Structural equation modeling in language testing and learning research: A review

Yo In’nami

Toyohashi University of Technology, Japan

Rie Koizumi

Tokiwa University, Japan

This is an electronic version of an article published in the Language Assessment Quarterly, Volume 8, Issue 3, 2011 © Taylor & Francis. Language Assessment Quarterly is available online at: www.tandfonline.com with the open access of our article: http://www.tandfonline.com/doi/abs/10.1080/15434303.2011.582203

Abstract

Despite the recent increase of structural equation modeling (SEM) in language testing and learning research and Kunnan’s (1998) call for the proper use of SEM to produce useful findings, there seem to be no reviews about how SEM is applied in these areas or about the extent to which the current application accords with appropriate practices. To narrow these gaps, we investigated the characteristics of the use of SEM in language testing and learning research. Electronic and manual searches of 20 journals revealed 50 articles containing a total of 360 models analyzed using SEM. We discovered that SEM was most often used to investigate learners’ strategy use and trait/test structure. Maximum likelihood methods were most often used to estimate parameters of a model; model fit indices of chi-squares, CFI, RMSEA, and TLI were often reported, but SRMR rarely was. Univariate and multivariate normality checks were infrequently reported, as was missing data treatment. Sample sizes, when judged according to Kline’s (2005) and Raykov and Marcoulides’ (2006) guidelines, were in most cases adequate, and LISREL was the most widely used program. Two recommendations are provided for the better practice of using and reporting SEM for language testing and learning research.

1 Introduction

Structural equation modeling (SEM), also known as covariance structure analysis or simultaneous equation modeling, is a statistical technique for examining the nature of the
relationships among observed and latent variables that applies a confirmatory, hypothesis-testing approach to the data (e.g., Bollen, 1989b; Byrne, 2006; Ullman, 2007). SEM has been widely used in many areas, such as marketing (e.g., Hulland, Chow, & Lam, 1996; Williams, Edwards, & Vandenberg, 2003) and psychology (e.g., Breckler, 1990; MacCallum & Austin, 2000). In language testing and learning research, SEM was introduced (Kunnan, 1998) and has been reviewed along with other statistical methods (e.g., Bachman, 1998; Bachman & Eignor, 1997; Dörnyei, 2007; Lumley & Brown, 2005). It has been used primarily to examine (a) the factor structure of the abilities assessed by tests or learner characteristics collected via questionnaires (e.g., Gorsuch, 2000; Lee, 2005; Purpura, 1997; Sasaki, 1996; Schoonen et al., 2003), (b) the effect of different constructs and test methods on test performance (e.g., Bachman & Palmer, 1981; Llosa, 2007; Sawaki, 2007), (c) the equivalency of models across different populations (e.g., Bae & Bachman, 1998; Ginther & Stevens, 1998; Kunnan, 1995; Shin, 2005), and (d) the effects of learning context on language development across time (Matsumura, 2003; Ross, 2005).

Reflecting the confirmatory, hypothesis-testing nature of SEM, Fornell (1982) stated that SEM be best used if researchers base their model on what previous research has found, and this process results in the effective accumulation and advancement of knowledge. Such accumulation and advancement of knowledge presupposes the proper use of SEM for the models tested. Further, Kunnan (1998) argued that methods and techniques used in any
academic field must be reviewed to determine if they are skillfully used in order to clarify problems in the field. Since there seem to be no reviews in internationally accessible journals about how SEM is used in the fields of language testing and learning, this paper attempts to fill this gap by reviewing articles using SEM in those fields.

2 Literature Review

2.1 SEM

SEM is a statistical method to examine various relationships that are hypothesized among sets of variables (e.g., Bollen, 1989b; Byrne, 2006; Ullman, 2007). In each analysis, a researcher hypothesizes, based on previous findings, a model showing relationships among variables, and tests whether the model is supported by sample data. According to Byrne (2006), SEM is widely used across various fields for four reasons. First, SEM takes a confirmatory, hypothesis-testing approach to the data, in contrast to traditional analysis, such as exploratory factor analysis, where analysis is data-driven. Second, SEM is designed to correct for measurement errors of variables. The results would allow a researcher to interpret the relationship among variables, separating the measurement errors. Third, SEM can analyze both unobserved (i.e., latent) and observed variables. This contrasts with path analysis that enables researchers to model only observed variables. Latent variables are used to define factors or constructs. Fourth, multivariate relations or indirect effects can be analyzed using
SEM, whereas no other statistical methods can easily do this. Investigation into multivariate relations may include models where correlations are hypothesized only among a certain set of variables. Investigating indirect effects may include determining whether an independent variable directly affects a dependent variable or whether it does so through a mediating variable. Path analysis can be used to model these multivariate relations or indirect effects with observed variables, but it cannot be used to conduct analyses using unobserved variables.

