YO IN'NAMI AND RIE KOIZUMI

2. STRUCTURAL EQUATION MODELING IN EDUCATIONAL RESEARCH: A PRIMER

INTRODUCTION

Structural equation modeling (SEM) is a collection of statistical methods for modeling the multivariate relationship between variables. It is also called covariance structure analysis or simultaneous equation modeling and is often considered an integration of regression and factor analysis. As SEM is a flexible and powerful technique for examining various hypothesized relationships, it has been used in numerous fields, including marketing (e.g., Jarvis, MacKenzie, & Podsakoff, 2003; Williams, Edwards, & Vandenberg, 2003), psychology (e.g., Cudeck & du Toit, 2009; Martens, 2005), and education (e.g., Kieffer, 2011; Teo & Khine, 2009; Wang & Holcombe, 2010). For example, educational research has benefited from the use of SEM to examine (a) the factor structure of the learner traits assessed by tests or questionnaires (e.g., Silverman, 2010; Schoonen et al., 2003), (b) the equivalency of models across populations (e.g., Byrne, Baron, & Balev, 1998; In'nami & Koizumi, 2012; Shin, 2005), and (c) the effects of learner variables on proficiency or academic achievement at a single point in time (e.g., Ockey, 2011; Wang & Holcombe, 2010) or across time (e.g., Kieffer, 2011; Marsh & Yeung, 1998; Tong, Lara-Alecio, Irby, Mathes, & Kwok, 2008; Yeo, Fearrington, & Christ, 2011). This chapter provides the basics and the key concepts of SEM, with illustrative examples in educational research. We begin with the advantages of SEM, and follow this with a description of Bollen and Long's (1993) five steps for SEM application. Then, we discuss some of the key issues with regard to SEM. This is followed by a demonstration of various SEM analyses and a description of software programs for conducting SEM. We conclude with a discussion on learning more about SEM. Readers who are unfamiliar with regression and factor analysis are referred to Cohen, Cohen, West, and Aiken (2003), Gorsuch (1983), and Tabachnick and Fidell (2007). SEM is an extension of these techniques, and having a solid understanding of them will aid comprehension of this chapter.

ADVANTAGES OF SEM

SEM is a complex, multivariate technique that is well suited for testing various hypothesized or proposed relationships between variables. Compared with a number of statistical methods used in educational research, SEM excels in four aspects (e.g., Bollen, 1989; Byrne, 2012b). First, SEM adopts a confirmatory,

M.S. Khine (ed.), Application of Structural Equation Modeling in Educational Research and Practice, 23–51.

hypothesis-testing approach to the data. This requires researchers to build a hypothesis based on previous studies. Although SEM can be used in a model-exploring, data-driven manner, which could often be the case with regression or factor analysis, it is largely a confirmatory method. Second, SEM enables an explicit modeling of measurement error in order to obtain unbiased estimates of the relationships between variables. This allows researchers to remove the measurement error from the correlation/regression estimates. This is conceptually the same as correcting for measurement error (or correcting for attenuation), where measurement error is taken into account for two variables by dividing the correlation by the square root of the product of the reliability estimates of the two instruments ($r_{xy} / \sqrt{[r_{xx} \times r_{yy}]}$). Third, SEM can include both unobserved (i.e., latent) and observed variables. This is in contrast with regression analysis, which can only model observed variables, and with factor analysis, which can only model unobserved variables. Fourth, SEM enables the modeling of complex multivariate relations or indirect effects that are not easily implemented elsewhere. Complex multivariate relations include a model where relationships among only a certain set of variables can be estimated. For example, in a model with variables 1 to 10, it could be that only variables 1 and 2 can be modeled for correlation. Indirect effects refer to the situation in which one variable affects another through a mediating variable.

FIVE STEPS IN AN SEM APPLICATION

The SEM application comprises five steps (Bollen & Long, 1993), although they vary slightly from researcher to researcher. They are (a) model specification, (b) model identification, (c) parameter estimation, (d) model fit, and (e) model respecification. We discuss these steps in order to provide an outline of SEM analysis; further discussion on key issues will be included in the next section.

Model Specification

First, model specification is concerned with formulating a model based on a theory and/or previous studies in the field. Relationships between variables – both latent and observed – need to be made explicit, so that it becomes clear which variables are related to each other, and whether they are independent or dependent variables. Such relationships can often be conceptualized and communicated well through diagrams.

For example, Figure 1 shows a hypothesized model of the relationship between a learner's self-assessment, teacher assessment, and academic achievement in a second language. The figure was drawn using the SEM program Amos (Arbuckle, 1994-2012), and all the results reported in this chapter are analyzed using Amos, unless otherwise stated. Although the data analyzed below are hypothetical, let us suppose that the model was developed on the basis of previous studies. Rectangles represent observed variables (e.g., item/test scores, responses to questionnaire items), and ovals indicate unobserved variables. Unobserved variables are also

called factors, latent variables, constructs, or traits. The terms *factor* and *latent variable* are used when the focus is on the underlying mathematics (Royce, 1963), while the terms *construct* and *trait* are used when the concept is of substantive interest. Nevertheless, these four terms are often used interchangeably, and, as such, are used synonymously throughout this chapter. Circles indicate measurement errors or residuals. Measurement errors are hypothesized when a latent variable affects observed variables, or one latent variable affects another latent variable. Observed and latent variables that receive one-way arrows are usually modeled with a measurement error. A one-headed arrow indicates a hypothesized one-way direction, whereas a two-headed arrow indicates a correlation between two variables. The variables that release one-way arrows are independent variables (also called exogenous variables), and those that receive arrows are dependent variables (also called endogenous variables). In Figure 1, self-assessment is hypothesized to comprise three observed variables of questionnaire items measuring self-assessment in English, mathematics, and science. These observed variables are said to *load on* the latent variable of self-assessment. Teacher assessment is measured in a similar manner using the three questionnaire items, but this time presented to a teacher. The measurement of academic achievement includes written assignments in English, mathematics, and science. All observed variables are measured using a 9-point scale, and the data were collected from 450 participants. The nine observed variables and one latent variable contained measurement errors. Self-assessment and teacher assessment were modeled to affect academic achievement, as indicated by a one-way arrow. They were also modeled to be correlated with each other, as indicated by a two-way arrow.

Additionally, SEM models often comprise two subsets of models: a measurement model and a structured model. A measurement model relates observed variables to latent variables, or, defined more broadly, it specifies how the theory in question is operationalized as latent variables along with observed variables. A structured model relates constructs to one another and represents the theory specifying how these constructs are related to one another. In Figure 1, the three latent factors – self-assessment, teacher assessment, and academic achievement – are measurement models; the hypothesized relationship between them is a structural model. In other words, structural models can be considered to comprise several measurement models. Since we can appropriately interpret relationships among latent variables only when each latent variable is well measured by observed variables, an examination of the model fit (see below for details) is often conducted on a measurement model before one constructs a structural model.

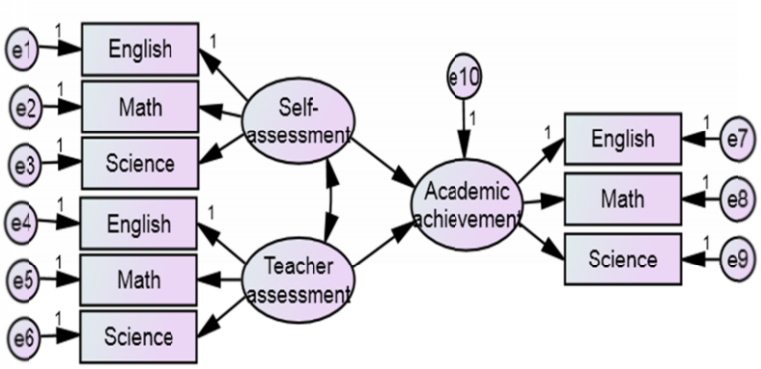

Figure 1. Example SEM model diagram.

