 (Dale et al., 2012; Ghonsooly et al., 2012), while one used up to 11 (Peng & Woodrow, 2010). Many SLA researchers may have used 5- or 6-point scales because they employed long-standing measures or questionnaires that are available publicly. These measures may have been created long ago (e.g., the SILL was created more than 20 years ago) with 5-point, Likert-scale questions; that is, before the optional number of points on Likert scales was investigated (and I will review this research shortly). Or it could be that SLA researchers are using 6-point Likert scales to match recommendations by Dörnyei and Taguchi (2009), who wrote about designing questionnaires for SLA research. In their appendix, all scales have 6 points.

Educational measurement researchers have published extensive research on Likert scales (Betts & Hartley, 2012; Cook, Heath, Thompson, & Thompson, 2001; Gorsuch, 1997; Hartley & Betts, 2010; Neuman, 2010). The recommendation from that literature is to have 7 to 9 points, or even 10 to 11 (c.f., Jones & Loe, 2013). Neuman (2010) reported on research that has shown that as the number of scale steps is increased from 2 up through 20, the increase in reliability is very rapid at first. It tends to level off at about 7, and after 11 steps, there is little gain in reliability from increasing the number of steps. Gorsuch (1997)

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

meticulously reviewed assumptions related to factor analysis. He explained that with multiple points, Likert-scale-item means can vary with a greater likelihood of a less skewed distribution. SEM is based on reliable variation in indicator scores (the cause and effect foundation of SEM is based on the notion that the indicators in the model have good, construct-caused or construct-defining variation). The main gist is that SLA researchers who employ Likert-scale surveys or questionnaires have the potential to create more variation and higher reliability in their indicator scores by simply changing their scales from 5–6 points to 7–8 points. However, they should also note that the order of the scale (with positive orientations on the left or right) affects how people respond to the items (Hartley & Betts, 2010), especially when the respondents are children (Betts & Hartley, 2012). Thus, researchers should consider the number of scale points and the orientation (direction) of the scale carefully. Likewise, they should report the number of points and the direction of the scale in their articles.

CONCLUSION

The purpose of this article was to introduce the general concept of SEM and report on how researchers in applied linguistics are using it. Over the last five years, I found 39 SLA studies with SEM being at least one of its analyses. The researchers of the majority of these articles (11 out of 39 articles) investigated how various knowledge structures (such as metacognitive knowledge or phonological awareness) impact language development. Similar to those 11 studies were 3 on L2-learning aptitude. The researchers of seven articles were experts in L2 testing, and they mostly investigated the underlying constructs of L2 tests and strategies that impacted L2-test outcomes. ID researchers also used SEM, with six articles on motivation and five studies on language learners' WTC in the target language. Attitudes and beliefs, another type of ID, were also investigated through SEM, with three articles on international posture (people's dispositions toward learning a language in relation to the language's perceived status and importance) and one article on attitudes toward CALL. One final SEM study stood out among the rest, in particular because it was the only one that explored acculturation, that is, how factors such as length of stay, linguistic competence, and in-country, target-language use affect success in the host country. Applied linguists working within study abroad environments have much to gain from reading this study; future SEM work could test theories about optimal study abroad experiences. Overall, the collective 39 studies showcase SEM well and are excellent road maps for SLA researchers, showing them how they could apply SEM to their own data.