2.2 Five Steps in an SEM Application

Bollen and Long (1993) described five steps involved in an SEM application: (a) model specification, (b) model identification, (c) parameter estimation, (d) model fit, and (e) model respecification. First, the model specification step requires a researcher to develop a model based on previous research in the area. This reflects the confirmatory, theory-driven nature of SEM. Second, the identification step concerns whether the model can obtain a unique value for each parameter in the model whose value is unknown using the variance/covariance matrix (or the correlation matrix and standard deviations) of the measured variables. If there are too many parameters to be estimated relative to the number of variances and covariances in the matrix, the model is underidentified and cannot be tested. Third, the parameter estimation step involves obtaining an estimate for each parameter. Several estimation methods are available, including the maximum likelihood method and the generalized least
squares method, both for multivariate normal data, and the robust maximum likelihood method for nonnormal data. The fourth step tests the degree to which the model fits the data.

Fifth, the model respecification step explores improving the model-data fit, for example, by deleting statistically nonsignificant paths or adding possible paths. Such model respecification must be theoretically supported in the same way as model specification.

Since the fourth step of testing model fit with the data is conducted by inspecting various types of fit indices and since it has been extensively discussed in the SEM literature, further explanation is necessary. A good fit can be indicated by a nonsignificant chi-square ($\chi^2$) value, but this index is known to be affected by sample size. To overcome this problem, many fit indices have been created, which can be classified into four types based on Byrne (2006) and Kline (2005). The first type, incremental or comparative fit indices, assesses the relative improvement in fit of the researcher’s model compared with the null model, which assumes no covariances among the observed variables. Examples are the comparative fit index (CFI), the normed fit index (NFI), and the Tucker-Lewis index (TLI; also known as the nonnormed fit index [NNFI]). The second type, absolute fit indices, calculates a proportion of variance explained by the model in the sample variance/covariance matrix. Examples include the goodness-of-fit index (GFI) and the adjusted goodness-of-fit index (AGFI). The third type, residual-based fit indices, assesses differences between observed and predicted variances and covariances. Examples are the standardized root mean square residual (SRMR) and the root
mean square error of approximation (RMSEA). The fourth type, predictive fit indices, assesses the likelihood of the model to fit in similar-size 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). For the first three types of fit indices, models consistently producing good results across many of these indices are considered good models. In contrast, the last type, predictive fit indices, is used to select the best model among competing models (i.e., among similar models using the same data). A common procedure is that if two models are nested, that is, if a path is deleted or added from Model A to form Model B, a chi-square difference test is used to compare the original and modified models; if two models are not nested, the predictive fit indices, including AIC, CAIC, and ECVI, are employed.

Although no agreed-upon guidelines exist regarding which fit indices should be reported, reviewing the SEM literature provides some clues as to the appropriate reporting practices of fit indices. Table 1 shows a list of works in English providing guidelines for fit indices that should be reported. Works reviewing the use of SEM but not presenting such guidelines explicitly are not listed. We found that the most often recommended indices are the chi-square, CFI, TLI, RMSEA (and its confidence interval), and SRMR. Note that it is assumed that chi-squares are reported along with degrees of freedom and $p$ values; these three statistics are a group of indices used together to evaluate model fit. Degrees of freedom and $p$
values are employed to examine whether the obtained chi-square values are statistically significant, and both of them cannot stand alone. Thus, although some authors (e.g., Widaman, 2010) encourage reporting chi-squares and do not mention reporting corresponding degrees of freedom and $p$ values, the need to report the latter two statistics together with chi-squares is implied.