Model Identification

The second step in an SEM application, namely model identification, is concerned with whether one can derive a unique value for each parameter (in the model) whose value is unknown (e.g., factor loadings, factor correlations, measurement errors) using the variance/covariance matrix (or the correlation matrix and standard deviations) of the measured variables that *are* known. Models are not identified when there are more parameters than can be estimated from the information available in the variance/covariance matrix. Models that are complex, even if theoretically sound, are likely to have identification problems, particularly when there are a large number of parameters to be estimated relative to the number of variances and covariances in the matrix. Two important principles are applicable to the identification of SEM models. First, latent variables must be assigned a scale (metric) because they are unobserved and do not have predetermined scales. This can be achieved by fixing either a factor variance, or one of the factor loadings, to be a specific value, usually 1. Second, the number of data points in the variance/covariance matrix – known information – must be at least equal to the number of parameters to be estimated in the model (i.e., free parameters) – unknown information. For example, for the academic achievement model, there are 21 estimated parameters: 8 factor loadings, 10 measurement error variances, 1 covariance, and 2 factor variances. Three of the factor loadings are each fixed to be 1 and do not have to be estimated. The number of data points is $p(p + 1)/2$, where p refers to the number of observed variables. For the academic achievement factor in Figure 1, there are nine observed variables, and therefore $9(9 + 1)/2 = 45$ data points. This is larger than the number of parameters to be estimated in the model, which is 21. Thus, this model is identifiable. The degrees of freedom (*df*) are the difference between the number of data points and the number of parameters to be estimated. In the current example, the *df* are 24. When *df* are positive (one or

above), models can be identified. When *df* are negative, models cannot be identified, and are called unidentified. When *df* are zero, models can be identified but cannot be evaluated using fit indices (for fit indices, see below).

Parameter Estimation

Third, once the model has been identified, the next step is to estimate parameters in the model. The goal of parameter estimation is to estimate population parameters by minimizing the difference between the observed (sample) variance/covariance matrix and the model-implied (model-predicted) variance/covariance matrix. Several estimation methods are available, including maximum likelihood, robust maximum likelihood, generalized least squares, unweighted least squares, elliptical distribution theory, and asymptotically distribution-free methods. Although the choice of method depends on many factors, such as data normality, sample size, and the number of categories in an observed variable, the most widely used method is maximum likelihood. This is the default in many SEM programs because it is robust under a variety of conditions and is likely to produce parameter estimates that are unbiased, consistent, and efficient (e.g., Bollen, 1989). Maximum likelihood estimation is an iterative technique, which means that an initially posited value is subsequently updated through calculation. The iteration continues until the best values are attained. When this occurs, the model is said to have converged. For the current example in Figure 1, the data were analyzed using maximum likelihood. The subsequent section entitled Data Normality provides more discussion on some recommendations for choice of estimation method.

Model Fit

Fourth, when parameters in a model are estimated, the degree to which the model fits the data must be examined. As noted in the preceding paragraph, the primary goal of SEM analysis is to estimate population parameters by minimizing the difference between the observed and the model-implied variance/covariance matrices. The smaller the difference is, the better the model. This is evaluated using various types of fit indices. A statistically nonsignificant chi-square (χ^2) value is used to indicate a good fit. Statistical nonsignificance is desirable because it indicates that the difference between the observed and the model-implied variance/covariance matrices is statistically nonsignificant, which implies that the two matrices cannot be said to be statistically different. Stated otherwise, a nonsignificant difference suggests that the proposed model cannot be rejected and can be considered correct. Note that this logic is opposite to testing statistical significance for analysis of variance, for example, where statistical significance is usually favorable.

Nevertheless, chi-square tests are limited in that, with large samples, they are likely to detect practically meaningless, trivial differences as statistically significant (e.g., Kline, 2011; Ullman, 2007). In order to overcome this

problem, many other fit indices have been created, and researchers seldom depend entirely on chi-square tests to determine whether to accept or reject the model. Fit indices are divided into four types based on Byrne (2006) and Kline (2011), although this classification varies slightly between researchers. First, incremental or comparative fit indices compare the improvement of the model to the null model. The null model assumes no covariances among the observed variables. Fit indices in this category include the comparative fit index (CFI), the normal fit index (NFI), and the Tucker-Lewis index (TLI), also known as the non-normed fit index (NNFI). Second, unlike incremental fit indices, absolute fit indices evaluate the fit of the proposed model without comparing it against the null model. Instead, they evaluate model fit by calculating the proportion of variance explained by the model in the sample variance/covariance matrix. Absolute fit indices include the goodness-of-fit index (GFI) and the adjusted GFI (AGFI). Third, residual fit indices concern the average difference between the observed and the model-implied variance/covariance matrices. Examples are the standardized root mean square residual (SRMR) and the root mean square error of approximation (RMSEA). Fourth, predictive fit indices examine the likelihood of the model to fit in similarly sized samples from the same population. Examples include the Akaike information criterion (AIC), the consistent Akaike information criterion (CAIC), and the expected cross-validation index (ECVI).

The question of which fit indices should be reported has been discussed extensively in SEM literature. We recommend Kline (2011, pp. 209-210) and studies such as Hu and Bentler (1998, 1999) and Bandalos and Finney (2010), as they all summarize the literature remarkably well and clearly present how to evaluate model fit. Kline recommends reporting (a) the chi-square statistic with its degrees of freedom and *p* value, (b) the matrix of correlation residuals, and (c) approximate fit indices (i.e., RMSEA, GFI, CFI) with the *p* value for the close-fit hypothesis for RMSEA. The close-fit hypothesis for RMSEA tests the hypothesis that the obtained RMSEA value is equal to or less than .05. This hypothesis is similar to the use of the chi-square statistic as an indicator of model fit and failure to reject it is favorable and supports the proposed model. Additionally, Hu and Bentler (1998, 1999), Bandalos and Finney (2010), and numerous others recommend reporting SRMR, since it shows the average difference between the observed and the model-implied variance/covariance matrices. There are at least three reasons for this. First, this average difference is easy to understand by readers who are familiar with correlations but less familiar with fit indices. Hu and Bentler (1995) emphasize this, stating that the minimum difference between the observed and the model-implied variance/ covariance matrices clearly signals that the proposed model accounts for the variances/covariances very well. Second, a reason for valuing the SRMR that is probably more fundamental is that it is a precise representation of the objective of SEM, which is to reproduce, as closely as possible, the model-implied variance/covariance matrix using the observed variance/covariance

matrix. Third, calculation of the SRMR does not require chi-squares. Since chi-squares are dependent on sample size, this indicates that the SRMR, which is not based on chi-squares, is not affected by sample size. This is in contrast with other fit indices (e.g., CFI, GFI, RMSEA), which use chi-squares as part of the calculation. For the assessment and academic achievement data, the chi-square is 323.957 with 24 degrees of freedom at the probability level of .464 ($p > .05$). The matrix of correlation residuals is presented in Table 1. If the model is correct, the differences between sample covariances and implied covariances should be small. Specifically, Kline argues that differences exceeding |0.10| indicate that the model fails to explain the correlation between variables. However, no such cases are found in the current data. Each residual correlation can be divided by its standard error, as presented in Table 2. This is the same as a statistical significance test for each correlation. The well-fitting model should have values of less than |2|. All cases are statistically nonsignificant. The RMSEA, GFI, and CFI are 0.000 (90% confidence interval: 0.000, 0.038), .989, and 1.000, respectively. The p value for the close-fit hypothesis for RMSEA is .995, and the close-fit hypothesis is not rejected. The SRMR is .025. Taken together, it may be reasonable to state that the proposed model of the relationship between self-assessment, teacher assessment, and academic achievement is supported.

The estimated model is presented in Figure 2. The parameter estimates presented here are all standardized as this facilitates the interpretation of parameters. Unstandardized parameter estimates also appear in an SEM output and these should be reported as in Table 3 because they are used to judge statistical significance of parameters along with standard errors. Factor loadings from the factors to the observed variables are high overall ($\beta = .505$ to .815), thereby suggesting that the three measurement models of self-assessment, teacher assessment, and academic achievement were each measured well in the current data. A squared factor loading shows the proportion of variance in the observed variable that is explained by the factor. For example, the squared factor loading of English for self-assessment indicates that self-assessment explains 53% of the variance in English for self-assessment ($.731 \times .731$). The remaining 47% of the variance is explained by the measurement error ($.682 \times .682$). In other words, the variance in the observed variable is explained by the underlying factor and the measurement error. Finally, the paths from the self-assessment and teacher assessment factors to the academic achievement factor indicate that they moderately affect academic achievement ($\beta = .454$ and .358). The correlation between self-assessment and teacher assessment is rather small (–.101), thereby indicating almost no relationship between them.