By reviewing research in SLA that has employed SEM, I have shown that SEM is a powerful resource for exploring the field's theories and hypotheses. SEM is, indeed, "increasingly popular" in SLA research (Schumacker & Lomax, 2010, p. 195), but at the same time, it is disappointingly "sporadically" used (Bollen & Hoyle, 2012, p. 668). The main goal now for those interested in fully exploiting SEM as a data analysis tool may be getting a larger number of SLA and language-testing researchers to be knowledgeable about SEM. This is important

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

because researchers who use SEM want (and need) those who are not experts in SEM to be able to read, understand, and intuit information from their published studies. However, increasing applied linguists' general knowledge of SEM may be an unrealistic goal because very few statistics or testing courses that center on research in applied linguistics broach the topic of SEM (Ho et al., 2012). When compared to all empirical publications in SLA, relatively few include SEM (Plonsky, 2013). Thus to learn more about SEM at this point in time, graduate students and researchers may need to attend or audit courses on SEM that are offered by SLA's cousin fields (statistics, educational measurement, or psychological measurement) or attend SEM workshops offered before conferences or at statistical institutes. Anyway one looks at it, the future is bright for SEM, and the time to start learning about it is now.

NOTES

1 SEM can also be used to perform CFA only. For examples of this within language testing and SLA, see Turner (1989) and Henry and Cliffordson (2013), respectively. For more on CFA within SEM, see Kline (2011), chapter 9.
2 CFA is confirmatory factor analysis, and this application of SEM is not a cause and effect investigation, as path analysis and latent trait model testing are. Path analysis is different from latent trait modeling because in path analysis, all variables in the model are indicator or manifest variables (represented in rectangles). A full latent trait (structural) model will include latent traits (in ovals) and their indicators (in rectangles) and may include a confirmatory factor analysis as a model subpart.
3 Two of the SEM studies (Andringa et al., 2012; Marx & Roick, 2012) that I referred to as investigations into knowledge structures had WM as independent variables (causal variables) within the structural models being tested. Thus, these two studies could be considered as studies on how aptitude affects language learning because many theorists posit that WM is an aptitude construct, like memory is. (See Winke, 2013, for a review.) However, the authors did not frame their studies as investigations into L2-learning aptitude; therefore, I did not include them here.
4 Gallagher (2013) researched *acculturative stress*. She defined this as "transactional stressors linked to adaption process" (p. 61). She measured this stress in 104 Chinese native speakers studying as international students at a British university. I classified this study as a WTC study because acculturation was only indirectly measured (through acculturative stress), but Gallagher's study is also a good example of how to employ SEM in study abroad research.

ANNOTATED BIBLIOGRAPHY

Kieffer, M. J., & Lesaux, N. K. (2012). Direct and indirect roles of morphological awareness in the English reading comprehension of native English, Spanish, Filipino, and Vietnamese speakers. *Language Learning*, *62*, 1170–1204. doi:10.1111/j.1467–9922.2012.00722.x

Kieffer and Lesaux investigated how derivational morphological awareness impacts English reading comprehension in sixth-grade students ($n = 952$) in southern California. The students came from different language backgrounds: native English, Spanish-speaking language minority, and Filipino-speaking language minority. The data were collected from the students over the course of two testing periods during their sixth-grade year. The authors tested their proposed model on all four language-background groups. The writing is clear, and the figures and tables in which the data

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

are presented are well crafted and easy to understand. These points in combination with excellent, multimodel theory-testing make this article a standout among the recent SLA SEM articles.

Miglietta, A., & Tartaglia, S. (2009). The influence of length of stay, linguistic competence, and media exposure in immigrants' adaptation. *Cross-Cultural Research*, *43*, 46–61. doi:10.1177/1069397108326289

Miglietta and Tartaglia studied 576 immigrants in Italy (196 from Romania, 179 from North Africa, and 201 from Latin America) to better understand what made them emotionally attached (acculturated) to life in Italy and what did not. The authors surveyed the individuals to find out the following information about them: how long they had been in Italy (i.e., their length of residency); their proficiency in Italian; how much they used Italian with friends and family; how much mass Italian media they consumed, and how much they knew about satellite TV programs from their homelands. The writing is this article is straightforward, and the data are well presented. Both of these help render this an exemplary SLA SEM study.