We found that Hu and Bentler (1998, 1999) were cited in 10 of the works listed in Table 1. Hu and Bentler (1998, 1999) recommended reporting SRMR, along with CFI, TLI, RMSEA, or other indices (i.e., Gamma Hat, Incremental Fit Index [IFI], McDonald’s Centrality Index [MC], Relative Noncentrality Index [RNI]). There are at least four reasons for valuing SRMR reporting. First, Bentler (2006) argued that SRMR’s intuitive interpretability is especially useful to readers who are familiar with correlations but not so much with fit indices. Hu and Bentler (1995) stated the following:

if the discrepancy between the observed correlations and the model-reproduced correlations are [sic] very small, clearly the model is good at accounting for the correlations no matter what the $\chi^2$ or fit indexes seem to imply. For example, if the average of the absolute values of the discrepancy between observed and reproduced correlations is .02, it is simply a fact that the model explains the correlations to [sic] within an average error of .02. . . . We suggest that this descriptive information always
accompany reports of model fit, to round out the more popularly used $\chi^2$ and fit indexes. (p. 98)

Second, SRMR concurs well with the objective of SEM, which is to reproduce, as closely as possible, observed correlation/covariance matrices using model-estimated correlation/covariance matrices (e.g., Bollen, 1989b). Third, unlike other fit indices where chi-squares are used as part of the calculation (e.g., CFI, GFI, RMSEA), SRMR is not based on chi-squares. Rather, it is the average difference between the correlations observed in the input matrix and the correlations predicted by the model. Thus, SRMR can provide a unique perspective for the model fit. Fourth, Bentler (2006) argued that SRMR is also sensitive to model misspecification, although Fan and Sivo (2005) and Yuan (2005) disagreed with this. Considering the advantages of SRMR, reporting SRMR together with other statistics seems to be appropriate practice.

In addition to Hu and Bentler (1998, 1999), several guidelines for reporting combinations of fit indices have been proposed. Hoyle and Panter (1995) recommended reporting chi-squares and Satorra-Bentler corrected chi-squares (and degrees of freedom and
associated with them), CFI, TLI, and GFI when using maximum likelihood estimation and IFI instead of TLI when using generalized least squares estimation. Russell (2002) recommended reporting SRMR and one of the following indices: CFI, IFI, MC, RNI, TLI, or RMSEA. Kashy, Donnellan, Ackerman, and Russell (2009) recommended reporting CFI or TLI along with the chi-square and RMSEA. Bandalos and Finney (2010) recommended the chi-square, CFI, TLI, RMSEA, and SRMR, whereas Mueller and Hancock (2010) recommended RMSEA and its confidence interval, SRMR, and at least one of CFI, NFI, and TLI. Widaman (2010) encouraged 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) recommended reporting CFI, Gamma Hat, and McDonald’s Noncentrality Index and interpreting reduction in each index as evidence for measurement invariance. Following Hu and Bentler (1998, 1999) and the overall recommendations of other researchers, we argue that SRMR should be reported, along with the chi-square, CFI, TLI, or RMSEA (with its confidence interval), unless one has other explicit rationales for fit index selection, and that reporting two or more fit indices is preferable.

2.3 Measurement Issues in SEM

According to Byrne (2006) and Kline (2005), the proper use of SEM must satisfy at
least three statistical requirements. First, most estimation methods used in SEM require univariate and multivariate normality of data. Nonnormality can affect standard error estimates, chi-squares, and consequently model fit indices based on chi-squares. It can be checked by calculating univariate skewness and kurtosis and Mardia’s multivariate kurtosis. Nonnormal data can be transformed and analyzed if normal distribution is expected in the population. If nonnormality is expected, it can be handled by using appropriate estimation methods (e.g., the robust maximum likelihood, which produces the Satorra-Bentler corrected chi-square statistic, or the robust weighted least squares).

Second, missing data can be handled by listwise deletion (removal of all cases with missing values from subsequent analysis) or pairwise deletion (removal of paired cases, either or both of which has missing values). However, a more sophisticated approach would be full information maximum likelihood estimation, where all available data are analyzed regardless of incompleteness (e.g., Enders, 2001). Missing data can also be imputed by, for example, mean substitution, regression imputation, or an expectation maximization algorithm. It should be noted that listwise and pairwise methods work only when data are missing completely at random, although this assumption is often violated in practice (e.g., Muthén, Kaplan, & Hollis, 1987). Both full information maximum likelihood estimation and expectation maximization algorithm methods are effective for situations where the data are missing completely at random or missing at random (for more on missing data analysis, see
Third, since parameter estimates and most fit indices are calculated using sample size, an adequate sample size is necessary. No definitive rules of a necessary sample size have been established because sample size and many variables intricately affect the fit (MacCallum & Austin, 2000). However, Kline (2005) heuristically stated that a sample size below 100 is often considered small, a sample size between 100 and 200 is medium, and a sample size exceeding 200 is large. Similarly, Ding, Velicer, and Harlow (1995) stated that most previous studies agreed that 100 to 150 participants is the minimum sample size adequate for analysis. Further, Raykov and Marcoulides (2006) and many others considered model complexity, which is reflected in the number of free parameters needed to be estimated in a model. They argued that a minimum desirable sample size would be 10 times the number of free model parameters. For example, a model with 30 free parameters would require 300 observations or participants (10 times 30).