Table 1. Correlation residuals

		Self-assessment			Teacher assessment			Academic achievement		
		English	Mathematics	Science	English	Mathematics	Science	English	Mathematics	Science
Self-assessment	English	–								
	Math	0.008	–							
	Science	–0.026	0.005	–						
Teacher	English	0.032	–0.003	–0.023	–					
assessment	Math	0.002	0.015	0.036	0.003	–				
	Science	–0.013	–0.065	–0.014	–0.006	–0.001	–			
Academic	English	0.002	–0.064	0.046	0.048	0.008	–0.023	–		
achievement	Math	0.012	–0.016	0.047	–0.011	0.003	–0.032	0.007	–	
	Science	0.081	0.010	0.029	–0.083	–0.009	0.058	0.000	–0.012	–

Table 2. Standardized correlation residuals

		Self-assessment			Teacher assessment			Academic achievement		
		English	Mathematics	Science	English	Mathematics	Science	English	Mathematics	Science
Self-assessment	English	–								
	Mathematics	0.102	–							
	Science	–0.380	0.092	–						
Teacher	English	0.389	–0.045	–0.358	–					
assessment	Mathematics	0.030	0.314	0.757	0.043	–				
	Science	–0.219	–1.297	–0.284	–0.085	–0.014	–			
Academic	English	0.029	–0.959	0.710	0.567	0.135	–0.371	–		
achievement	Mathematics	0.211	–0.349	1.037	–0.186	0.062	–0.718	0.106	–	
	Science	1.340	0.195	0.608	–1.304	–0.184	1.224	–0.005	–0.261	–

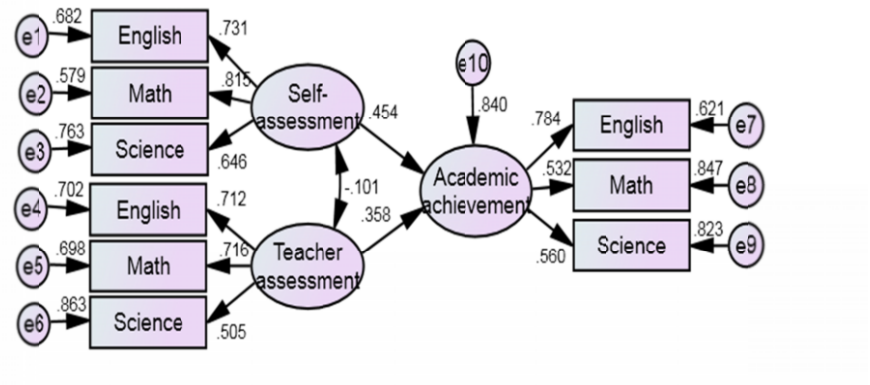

Figure 2. Example of an SEM model with standardized estimates

Model Respecification

Fifth, model re-specification is concerned with improving the model-data fit, for example, by deleting statistically nonsignificant paths or adding paths to the model. Any decision must be theoretically defensible and should not be statistically driven. The results are no longer confirmatory and should be viewed as explanatory. For the assessment and academic achievement data, we could, for example, delete the correlation between self-assessment and teacher assessment as it is very small in size ($r = -.101$) and statistically nonsignificant. This could be done only if it were supported by previous studies. Since this is not the case, no change is made in the model.

Table 3. Unstandardized and standardized estimates

Parameter		B	Standard error	β
Self-assessment –>	English	1.000[a]	–	.731
	Mathematics	.910*	.073	.815
	Science	.703*	.060	.646
Teacher assessment –>	English	1.000[a]	–	.712
	Mathematics	.736*	.086	.716
	Science	.528*	.066	.505
Academic achievement –>	English	1.000[a]	–	.784
	Mathematics	.483*	.060	.532
	Science	.534*	.065	.560
Self-assessment –>	Academic achievement	.498*	.072	.454
Teacher assessment –>	Academic achievement	.380*	.073	.358
Self-assessment <–>	Teacher assessment	–0.092	.058	-.101

Note. [a]*Fixed to 1.000 for scale identification. *p < .05. B refers to unstandardized estimates. β refers to standardized estimates.*

SOME KEY ISSUES

Thus far, we have discussed an SEM analysis with minimal details. In practice, there are several other issues that must be considered in order to use SEM appropriately. We will discuss these issues surrounding data screening, model fit indices, and sample size because of their prevalence in SEM.

Data Screening

Before being put to appropriate use, SEM must undergo data screening. Such preliminary analysis may initially seem tedious; however, if it is done properly, it often saves time and leads to a more precise understanding of the results. Data screening is often discussed in terms of linearity, data normality, outliers, and missing data. Researchers examine these issues in slightly different ways. Readers are referred to Byrne (2006, 2010), Kline (2011), and Tabachnick and Fidell (2007) for further details.

Linearity. SEM models are estimated by examining the relationship – usually a linear one – among measured variables that are represented in the variance/covariance matrix (or the correlation matrix and standard deviations). Such a linear relationship between variables is called linearity: One variable increase/decreases in proportion to a change in another variable. Figure 3A shows an example of this relationship. As with regression and factor analysis, excessive linearity is problematic. This can be examined through inspection of scatterplots or correlation matrices. For example, high correlations among variables (e.g., +/–.90; Tabachnick & Fidell, 2007) – also called multicollinearity – are troublesome. Table 4 shows that the correlations between the observed variables range from –.103 to .601. They are not high enough to cause a problem. Statistical tests for multicollinearity are also available, which include squared multiple correlations, tolerance, and the variance inflation factor. These tests are also used in statistical analysis in general and are not limited to SEM. High linearity can be adjusted for by deleting or aggregating redundant variables.

Nonlinear relationships can also be examined in quadratic or cubic models. A quadratic relationship is one in which one variable affects another up to some point, after which the effect levels off or decreases. Figure 3B shows a data distribution that looks like an inverse U-shape, where as one variable increases (1, 2, 3, 4, 5, 6, 7, 8) the other increases and then decreases (2, 3, 4, 5, 4, 3, 2, 1). A cubic relationship is similar to a quadratic relationship—one variable affects another up to some point, the effect levels off or decreases beyond that point, but this time comes back to influence once again after a certain point. Figure 3C shows a cubic relationship. Quadratic and cubic relationships are also called curvilinear relationships. Figure 3D shows an interactive relationship, in which scores in one group increase while those in the other group decrease. It is possible that a moderator variable is at play. It should be noted that there are a variety of nonlinear relationships in addition to those presented in Figures 3B, 3C, and 3D

(e.g., U-shaped relationship for a quadratic one). As a standard SEM assumes linear relations, modeling a nonlinear effect requires advanced techniques (see Kline, 2005, 2011; Marsh, Wen, Nagengast, & Hau, 2012).

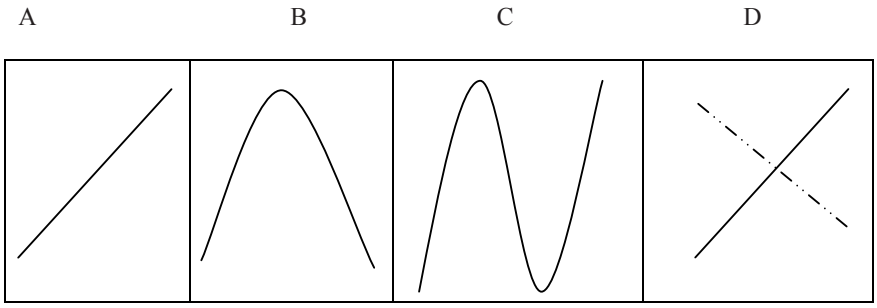

Figure 3. Linear, quadratic, cubic, and interactive relationships

Data normality. Data normality is divided into univariate normality and multivariate normality. Univariate normality refers to the situation in which one variable is normally distributed. Multivariate normality refers to the situation in which, in addition to the normality of each variable, each variable is also normally distributed for each other variable (Tabachnick & Fidell, 2007). Numerous SEM application studies use the maximum likelihood estimation method. This method assumes multivariate normal distribution of the data for the dependent (i.e., endogenous) variable. Although maximum likelihood methods are robust against non-normality, it is still important to assess whether the data satisfy the assumption of normality. Since multivariate normality is related to univariate normality, both types of normality need to be examined.

Univariate normality can be examined by inspecting absolute skewness and kurtosis values or the statistical significance of those values. First, with regard to the inspection of skewness and kurtosis, data normality is ensured when both values are zero. Unfortunately, there are no clear-cut guidelines on the degree of non-normality. Kline (2011) reviewed relevant studies (e.g., Curran, West, & Finch, 1996) and suggested viewing skewness and kurtosis exceeding 3 and 20 respectively as extremely non-normal. Note that this is a rule-of-thumb and is not an agreed-upon definition. For example, Curran et al. (1996) consider a skewness of 2 and a kurtosis of 7 as moderately non-normal, and a skewness of 3 and a kurtosis of 21 as severely non-normal. Chou and Bentler (1995) and Muthén and Kaplan (1985) argue that skewness and kurtosis values approaching 2 and 7, respectively, indicate problems. Table 4 shows that skewness and kurtosis values for all the observed variables are well below these cut-offs and are in fact very near to zero, thereby indicating that the data are univariately normal.