Rivers, D. J. (2010). National identification and intercultural relations in foreign language learning. *Language and Intercultural Communication*, *10*, 318–336. doi:10.1080/14708477.2010.502234

Rivers investigated Japanese university students' attitudes toward learning English. He did this by surveying 377 of them regarding their levels of internationalism, patriotism, and Japanese nationalism. He also surveyed them in terms of how vital they considered English-speaking nations and whether they believed English speakers have intercultural appeal. This article is reader friendly and interesting. Rivers's work represents well how researchers in SLA using SEM should go about describing and reporting their research.

REFERENCES

Aidinlou, N. A., & Ghobadi, S. (2012). Examination of relationships between factors affecting on oral participation of ELT students and language development: A structural equation modeling approach. *International Journal of English Linguistics*, *2*, 131–141.

Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, *62*, 49–78.

Aryadoust, V. (2010). Investigating writing sub-skills in testing English as a foreign language: A structural equation modeling study. *TESL-EJ: Teaching English as a Second or Foreign Language*, *13*, 1–20.

Betts, L., & Hartley, J. (2012). The effects of changes in the order of verbal labels and numerical values on children's scores on attitude and rating scales. *British Educational Research Journal*, *38*, 319–331. doi:10.1080/01411926.2010.544712

Bollen, K. A., & Hoyle, R. H. (2012). Latent variables in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 56–67). New York, NY: Guilford Press.

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

Byrne, B. M. (2012). Choosing structural equation modeling computer software: Snapshots of LISREL, EQS, Amos, and Mplus. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 307–324). New York, NY: Guilford Press.

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

Cochran, J. L., McCallum, R. S., & Mee Bell, S. (2010). Three A's: How do attributions, attitudes, and aptitude contribute to foreign language learning? *Foreign Language Annals*, *43*, 566–582.

Cook, C., Heath, F., Thompson, R. L., & Thompson, B. (2001). Score reliability in web- or Internet-based surveys: Unnumbered graphic rating scales versus Likert-type scales. *Educational and Psychological Measurement*, *61*, 697–706. doi:10.1177/00131640121971356

Csizér, K., & Dörnyei, Z. (2005). The internal structure of language learning motivation and its relationship with language choice and learning effort. *The Modern Language Journal*, *89*, 19–36.

Csizér, K., & Kontra, E. H. (2012). ELF, ESP, ENL and their effect on students' aims and beliefs: A structural equation model. *System*, *40*, 1–10.

Csizér, K., & Kormos, J. (2008). The role of intercultural contact in learning German in Hungary: A structural equation modelling approach. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, *13*, 1–12.

Csizér, K., & Kormos, J. (2009). Modelling the role of inter-cultural contact in the motivation of learning English as a foreign language. *Applied Linguistics*, *30*, 166–185.

Dale, P. S., Harlaar, N., & Plomin, R. (2012). Nature and nurture in school-based second language achievement. *Language Learning*, *62*, 28–48.

Dooly, M. (2007). Joining forces: Promoting metalinguistic awareness through computer-supported collaborative learning. *Language Awareness*, *16*, 57–74.

Dooly, M. (2011). Divergent perceptions of telecollaborative language learning tasks: Task-as-workplan vs. task-as-process. *Language Learning & Technology*, *15*, 69–91.

Dörnyei, Z. (2009). *The psychology of the second language learner*. Oxford, UK: Oxford University Press.

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). New York, NY: Routledge.

Du, H. (2013). The development of Chinese fluency during study abroad in China. *The Modern Language Journal*, *97*, 131–143.

Fushino, K. (2010). Causal relationships between communication confidence, beliefs about group work, and willingness to communicate in foreign language group work. *TESOL Quarterly*, *44*, 700–724.

Gallagher, H. C. (2013). Willingness to communicate and cross-cultural adaptation: L2 communication and acculturative stress as transaction. *Applied Linguistics*, *34*, 53–73.

Gardner, R. C., Tremblay, P. F., & Masgoret, A.-M. (1997). Towards a full model of second language learning: An empirical investigation. *The Modern Language Journal*, *81*, 344–362.