Nevertheless, Brown (2006) stressed that these guidelines are crude and do not necessarily apply to the researcher’s data and model as requisite sample size varies with many variables including the amount and patterns of missing data, strength of the relationships among the indicators, types of indicators (e.g., categorical or continuous) and estimators (e.g., [robust] maximum likelihood, robust weighted least squares), and reliability of the indicators. Among more accurate methods for determining appropriate sample sizes,
perhaps the most practical and accessible one for researchers of language testing and learning is Monte Carlo analysis where the appropriateness of sample size is estimated under various conditions as described above by taking into account the statistical power and precision of model parameter estimates (for details, see Muthén & Muthén, 2002).

In addition to the three requirements for SEM, there are two other issues that researchers need to keep in mind when using SEM. First, it is recommended that the variance/covariance matrix or the correlation matrix with standard deviations be provided in the article or be made available upon request. Access to such information allows researchers to reproduce the original model and examine alternative models not tested in the primary study (see In’nami & Koizumi, 2010, for further details). Second, the name and version of software used should be reported since calculation formulas and default methods sometimes differ across software programs (e.g., calculation formula of AIC between Amos [Arbuckle, 1994–2011] and EQS [Bentler, 1994–2011], as explained in Brown, 2006) and even within different versions of the same software programs.

2.4 Current Study

Despite the increasing use of SEM in language testing and learning research and Kunnan’s (1998) call for the proper use of SEM to produce useful findings, there seem to be no reviews in internationally accessible journals about how SEM is applied in these areas or
about the extent to which the current application accords with appropriate practices. To narrow these gaps, we addressed the following research question: What are the characteristics of use of structural equation modeling in language testing and learning research?

We investigated this in terms of (a) content analysis, (b) parameter estimation methods, (c) the types and frequency of model fit indices used, (d) normality checks, (e) missing data treatment, (f) comparison of actual and adequate sample sizes, and (g) software used. We conducted content analysis to examine the topics in language testing and learning that have been researched using the SEM approach. Further, we focused on two of the five aforementioned steps in an SEM application—parameter estimation and model fit indices. This is because we did not have the range of content expertise to evaluate the theoretical justification of model specification and model respecification, and because all the models reported in articles are usually statistically identified (i.e., it is possible to compute a unique estimate of each parameter) so that it made little sense to incorporate this step in the review. By conducting this review, we hope to raise awareness about the proper use and reporting of SEM among researchers in the fields of language testing and learning.

3 Method

3.1 Article Collection

We searched for empirical studies using SEM in language testing and learning in
October 2008 in two ways. First, we retrieved studies from the following 20 representative journals using Education Resources Information Center (ERIC), Linguistics and Language Behavior Abstracts (LLBA), and search engines available at journal homepages: Annual Review of Applied Linguistics, Applied Language Learning, Applied Linguistics (AL), Assessing Writing, ELT Journal, Foreign Language Annals, International Journal of Applied Linguistics, International Review of Applied Linguistics in Language Teaching, Language Assessment Quarterly (LAQ), Language Learning (LL), Language Learning & Technology, Language Teaching, Language Teaching Research, Language Testing (LT), Modern Language Journal (MLJ), RELC Journal (RELC), Second Language Research, Studies in Second Language Acquisition (SSLA), System, and TESOL Quarterly (TQ). The following keywords were independently used: causal analysis (analyses), causal model(s), causal modeling (modelling), confirmatory factor analysis (analyses), covariance structure(s), covariance structure analysis (analyses), covariance structure model(s), simultaneous equation model(s), simultaneous equation modeling (modelling), structural equation model(s), and structural equation modeling (modelling). We compiled these sets of keywords based on the keywords and synonyms found in books, articles, our experiences, and feedback from colleagues. Abstract, title, and article keyword searches were used. A date range restriction was not imposed. Second, we conducted manual searches for relevant studies in these journals to crosscheck studies identified electronically. The reference list of empirical,
theoretical, and review papers was further inspected for additional relevant materials.