Table 4. Correlations between variables and their descriptive statistics

		Self-assessment			Teacher assessment			Academic achievement		
		English	Mathematics	Science	English	Mathematics	Science	English	Mathematics	Science
Self-assessment	English	–								
	Mathematics	.601	–							
	Science	.452	.531	–						
Teacher assessment	English	–.034	–.061	–.063	–					
	Mathematics	–.052	–.044	–.011	.513	–				
	Science	–.048	–.103	–.046	.356	.361	–			
Academic achievement	English	.241	.220	.246	.202	.182	.106	–		
	Mathematics	.172	.164	.193	.109	.122	.050	.423	–	
	Science	.235	.200	.180	.063	.117	.146	.439	.285	–
Mean		4.089	4.102	4.088	4.077	4.158	4.078	4.177	4.072	4.063
SD		1.284	1.048	1.021	1.361	0.996	1.013	1.313	0.935	0.981
Minimum		0.412	1.099	1.281	0.040	1.724	1.061	0.468	1.601	1.171
Maximum		8.423	6.984	6.776	8.625	7.549	7.383	7.712	7.230	6.736
Skewness		0.095	0.035	0.016	–0.065	0.148	–0.035	–0.119	0.128	–0.072
z		0.819	0.299	0.137	–0.565	1.273	–0.298	–1.031	1.106	–0.623
Kurtosis		0.135	0.006	–0.326	–0.360	0.208	0.164	–0.166	–0.218	–0.013
z		0.520	–0.031	–1.453	–1.599	0.831	0.644	–0.768	–0.992	–0.112

Second, the statistical significance of skewness and kurtosis also serves as an indicator of data normality. In particular, the critical ratio or *z* value is computed by dividing either skewness or kurtosis by its standard error. Data normality is ensured when the absolute value is within +/– 2.58 ($p < .01$) or 3.29 ($p < .001$). However, as emphasized by Kline (2011) and Tabachnick and Fidell (2007), the standard errors of skewness and kurtosis shrink in large sample sizes, which can produce statistically significant skewness and kurtosis values even when the data distribution appears normal. Thus, with large samples, making substantive decisions on the basis of the visual inspection of the data – for example, using histograms or box plots – is preferred. However, it is difficult to define what is meant by a large sample. For example, Byrne (2006, 2010) only uses absolute skewness and kurtosis values for her dataset with a sample size of 372. Ullman (2007) uses both absolute values and statistical significance of skewness and kurtosis for her two datasets with sample sizes 175 and 459. In actuality, it appears that researchers are more likely to use estimation methods that are robust against non-normality, such as Satorra-Bentler correction or weighted least square parameter estimate methods. In any case, Table 4 shows that *z* values for skewness and kurtosis are all within +/–2.58 ($p < .01$) or 3.29 ($p < .001$), thereby suggesting data normality.

Additionally, multivariate normality can be measured using Mardia's coefficient of multivariate kurtosis. The statistical significance of Mardia's coefficient is examined using a *z* value, but this time using the *z* values of 5 or 6, not +/–2.58 ($p < .01$) or 3.29 ($p < .001$), since Bentler (2005) argues that multivariate non-normality would not affect the model in practice unless its values were 5, 6, or above. Univariate normality can be estimated using general-purpose software programs (e.g., SAS or SPSS) or SEM programs, whereas multivariate normality can only be estimated using SEM programs (for SEM programs, see the Software section). Mardia's coefficient for the current data is –.157 with a *z* value of –.119. This indicates the multivariate normality of the data.

As seen above, numerous issues surrounding the treatment of non-normal data complicate decision making during data analysis. We reviewed previous studies and found Finney and DiStefano (2006) the most accessible, synthetic, and up to date. They summarize relevant studies and recommend that, for continuous data, if the variables are approximately normally distributed, the maximum likelihood estimation is recommended; if the variables are moderately non-normal (skewness < 2 and kurtosis < 7) the maximum likelihood estimation or Satorra-Bentler correction method are recommended; if the variables are severely non-normal (skewness > 2 and kurtosis > 7), the Satorra-Bentler correction or bootstrapping methods is recommended. For categorical data, regardless of the number of categories, they recommend using weighted least square parameter estimates (WLSMV), available in the SEM program Mplus. If Mplus is not available, they recommend that if the variables are approximately normally distributed, the maximum likelihood estimation should be used for scales with five or more categories and the Satorra-Bentler correction method for scales with four or more categories. This also applies to moderately non-normal data (skewness < 2 and

kurtosis < 7). If the variables are severely non-normal (skewness > 2 and kurtosis > 7), the Satorra-Bentler correction method is recommended.

Outliers. An outlier is an extremely large or small value of one variable (a univariate outlier) or a combination of such values of two or more variables (a multivariate outlier). Univariate outliers can be detected by drawing a histogram or inspecting the z values of variables using, for example, the SPSS EXPLORE or DESCRIPTIVES functions. Multivariate outliers can be detected using the Mahalanobis distance (i.e., Mahalanobis d-squared) statistic. It shows how one observation in the data is distantly located from the others. It is distributed as a chi-square statistic with degrees of freedom equal to the number of observed variables. Observations are arranged according to the size of the statistics, and those exceeding the critical value of the chi-square given degrees of freedom (e.g., $p < .001$) can be judged as outliers. For the current data, the histograms appear normal. There are five responses out of 4050 (450 × 9 items) exceeding the z value 3.29 ($p < .001$). As this is just 0.001% of the total responses, it is considered negligible. With regard to multivariate outliers, the critical value of chi-square for 24 degrees of freedom is 51.179. The most deviated case was participant 4, whose responses produced a Mahalanobis distance of 27.192—still below 51.179. Taken together, it is reasonable to say that the current dataset does not include univariate or multivariate outliers.

Missing data. The ideal situation is to be able to analyze a complete dataset that contains all examinees' responses to all items. In reality, this rarely occurs and one often has to analyze a dataset with missing values. Therefore, how to treat missing data is a widely discussed issue in the application of statistics, including SEM. Missing data treatment is classified into three types: (a) the deletion of those data, (b) the estimation of those data, and (c) the use of parameter estimation methods that take missingness into consideration. Deletion of missing data is a traditional approach, and includes listwise deletion (elimination of all cases with missing values from subsequent analysis) or pairwise deletion (removal of paired cases in which at least one case has a missing value). Although both methods are easy to implement, they may result in substantial loss of data observations. More importantly, Muthén, Kaplan, and Hollis (1987) argue that the two methods work only when data are missing completely at random, a case that is often violated in practice. Thus, both listwise and pairwise deletion methods may bias results if data missingness is not randomly distributed through the data (Tabachnick & Fidell, 2007).

A preferred approach is to estimate and impute missing data. Methods abound, such as mean substitution, regression, and expectation maximization methods; however, according to Tabachnick and Fidell (2007), the most recommended method is multiple imputation (Rubin, 1987). It replaces missing values with plausible values that take into account random variation.

Another way to address missing data is to use parameter estimation methods that take missingness into consideration. This is implemented in (full information)

maximum likelihood estimation, which uses all available data regardless of completeness (e.g., Enders, 2001). Both expectation maximization and maximum likelihood estimation methods are available in SEM programs. As the current data do not include missing responses, it is not necessary to eliminate, estimate, or impute such responses.

Model Fit Indices

Although no agreed-upon guidelines exist regarding which fit indices should be reported, some recommendations can be found in the literature. In an often-cited article, Hu and Bentler (1998, 1999) recommend reporting (a) the SRMR, and (b) the CFI, TLI, RMSEA, or other indices (e.g., Gamma Hat, Incremental Fit Index [IFI], McDonald's Centrality Index [MCI], Relative Noncentrality Index [RNI]). Similarly, Kashy, Donnellan, Ackerman, and Russell (2009) recommend reporting the CFI or TLI along with the chi-square and RMSEA. Bandalos and Finney (2010) recommend the chi-square, CFI, TLI, RMSEA, and SRMR, whereas Mueller and Hancock (2010) recommend RMSEA and its confidence interval, the SRMR, and at least one of CFI, NFI, and TLI. Widaman (2010) encourages reporting the chi-square, CFI, TLI, and RMSEA. For testing measurement invariance across groups (e.g., whether the factor loadings are the same across groups), Cheung and Rensvold (2002) recommend reporting the CFI, Gamma Hat, and McDonald's Noncentrality Index and interpreting a reduction in each index as evidence of measurement invariance. Summarizing studies that provide guidelines for reporting fit indices, In'nami and Koizumi (2011) report that the indices recommended most often are the chi-square (with degrees of freedom and *p* values), CFI, TLI, RMSEA (and its confidence interval), and the SRMR.