Ghonsooly, B., Khajavy, G. H., & Asadpour, S. F. (2012). Willingness to communicate in English among Iranian non-English major university students. *Journal of Language and Social Psychology*, *31*, 197–211. doi:10.1177/0261927′12438538

Godwin-Jones, R. (2011). Autonomous language learning. *Language Learning & Technology*, *15*, 4–11.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment*, *68*, 532–560. doi:10.1207/s15327752jpa6803_5

Guo, Y., & Roehrig, A. D. (2011). Roles of general versus second language (L2) knowledge in L2 reading comprehension. *Reading in a Foreign Language*, *23*, 42–64.

Hartley, J., & Betts, L. R. (2010). Four layouts and a finding: The effects of changes in the order of the verbal labels and numerical values on Likert-type scales. *International Journal of Social Research Methodology*, *13*, 17–27. doi:10.1080/13645570802648077

Henry, A., & Cliffordson, C. (2013). Motivation, gender, and possible selves. *Language Learning*, *63*, 271–295. doi:10.1111/lang.12009

Hershberger, S. L. (2003). The growth of structural equation modeling: 1994–2001. *Structural Equation Modeling*, *10*, 53–46.

Hiromori, T. (2009). A process model of L2 learners' motivation: From the perspectives of general tendency and individual differences. *System*, *37*, 313–321.

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

Ho, M. R., Stark, S., & Chernyshenko, O. (2012). Graphical representation of structural equation models using path diagrams. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 43–55). New York, NY: Guilford Press.
Hoyle, R. H. (Ed.): (2012a). *Handbook of structural equation modeling*. New York, NY: Guilford Press.
Hoyle, R. H. (2012b). Model specification in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 126–144). New York, NY: Guilford Press.
Hu, L., & 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. doi:10.1080/10705519909540118
In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System*, *34*, 317–340.
In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly*, *8*, 250–276.
Isemonger, I. M. (2007). Operational definitions of explicit and implicit knowledge: Response to R. Ellis (2005) and some recommendations for future research in this area. *Studies in Second Language Acquisition*, *29*, 101–118.
Jones, W. P., & Loe, S. A. (2013). Optimal number of questionnaire response categories: More may not be better. *SAGE Open, April-June*, 1–10. doi:10.1177/2158244013489691
Kieffer, M. J. (2011). Converging trajectories: Reading growth in language minority learners and their classmates, kindergarten to grade 8. *American Educational Research Journal*, *48*, 1187–1225. doi:10.3102/0002831211419490
Kieffer, M. J., & Lesaux, N. K. (2012). Direct and indirect roles of morphological awareness in the English reading comprehension of native English, Spanish, Filipino, and Vietnamese speakers. *Language Learning*, *62*, 1170–1204. doi:10.1111/j.1467-9922.2012.00722.x
Kinginger, C. (2009) *Language learning and study abroad: A critical reading of research*. New York, NY: Palgrave Macmillan.
Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
Kline, R. B. (2012). Assumptions in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111–125). New York, NY: Guilford Press.
Kunnan, A. J. (1998). An introduction to structural equation modelling for language assessment research. *Language Testing*, *15*, 295–332.
Kunnan, A. J. (1999). Recent developments in language testing. *Annual Review of Applied Linguistics*, *19*, 235–253.
Kunnan, A. J. (2010). Statistical analyses for test fairness. *Revue française de linguistique appliquée*, *15*, 39–48.
Lewis, M. B., & Vladeanu, M. (2006). What do we know about psycholinguistic effects? *The Quarterly Journal of Experimental Psychology*, *59*, 977–986.
Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *The Modern Language Journal*, *97*, 634–654. doi:10.1111/j.1540-4781.2013.12030.x
Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.
MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 200–226. doi:10.1146/annurev.psych.51.1.201
MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.
MacIntyre, P. D., Dörnyei, Z., Clément, R., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, *82*, 545–562.
MacIntyre, P. D., MacMaster, K., & Baker, S. C. (2001). The convergence of multiple models of motivation for second language learning: Gardner, Pintrich, Kuhl, and McCroskey. In