The literature search was limited to published journal articles, as we intended to examine the status quo based on internationally accessible published sources, although we were aware of many unpublished SEM studies such as Liao (2009) and Yamashiro (2002).

### 3.2 Analyses

The article was used as a unit of analysis. When multiple models existed in one article, information on the final model was coded. When several final models existed (i.e., multiple research questions were examined in one paper, and each research question was accompanied by a final model, resulting in multiple final models in one article), one of the final models was randomly selected. However, results based on the model as a unit of analysis (i.e., counting all models included in a paper) were also reported for reference.

In order to calculate the minimum desirable sample size for SEM analyses, the number of free parameters in each model was calculated and multiplied by ten, following Raykov and Marcoulides (2006). The number of free parameters in a model was calculated by drawing a model with Amos 4.01 (Arbuckle, 1999) and entering (a) the corresponding correlation (with SDs) or variance/covariance matrix, or (b) the corresponding size (row x column) of a hypothetical matrix for a model. Conducting Monte Carlo studies to evaluate the appropriateness of sample size for each model was beyond the scope of this paper, although we plan to undertake these in the near future.
In order to check the consistency of coding many kinds of information on articles and models, a sample of 8% (30 models) of the 360 models (see section 4.1) was independently examined by both authors. The agreement percentage and kappa coefficient for all the coding were 95 and .90, respectively. When the two authors disagreed, they discussed their decisions and rationale for giving their ratings while reconsidering the models for which coding discrepancy occurred. Most of the coded information—especially, parameter estimation and model fit indices—appeared in the method, result, or discussion sections of a paper and was explicitly reported. The remaining models were examined by the first author.

4 Results and Discussion

4.1 SEM in Language Testing and Learning Studies

Nine of the 20 journals reviewed included articles using SEM. As Table 2 indicates, most articles using SEM appeared in Language Testing (42% of the articles; 21/50), followed by Language Learning (20%; 10/50). This suggests that SEM has been used particularly in studies on language testing compared to studies on language education in general. Table 3 shows that the first article using SEM appeared in 1981. Since then, an increasing number of articles have used SEM. The total number of models from the 50 articles in the nine journals was 360. Those 50 articles are marked with asterisks (*) in the references.
4.2 Research Question: What are the Characteristics of Use of Structural Equation Modeling in Language Testing and Learning Research?

4.2.1 Content Analysis

Table 4 shows the frequency of articles by content area. These data were broadly divided into test-taker variables and test-/task-related variables and reflected the diverse areas of language testing and learning that have been subjected to SEM. The most researched area among test-taker variables was strategy (26%; 13/50), followed by motivation (16%) and social milieu (12%), and among test-/task-related variables, it was trait/test structure (20%). It should be noted that these three variables—motivation, social milieu, and trait/test structure—are those stated in Kunnan (1998) as promising lines of research and that an increasing number of research works regarding these variables indicate advances in our field. In fact, more than half of the studies investigating these variables were published in 1999 or later. Table 4 also displays areas where SEM has been used but where the number of such studies is limited as represented by explicit and implicit knowledge and self-rating. It should be noted that Table 4 was developed inductively and does not list all topics of substantive interest that can benefit from application of SEM. Examples of such topics of interest may include investigation into developmental patterns of L2 acquisition, L1 transfer effects on L2 learning, the relationship between personality and test scores, and the comparison between
instructed and natural settings for L2 learning. Finally, although not shown in Table 4, nine of the 10 articles investigating trait/test structure were published in *Language Testing*, suggesting that SEM is an essential methodology for examining trait/test structure among language testers.

[Insert Table 4 about here]

To explore whether researchers who examined certain content areas selected different parameter estimation methods, types and frequencies of model fit indices, normality checks, missing data treatments, ways of comparing actual and adequate sample sizes, and software, we selected the top four topics in Table 4 (i.e., strategy, motivation, social milieu, and trait/test structure) and presented the results separately below.