Sample Size

One rule-of-thumb is that a sample size below 100, between 100 and 200, and over 200 is often considered small, medium, and large, respectively (Kline, 2005). Similarly, Ding, Velicer, and Harlow (1995) argue that the minimum sample size adequate for analysis is generally 100 to 150 participants. Another approach is to consider model complexity in terms of the ratio of the sample size to the number of free parameters to be estimated in a model. A minimum sample size is at least 10 times the number of free model parameters (Raykov & Marcoulides, 2006). For example, a model with 30 free parameters would require at least 300 observations (30×10). Nevertheless, as the authors of the aforementioned articles emphasize, these are only rough guidelines. This is particularly because the requisite sample size depends on numerous factors, including the number and patterns of missing data, strength of the relationships among the indicators, types of indicators (e.g., categorical or continuous), estimation methods (e.g., [robust] maximum likelihood, robust weighted least squares), and reliability of the indicators. Complex issues surrounding sample size determination seem to hamper creating definitive rules –

or even rules of thumb – concerning necessary sample size (e.g., Mundfrom, Shaw, & Ke, 2005).

Instead of elaborating on general guidelines for sample size, more empirically grounded, individual-model-focused approaches to determining sample size in relation to parameter precision and power have been proposed. These approaches include Satorra and Saris (1985), MacCallum, Browne, and Sugawara (1996), and Muthén and Muthén (2002). The methods of both Satorra and Saris (1985) and MacCallum et al. (1996) estimate sample size in terms of the precision and power of an entire model using the chi-square statistic and RMSEA, respectively. In contrast, Muthén and Muthén (2002) evaluate sample size in terms of the precision and power of individual parameters in a model, while allowing the modeling of various conditions that researchers frequently encounter in their research, such as non-normality or type of indicator. Such modeling flexibility is certainly useful for estimating sample size, given that sample size and many variables affect each other in intricate ways.

In order to evaluate sample size, Muthén and Muthén (2002) use four criteria. First, parameter bias and standard error bias should not exceed |10%| for any parameter in the model. Second, the standard error bias for the parameter for which power is of particular interest should not exceed |5%|. Third, 95% coverage – the proportion of replications for which the 95% confidence interval covers the population parameter value – should fall between 0.91 and 0.98. One minus the coverage value equals the alpha level of 0.05. Coverage values should be close to the correct value of 0.95. Finally, power is evaluated in terms of whether it exceeds 0.80 – a commonly accepted value for sufficient power.

An analysis of the sample size of the current data based on Muthén and Muthén (2002) is presented in Table 5. Columns 2 and 3 show population and sample parameters. Population parameters are unstandardized parameters in Table 3. They are viewed as correct, true parameters from which numerous samples (replications) are generated in each run, and results over the replications are summarized. For example, using these values, the parameter bias for self-assessment measured by mathematics is calculated in the following manner: |0.9130 – 0.910|/|0.910| = 0.00330, or in other words, 0.330%. This is far below the criterion of 10%, thereby suggesting a good estimation of the parameter. The result is presented in Column 4. Column 5 shows the standard deviation of the parameters across replications. Column 6 shows the average of the standard errors across replications. The standard error bias for self-assessment measured by mathematics is |0.0743 – 0.0754|/|0.0754| = 0.01459, or in other words, 1.459%. This is again far below the criterion of 10%, thereby suggesting a good estimation of the parameter. The result is presented in Column 7. In particular, we are interested in the effect of self-assessment and teacher assessment on academic achievement. The standard error biases for these parameters of interest are 0.413% and 0.545%, respectively. Neither exceeds 5%, thereby suggesting a good estimation of the parameter. Column 8 provides the mean square error of parameter estimates, which equals the variance of the estimates across replications plus the squared bias (Muthén & Muthén, 2007). Column 9 shows coverage, or the proportion of

replications where the 95% confidence interval covers the true parameter value. The value of 0.947 for self-assessment measured by mathematics is very close to 0.95, thereby suggesting a good estimation of the parameter. The last column shows the percentage of replications for which the parameter is significantly different from zero (i.e., the power estimate of a parameter). Column 10 shows that the power for self-assessment measured by mathematics is 1.000, which exceeds 0.80 and suggests sufficient power for the parameter. Together, these results provide good evidence for parameter precision and power for self-assessment measured by mathematics and suggest that the sample size for self-assessment measured by mathematics is sufficient. The same process is repeated for the remaining parameters. It should be noted that the power for the correlation between self-assessment and teacher assessment is low (0.339; see the last row). This suggests that the current sample size of 450 is not enough to distinguish the correlation from zero. Thus, although the sample size for the current model is adequate overall, the underpowered correlation indicates that caution should be exercised when interpreting it. The Appendix shows the Mplus syntax used for the current analysis.

Table 5. Mplus output for the Monte Carlo analysis to determine the precision and power of parameters

	Population parameter	Sample parameters averaged	Parameter bias	SD of sample parameters	Standard error of sample parameters	Standard error bias	Mean square error of parameters	95% coverage	Power
Self-assessment by									
English	1	1	0	0	0		0	1	0
Mathematics	0.910	0.9130	0.330	0.0754	0.0743	1.459	0.0057	0.947	1.000
Science	0.703	0.7024	0.085	0.0613	0.0609	0.653	0.0038	0.950	1.000
Teacher assessment by									
English	1	1	0	0	0	0	0	1	0
Science	0.736	0.7451	1.236	0.0890	0.0876	1.573	0.0079	0.951	1.000
Mathematics	0.528	0.5318	0.720	0.0662	0.0666	0.604	0.0044	0.953	1.000
Academic achievement by									
English	1	1	0	0	0	0	0	1	0
Mathematics	0.483	0.4811	0.393	0.0613	0.0604	1.468	0.0038	0.945	1.000
Science	0.534	0.5310	0.562	0.0658	0.065	1.216	0.0043	0.947	1.000
Academic achievement On									
Self-Assessment	0.498	0.5025	0.904	0.0726	0.0723	0.413	0.0053	0.949	1.000
Teacher assessment	0.380	0.3810	0.263	0.0734	0.0730	0.545	0.0054	0.945	1.000
Self-assessment With									
Teacher assessment	–0.092	–0.0894	2.826	0.0586	0.0578	1.365	0.0034	0.951	0.339

Note. The column labels were slightly changed from original Mplus outputs to enhance clarity. Self-assessment by English refers to a path from the self-assessment factor to the English variable. Self-assessment with Teacher assessment refers to the correlation between these two factors.

VARIOUS SEM ANALYSES

Various types of models can be analyzed within the SEM framework. In addition to the models presented in Figures 1 and 2, we describe models often used in educational studies: confirmatory factor analysis, multiple-group analysis, and latent growth modeling. First, confirmatory factor analysis is used to examine whether the factor structure of a set of observed variables is consistent with previous theory or empirical findings (e.g., Brown, 2006). The researcher constructs a model using knowledge of the theory and/or empirical research, postulates the relationship pattern, and tests the hypothesis statistically. This reinforces the importance of theory in the process of model building. The models of self-assessment, teacher assessment, and academic achievement in Figures 1 and 2 represent different measurement models and must be verified through confirmatory factor analysis in terms of whether each of the three constructs are well represented by the three measurements of English, mathematics, and science. Unfortunately, each measurement model has only three observed variables, and this results in zero degrees of freedom (6 parameters to estimate – two factor loadings, three measurement errors, and one factor variance – and $3(3 + 1)/2 = 6$ data points). The measurement models cannot be evaluated in the current model specification (see model identification in the Five Steps in an SEM Application above).