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

R. Schmidt (Ed.), *Motivation and second language acquisition* (pp. 461–492). Honolulu, HI: University of Hawai'i Press.
Mancilla-Martinez, J., & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology*, *102*, 701–711.
Marx, A., & Roick, T. (2012). Determinants of listening comprehension in German L1 and L2 students. *Zeitschrift für Pädagogische Psychologie*, *26*, 121–134.
Matsumura, S. (2003). Modelling the relationships among interlanguage pragmatic development: L2 proficiency and exposure to L2. *Applied Linguistics*, *24*, 465–491.
McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64–82.
Mererdith, R. A. (2010). Acquiring cultural perceptions during study abroad: The influence of youthful associates. *Hispania*, *93*, 686–702.
Miglietta, A., & Tartaglia, S. (2009). The influence of length of stay, linguistic competence, and media exposure in immigrants' adaptation. *Cross-Cultural Research*, *43*, 46–61. doi:10.1177/1069397108326289
Neuman, L. (2010). *Social research methods: Qualitative and quantitative approaches* (7th ed.). Harlow, UK: Pearson Education.
Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. Boston, MA: Heinle & Heinle.
Papi, M. (2010). The L2 motivational self system, L2 anxiety, and motivated behavior: A structural equation modeling approach. *System*, *38*, 467–479.
Pasquarella, A., Chen, X., Lam, K., Luo, Y. C., & Ramirez, G. (2011). Cross-language transfer of morphological awareness in Chinese-English bilinguals. *Journal of Research in Reading*, *34*, 23–42.
Pearl, J. (2012). The causal foundations of structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 68–91). New York, NY: Guilford Press.
Peng, J.-E., & Woodrow, L. (2010). Willingness to communicate in English: A model in the Chinese EFL classroom context. *Language Learning*, *60*, 834–876. doi:10.1111/j.1467-9922.2010.00576.x
Perfetti, C. A., & Lesgold, A. M. (1977). Discourse comprehension and sources of individual differences. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 141–183). Hillsdale, NJ: Erlbaum.
Phakiti, A. (2008a). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing*, *25*, 237–272.
Phakiti, A. (2008b). Strategic competence as a fourth-order factor model: A structural equation modeling approach. *Language Assessment Quarterly*, *5*, 20–42.
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.
Proctor, C. P., August, D., Snow, C., & Barr, C. D. (2010). The interdependence continuum: A perspective on the nature of Spanish-English bilingual reading comprehension. *Bilingual Research Journal*, *33*, 5–20.
Pulido, D., & Hambrick, D. Z. (2008). The virtuous circle: Modeling individual differences in L2 reading and vocabulary development. *Reading in a Foreign Language*, *20*, 164–190.
Purpura, J. (1998). Investigating the effects of strategy use and second language test performance with high and low ability test takers: A structural equational modeling approach. *Language Testing*, *15*, 333–379.
Raykov, T. (2012). Scale construction and development using structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 472–492). New York, NY: Guilford Press.
Raykov, T., Dimitrov, D. M., & Asparouhov, T. (2010). Evaluation of scale reliability with binary measures using latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*, 265–279. doi:10.1080/10705511003659417