Various models can be analyzed using confirmatory factor analysis. For example, the often-cited study Holzinger and Swineford (1939) administered a battery of tests to measure seventh- and eighth-grade students in two Chicago schools. The tests were designed to measure mental ability, hypothesized to comprise spatial, verbal, speed, memory, and mathematics abilities. Although Holzinger and Swineford (1939) did not use SEM, the model closest to the one they hypothesized is shown in Figure 4A, and competing models that we postulated are shown in Figures 4B, 5A, and 5B. Figure 4A shows that mental ability comprises a general ability and five sub-abilities. Figure 4B is similar to Figure 4A but assumes a hierarchical relationship between a general ability and sub-abilities. Figure 5A assumes only a single general ability. Figure 5B hypothesizes no general ability and instead assumes correlated sub-abilities. A series of models can be tested on a single dataset using SEM by comparing model fit indices or using a chi-square difference test (see, for example, Brown, 2006; Shin, 2005).

Second, multiple-group or multiple-sample analysis aims to fit a model to two or more sets of data simultaneously. It allows us to test whether and to what extent measurement instruments (tests and questionnaires) function equally across groups, or, put another way, whether and to what extent the factor structure of a measurement instrument or theoretical construct of interest holds true across groups (e.g., Bollen, 1989). Multiple-group analysis involves testing across the samples whether factor loadings, measurement error variances, factor variances, and factor covariances are the same. Equivalence across groups suggests the cross-

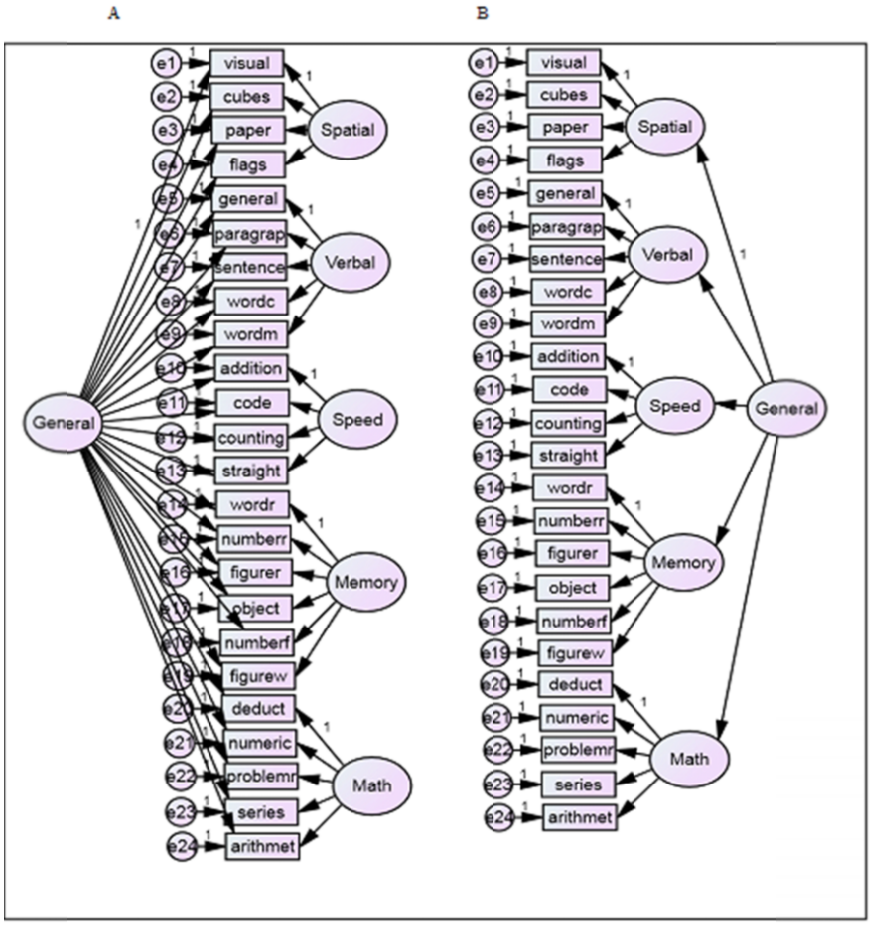

Figure 4. Confirmatory factor analysis of a model of mental ability: Bi-factor model (left) and higher-order model (right). The spatial test battery comprises (1) visual perception, (2) cubes, (3) paper form board, and (4) flags. The verbal test battery comprises (5) general information, (6) paragraph comprehension, (7) sentence completion, (8) word classification, and (9) word meaning. The speed test battery comprises (10) addition, (11) coding, (12) counting groups of dots, and (13) straight and curved capitals. The memory test battery comprises (14) word recognition, (15) number recognition, (16) figure recognition, (17) object-number, (18) number-figure, and (19) figure-word. The math test battery comprises (20) deduction, (21) numerical puzzles, (22) problem reasoning, (23) series completion, and (24) Woody-AcCall Mixed Fundamentals.

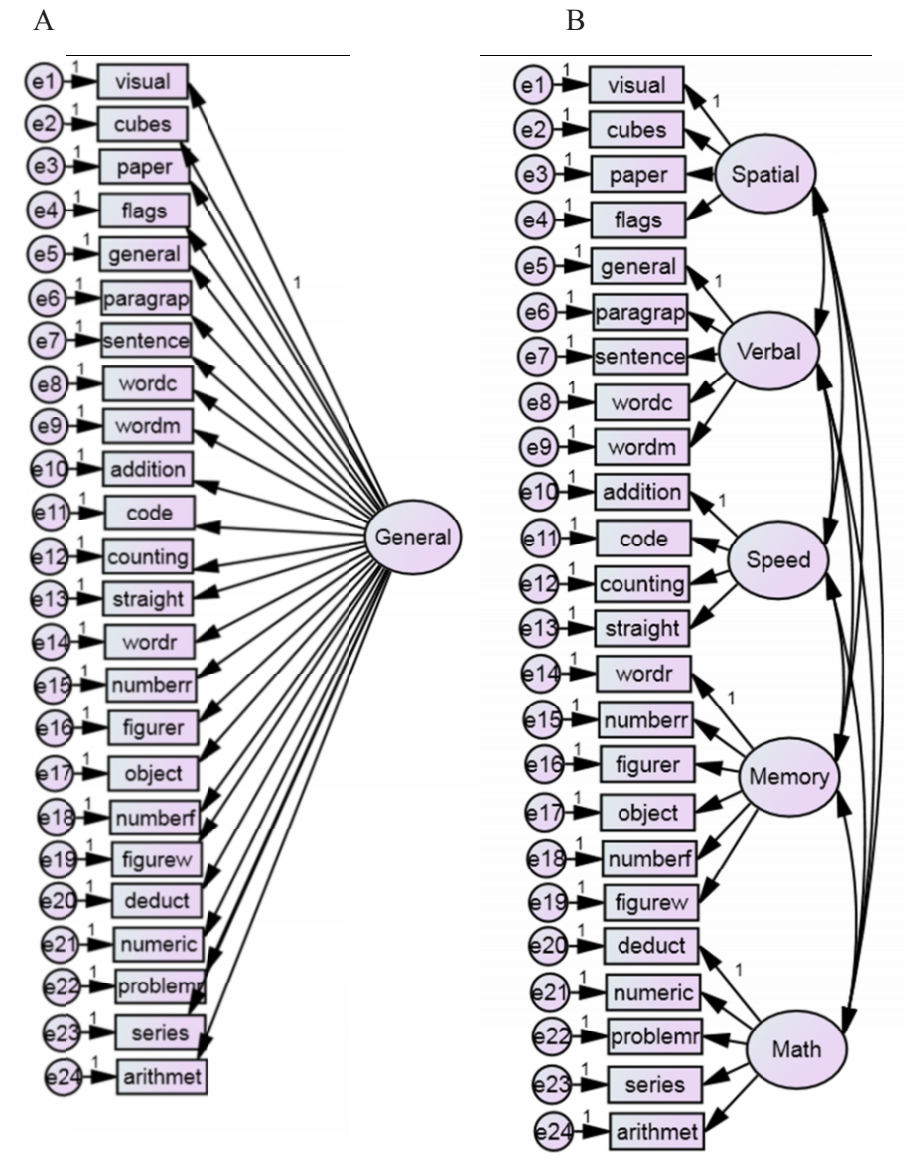

Figure 5. Confirmatory factor analysis of a model of mental ability: Single-factor model (left) and correlated-factor model (right)

validation or generalizability of findings. It is possible that factor loadings are similar in size across groups, while factor covariances are different. For example, Holzinger and Swineford's (1939) data include seventh- and eighth-grade students of both genders from two different schools. It would be of interest to examine whether the bi-factor model of mental ability in Figure 4A is stable across grades, gender, and/or schools. For examples of applications of multiple-group analysis, see Byrne, Baron and Balev (1998), In'nami and Koizumi (2012), and Shin (2005).

Third, latent growth modeling is useful for evaluating longitudinal changes in aspects of individuals over time. It provides a great deal of information, including change at the individual and the group levels, pattern of change (e.g., linear, quadratic), and variables related to change, such as age, gender, motivation, and socioeconomic status (i.e., income, education, and occupation). For example, Tong et al. (2008) hypothesized that second language oral proficiency develops linearly when measured by vocabulary and listening tests at three time points over two years. Their model is presented in Figure 6A. Initial status, also called intercept,

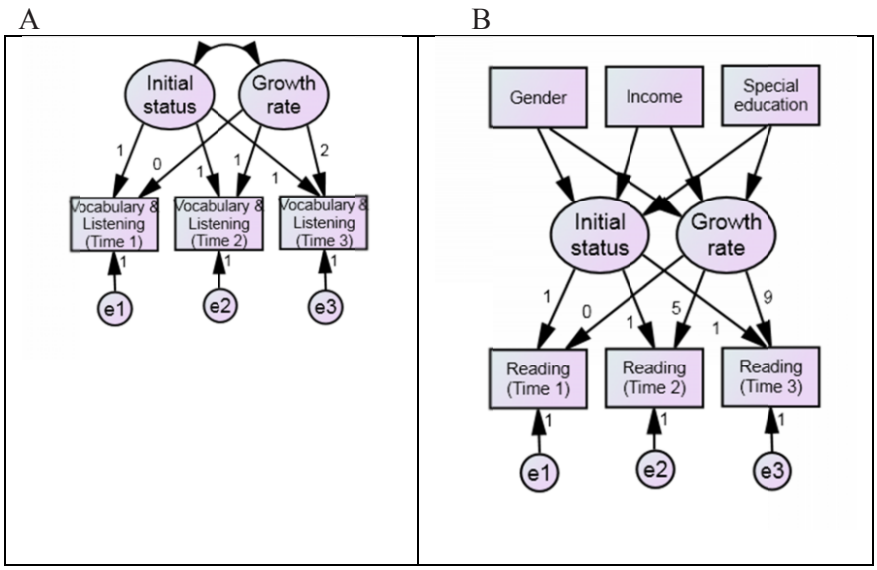

Figure 6. Latent growth model of oral proficiency (left) and of reading proficiency with external variables (right)

indicates the level of proficiency at the beginning of the study. Growth rate, also called slope, indicates the speed at which change is observed at each measurement point. The loadings for the initial status factor are all fixed to be 1, whereas those for the growth rate are fixed to be 0, 1, and 2 to model a linear growth rate. Note

that all factor loadings are fixed, unlike confirmatory factor analysis and multiple-group analysis.

More complex models can also be analyzed using latent growth modeling. Yeo, Fearrington, and Christ (2011) investigated how demographic variables – gender, income, and special education status – affect reading growth at school. Their model, shown in Figure 6B, differs primarily from the model in Figure 6A in two ways. First, the loadings for the growth rate factor are fixed to be 0, 5, and 9 – three time points of data collection (August = 0, January = 5, and May = 9) – because we assume that the authors were interested in nine-month growth and rescaled the slope factor loadings accordingly. It should be noted that growth rate factors, whether fixed to be 1, 2, and 3, or 0, 5, and 9, do not change the data-model fit (e.g., Hancock & Lawrence, 2006). Second, the three demographic variables are incorporated into the model as predictors of initial status and growth rate. The results indicate the relative impact of the external variables on the initial level of reading proficiency and on the growth rate of reading proficiency over nine months. For further examples of latent growth modeling, see Kieffer (2011) and Marsh and Yeung (1998).

SOFTWARE

Since Byrne (2012a) provides a detailed, comparative review of SEM software, we will present only a brief treatment of SEM software programs (also see Narayanan, 2012). There are several major commercial programs for performing SEM, including Amos (Analysis of Moment Structures; Arbuckle, 1994-2012), CALIS (SAS Institute, 1990-2012), EQS (Equations; Bentler, 1994-2011), LISREL (Linear Structural Relationships; Jöreskog & Sörbom, 1974-2012), and Mplus (Muthén & Muthén, 1998-2012). Free programs are also available, including Mx (Neale, Boker, Xie, & Maes, 2003) and three R-language packages: the OpenMx package (Boker, Neale, Maes, Wilde, Spiegel, Brick, et al., 2007-2012), the "sem" package (Fox, Nie, Byrnes, Culbertson, Friendly, Kramer, & Monette, 2012), and the "lavaan" package (Rosseel, 2012). The choice of software depends on the purpose of the SEM analysis and the proficiency of the user's computing skills. Byrne (2012a) indicates three aspects related to deciding on the best software: (a) familiarity with SEM concepts and application, (b) the types of SEM model to be tested, and (c) preference concerning manual or graphic interface. She argues that beginners may find Amos or EQS the easiest to use, and that more advanced learners may prefer to use EQS, LISREL, or Mplus. Unlike Amos, EQS, and LISREL, Mplus requires command-based inputs, and learners who are used to graphic interfaces may need some time to become comfortable with the program. In order to familiarize themselves with software, novice learners are referred to Byrne (1998, 2006, 2010, 2012b), whereas advanced learners wishing to use R-based packages are referred to Fox, Byrnes, Boker, and Neale (2012).

SOME DIRECTIONS FOR LEARNING MORE ABOUT SEM

Since SEM is a versatile technique, a single book chapter would not be able to cover a wide range of analyses that can be modeled using SEM. In order to deepen learning regarding SEM, we recommend reading through Byrne (1998, 2006, 2010, 2012b) for LISREL, EQS, Amos, and Mplus, trying to analyze the accompanying datasets, and ensuring that one can replicate findings. Based on our own experience with Byrne (2010) for Amos, and Byrne (2006) for EQS datasets, as well as on discussion with skilled SEM users, we believe that this is probably the best approach to familiarize oneself with SEM and apply the techniques to one's own data.

For providing answers to questions that may arise with regard to particular issues related to SEM, the following recent references may be useful: Bandalos and Finney (2010), Brown (2006), Cudeck and du Toit (2009), Hancock and Mueller (2006), Hoyle (2012), Kaplan (2009), Kline (2011), Lomax (2010), Mueller and Hancock (2008, 2010), Mulaik (2009), Raykov and Marcoulides (2006), Schumacker and Lomax (2010), Teo and Khine (2009), and Ullman (2007). For more on how researchers should report SEM results, see Boomsma, Hoyle, and Panter (2012); Gefen, Rigdon, and Straub (2011); Jackson, Gillaspy Jr., and Purc-Stephenson (2009); Kahn (2006); Kashy, Donnellan, Ackerman, and Russell (2009); Martens (2005); McDonald and Ho (2002); Schreiber, Nora, Stage, Barlow, and King (2006); and Worthington and Whittaker (2006). Reporting a correlation matrix with means and standard deviations is strongly recommended as this allows one to replicate a model, although replication of non-normal and/or missing data requires raw data (for example, see In'nami & Koizumi, 2010). Of particular interest is the journal *Structural Equation Modeling: An Interdisciplinary Journal* published by Taylor & Francis, which is aimed at those interested in theoretical and innovative applied aspects of SEM. Although comprising highly technical articles, it also includes the Teacher's Corner, which features instructional modules on certain aspects of SEM, and book and software reviews providing objective evaluation of current texts and products in the field.

For questions pertaining to particular features of SEM programs, user guides are probably the best resource. In particular, we find the EQS user guide (Bentler & Wu, 2005) and manual (Bentler, 2005) outstanding, as they describe underlying statistical theory in a readable manner as well as stepwise guidance on how to use the program. A close look at manuals and user guides may provide answers to most questions. LISREL and Mplus users should take full advantage of technical appendices, notes, example datasets, and commands, which are all available online free of charge (Mplus, 2012; Scientific Software International, 2012). The Mplus website also provides recorded seminars and workshops on SEM and a schedule listing of upcoming courses.

For problems not addressed by the abovementioned resources, we suggest consulting the Structural Equation Modeling Discussion Network (SEMNET). It was founded in February 1993 (Rigdon, 1998) and archives messages by month. Because of the large number of archived messages collected over the past two

decades (thanks to the mushrooming popularity of SEM across many disciplines), SEMNET is a treasure trove of questions and answers on virtually every aspect of SEM. Questions should only be posted if answers to them cannot be found in the archive. As with any other academic online discussion forum, contributors to SEMNET take questions seriously and spend precious time responding to them. We recommend that anyone wishing to receive a good reply should mention that answers were not found in the archive and articulate problems in enough detail for others to respond. Posting a command/script/syntax file is a good idea.