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah, NJ: Erlbaum.
Raykov, T., & Marcoulides, G. A. (2013). Meta-analysis of scale reliability using latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 338–353. doi:10.1080/10705511.2013.769396
Razagifard, P., & Rahimpour, M. (2010). The effect of computer-mediated corrective feedback on the development of second language learners' grammar. *International Journal of Instructional Technology & Distance Learning*, *7*. Retrieved from http://itdl.org/Journal/May_2010/article2002.htm
Rivers, D. J. (2010). National identification and intercultural relations in foreign language learning. *Language and Intercultural Communication*, *10*, 318–336. doi:10.1080/14708477.2010.502234
Robinson, J. P., Rivers, W. P., & Brecht, R. D. (2006). Demographic and sociopolitical predictors of American attitudes toward foreign language policy. *Language Policy*, *5*, 419–440. doi:10.1007/s10993-006-9033-4
Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach. *Language Learning*, *43*, 313–344.
Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*, 1–30.
Schoonen, R., Van Gelderen, A., De Glopper, K., Hulstijn, J., Simis, A., Snelling, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing and metacognitive knowledge. *Language Learning*, *53*, 165–202.
Schoonen, R., van Gelderen, A., Stoel, R. D., Hulstijn, J., & de Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning*, *61*, 31–79. doi:10.1111/j.1467–9922.2010.00590.x
Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). New York, NY: Routledge.
Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*, 173–199.
Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 69–93). Philadelphia, PA: John Benjamins.
Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, *25*, 435–464.
Song, M.-Y. (2012). Note-taking quality and performance on an L2 academic listening test. *Language Testing*, *29*, 67–89. doi:10.1177/0265532211415379
Stracke, E. (2007). A road to understanding: A qualitative study into why learners drop out of a blended language learning (BLL) environment. *ReCALL*, *19*, 57–78.
Tseng, W.-T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*, 357–400.
Tseng, Y.-C., & Liou, H.-C. (2006). The effects of online conjunction materials on college EFL students' writing. *System*, *34*, 270–283.
Tremblay, P. F., & Gardner, R. C. (1995). Expanding the motivation construct in language learning. *The Modern Language Journal*, *79*, 505–518.
Tremblay, P. F., & Gardner, R. C. (1996). On the growth of structural equation modeling in psychological journals. *Structural Equation Modeling*, *3*, 93–104.
Turner, C. E. (1989). The underlying factor structure of L2 close test performance in Francophone, university-level students: Causal modelling as an approach to construct validation. *Language Testing*, *6*, 172–197. doi:10.1177/026553228900600205
Ushioda, E., & Dörnyei, Z. (2012). *Motivation.* In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 396–409). New York, NY: Routledge.
Vandewaetere, M., Desmet, P., & Leuven, K. U. (2010). The influence of learners' attitudes to and perceptions of CALL: A methodological approach. *ITL—International Journal of Applied Linguistics*, *160*, 1–23.

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press

Verhoeven, L. (2007). Early bilingualism, language transfer, and phonological awareness. *Applied Psycholinguistics*, *28*, 425–439. doi:10.1017/S0142716407070233

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

Wolfgramm, C., Rau, M., Zander-Musić, L., Neuhaus, J., & Hannover, B. (2010). On the connection between collective self-worth and the motivation to learn German: A study of students with a migration background in Germany and Switzerland. *Zeitschrift für Pädagogik*, *56*, 59–77.

Woodrow, L. J. (2006). A model of adaptive language learning. *The Modern Language Journal*, *90*, 297–319.

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with structural equation modeling. *Language Testing*, *30*, 49–70.

Yaghoub Zadeh, Z., Farnia, F., & Geva, E. (2012). Toward modeling reading comprehension and reading fluency in English language learners. *Reading and Writing*, *25*, 163–187.

Yang, H.-C. (2012). Modeling the relationships between test-taking strategies and test performance on a graph-writing task: Implications for EAP. *English for Specific Purposes*, *31*, 174–187.

Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *The Modern Language Journal*, *86*, 54–66.

Yu, B. (2010). Learning Chinese abroad: The role of language attitudes and motivation in the adaptation of international students in China. *Journal of Multilingual and Multicultural Development*, *31*, 301–321.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal*, *96*, 558–575.

https://doi.org/10.1017/S0267190514000075 Published online by Cambridge University Press