SEM is constantly evolving and expanding. The development and application of new techniques are causing numerous academic disciplines to move increasingly toward a better understanding of various issues that require tools that are more precise. SEM analysis offers powerful options for analyzing data from educational settings, and techniques discussed in this chapter will enable educational researchers to be in a better position to address a wide range of research questions. By employing SEM analysis appropriately, we will be able to contribute much in years to come.

APPENDIX

Mplus Input for the Monte Carlo Analysis for Determining the Precision and Power of Parameters

```
TITLE:          THREE-FACTOR, NORMAL DATA, NO MISSING
MONTECARLO:
            NAMES ARE X1-X9;
            NOBSERVATIONS = 450; ! SAMPLE SIZE OF INTEREST
            NREPS = 10000;
            SEED = 53567;
MODEL POPULATION:
            f1 BY X1@1 X2*.91 X3*.70;
            f2 BY X4@1 X5*.74 X6*.53;
            f3 BY X7@1 X8*.48 X9*.53;
            X1*.77; X2*.37; X3*.61; X4*.91; X5*.48; X6*.76; X7*.66; X8*.63;
            X9*.66;
            f1*.88; f2*.94; f3*.74;
            f3 ON f1*.50; f3 ON f2*.38;
            f1 WITH f2*-.09;
MODEL:
            f1 BY X1@1 X2*.91 X3*.70;
            f2 BY X4@1 X5*.74 X6*.53;
            f3 BY X7@1 X8*.48 X9*.53;
            X1*.77; X2*.37; X3*.61; X4*.91; X5*.48; X6*.76; X7*.66; X8*.63;
            X9*.66;
            f1*.88; f2*.94; f3*.74;
            f3 ON f1*.50; f3 ON f2*.38;
            f1 WITH f2*-.09;
ANALYSIS:       ESTIMATOR = ML;
OUTPUT:         TECH9;
```

REFERENCES

Arbuckle, J. L. (1994–2012). *Amos* [Computer software]. Chicago, IL: SPSS.

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93-114). New York: Routledge.

Bentler, P. M. (1994-2011). EQS for Windows [Computer software]. Encino, CA: Multivariate Software.

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

Bentler, P. M., & Wu, E. J. C. (2005). *EQS 6.1 for Windows user's guide*. Encino, CA: Multivariate Software.

Boker, S. M., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., et al. (2007–2012). OpenMx [Computer software]. Retrieved from http://openmx.psyc.virginia.edu/installing-openmx.

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

Bollen, K. A., & Long, J. S. (1993). Introduction. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 1-9). Newbury Park, CA: Sage.

Boomsma, A., Hoyle, R. H., & Panter, A. T. (2012). The structural equation modeling research report. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 341-358). New York: Guilford Press.

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

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

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

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Erlbaum.

Byrne, B. M. (2012a). 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: Guilford Press.

Byrne, B. M. (2012b). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.

Byrne, B. M., Baron, P., & Balev, J. (1998). The Beck Depression Inventory: A cross-validated test of second-order factorial structure for Bulgarian adolescents. *Educational and Psychological Measurement*, *58*, 241-251.

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

Chou, C., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37-55). Thousand Oaks, CA: Sage.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Cudeck, R., & du Toit, S. H. C. (2009). General structural equation models. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 515-539). London, UK: SAGE.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16-29.

Ding, L., Velicer, W. F., & Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor and improper solutions on structural equation modeling fit indices. *Structural Equation Modeling*, *2*, 119-143.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*, 128-141.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269-314). Greenwich, CT: Information Age.

Fox, J., Byrnes, J. E., Boker, S., & Neale, M. C. (2012). Structural equation modeling in R with the sem and OpenMx packages. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 325-340). New York: Guilford Press.

Fox, J., Nie, X., Byrnes, J., Culbertson, M., Friendly, M., Kramer, A., & Monette, G. (2012). sem: Structural equation modeling. Available from http://cran.r-project.org/web/packages/sem/index.html

Gefen, D., Rigdon, E. E., & Straub, D. (2011). An update and extension to SEM guidelines for administrative and social science research. *Management Information Systems Quarterly*, *35*, A1-A7.

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Erlbaum.

Hancock, G. R., & Lawrence, F. R. (2006). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course*. Greenwich, CT: Information Age.

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

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution* [Monograph]. Chicago: University of Chicago Press.

Hoyle, R. H. (Ed.). (2012). *Handbook of structural equation modeling*. New York: Guilford Press.

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.

Hu, L.-T., & Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424-453.

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

In'nami, Y., & Koizumi, R. (2010). Can structural equation models in second language testing and learning research be successfully replicated? *International Journal of Testing*, *10*, 262-273.

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly, 8,* 250-276.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing*, *29*, 131-152.

Jackson, D. L., Gillaspy Jr., J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*, 6-23.

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*, 199-218.

Jöreskog, K. G., & Sörbom, D. (1974–2012). *LISREL for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International.

Kahn, J. H. (2006). Factor analysis in counseling psychology research, training, and practice: Principles, advances, and applications. *The Counseling Psychology*, *34*, 684-718.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*, 1131-1142.

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.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Lomax, R. G. (2010). Structural equation modeling: Multisample covariance and mean structures. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 385-395). New York: Routledge.

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.

Marsh, H. W., Wen, Z., Nagengast, B., & Hau, K.-T. (2012). Structural equation models of latent interaction. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 436-458). New York: Guilford Press.

Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal*, *35*, 705-738.

Martens, M. P. (2005). The use of structural equation modeling in counseling psychology research. *The Counseling Psychologist*, *33*, 269-298.

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

Mplus. (2012). Retrieved from http://www.statmodel.com/.

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

Mueller, R. O., & Hancock, G. R. (2010). Structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 371-383). New York: Routledge.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton, FL: Chapman & Hall/CRC.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*, 159-168.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171-189.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *52*, 431-462.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus* [Computer software]. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599-620.

Narayanan, A. (2012). A review of eight software packages for structural equation modeling. *The American Statistician*, *66*, 129-138.

Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical modeling* (6th ed.). Richmond, VA: Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University.

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

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

Rigdon, E. (1998). SEMNET. Retrieved from http://www2.gsu.edu/~mkteer/semnet.html.

Rosseel, Y. (2012). lavaan: Latent variable analysis. Available from http://cran.r-project.org/web/packages/lavaan/.

Royce, J. R. (1963). Factors as theoretical constructs. In D. N. Jackson & S. Messick (Eds.), *Problems in human assessment*. New York: McGraw Hill.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

SAS Institute. (1990–2012). *SAS PROC CALIS* [Computer software]. Cary, NC: Author.

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

Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Simis, A., Snellings, 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.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis: A review. *Journal of Educational Research*, *99*, 323-337.

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling* (3rd ed.). New York: Routledge.

Scientific Software International. (2012). LISREL. Retrieved from http://www.ssicentral.com/lisrel/.

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

Silverman. S. K. (2010). What is diversity?: An inquiry into preservice teacher beliefs. *American Educational Research Journal*, *47*, 292-329.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Needham Heights, MA: Allyn & Bacon.

Teo, T., & Khine, M. S. (Eds.). (2009). *Structural equation modeling in educational research: Concepts and applications*. Rotterdam, the Netherlands: Sense Publishers.

Tong, F., Lara-Alecio, R., Irby, B., Mathes, P., & Kwok, O.-m. (2008). Accelerating early academic oral English development in transitional bilingual and structured English immersion programs. *American Educational Research Journal*, *45*, 1011-1044.

Ullman, J. B. (2007). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell, *Using multivariate statistics* (5th ed., pp. 676-780). Needham Heights, MA: Allyn & Bacon.

Wang, M.-T., & Holcombe, R. (2010). Adolescents' perceptions of school environment, engagement, and academic achievement in middle school. *American Educational Research Journal*, *47*, 633-662.

Widaman, K. F. (2010). Multitrait-multimethod analysis. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 299-314). New York: Routledge.

Williams, L., Edwards, J., & Vandenberg, R. (2003). Recent advances in causal modeling methods for organizational and management research. *Journal of Management*, *29*, 903-936.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*, 806-838.

Yeo, S., Fearrington, J., & Christ, T. J. (2011). An investigation of gender, income, and special education status bias on curriculum-based measurement slope in reading. *School Psychology Quarterly*, *26*, 119-130.

Yo In'nami
Shibaura Institute of Technology
Japan

Rie Koizumi
Juntendo University
Japan