### 4.2.2 Parameter Estimation

As Table 5 shows, estimation was in most cases conducted using the maximum likelihood method (52%; 26/50). This was not surprising since the maximum likelihood method is the default in many SEM programs. The second highest percentage was explained by articles not reporting estimation methods (30%), which is problematic since the choice of estimation method would affect the validity of the results. A similar trend was also seen
One question of significant interest concerns the extent to which parameter estimation was appropriate in the previous applications while considering various measurement issues, including the type of data being analyzed (e.g., categorical vs. continuous) and multivariate normality of data, which we summarized in Table 6. From all the articles and books we reviewed, we found Finney and DiStefano (2006) the most accessible, synthetic, authoritative, and up-to-date. They stated that if the variables have at least five ordered categories and the data are simultaneously normally distributed, the data can be treated as continuous in nature and the maximum likelihood estimation is recommended; if the variables have at least five ordered categories and the data are nonnormally distributed, the robust maximum likelihood estimation is recommended; if the variables have four or less categories, the robust weighted least squares estimator is recommended. The numerals representing the number of articles and models following these recommendations are italicized in Table 6 and this number was found to be very small: Maximum likelihood (12%; 6/50) and robust maximum likelihood (2%). This showed that a majority of SEM applications were not appropriate with regard to the parameter estimation method. The nil use of robust weighted least squares estimation was not desirable but understandable because it is available only in Mplus (Muthén & Muthén, 1998–2011) which, as discussed below, has been rarely used in our field (see Table 8).
4.2.3 Model Fit Indices

4.2.3.1 Types of Fit Indices Used

Table 5 shows that chi-squares (84%; 42/50), $p$ values (70%), and degrees of freedom (84%) were, overall, often reported. Other widely reported fit indices were CFI (64%), RMSEA (50%), and TLI (44%). These practices seem appropriate because the SEM literature has frequently recommended the use of these indices. Confidence intervals of RMSEA were rarely reported (4%), despite long-standing SEM literature advice to do so (e.g., MacCallum & Austin, 2000; Steiger, 1990). A similar pattern of fit indices reporting was also observed across the top four topics.

These results are consistent with Jackson, Gillaspy, and Purc-Stephenson (2009), who stated that in psychology, chi-squares (89%), CFI (78%), RMSEA (65%), and TLI (46%) were the most commonly reported measures of fit. This indicates that these indices tended to be used often across fields. One exception was SRMR, which was rarely reported (2%) in the language testing and learning literature, but, as reported by Jackson et al., was more often reported in psychology (23%). Thus, Hu and Bentler’s (1998, 1999) recommendation to report SRMR is not well practiced in our field.

To examine the reason for the infrequent use of SRMR in the language testing and
learning literature, the first appearance of SRMR in each SEM computer program was investigated. To the best of our knowledge, SRMR is calculated in Amos (1995, version 3.1 or later),\(^1\) CALIS (SAS PROC CALIS; at least in the 2009 version or later), EQS (1995, version 4.0 or later), LISREL (1986, version VI or later), and Mplus (2001, version 2 or later).

From the 50 articles, those in which the models were analyzed using these or later versions of the above-mentioned programs were extracted. Articles not reporting software versions were excluded. We found 11 articles using Amos, 0 using CALIS (SAS Institute, 1990–2011), 9 using EQS, 7 using LISREL (Jöreskog & Sörbom, 1974–2011), and 1 using Mplus, totaling 28 articles. Given that such a large number of articles could have been reported with SRMR but that only one article reported this index, the reason for the infrequent use of SRMR was at least not a technical one.\(^2\) Further, it seems that most language testers and researchers who used SEM were not very familiar with Hu and Bentler’s (1998, 1999) recommendation to report SRMR because only six of the 37 articles published in 1998 or later cited Hu and Bentler (1999) and none cited their 1998 paper. Moreover, the six papers that cited Hu and Bentler (1999) referred to the cutoff criteria for fit indices (e.g., RMSEA values of approximately .06 or below indicate a good fit) but did not refer to SRMR.

---

1. It should be noted that calculation of SRMR in Amos requires the use of its plug-in function (see Garson, 2009).
2. In EQS, SRMR is reported under normality assumption but not as part of robust statistics for maximum likelihood (ML), least squares (LS), or generalized least squares (GLS) estimation method. Even when using these robust statistics, SRMR that is reported under normality assumption needs to be reported. This is because it is not necessary to adjust SRMR for nonnormality, as it is simply the standardized difference between the observed and model-estimated covariances (EQS Technical Support Team [Kevin H. Kim], personal communication, September 9, 2010).
We also found one instance of misuse of model fit indices. In two articles, different criteria were used to evaluate different models, even though the same criteria should be used to compare all models in a study. In one article containing five models, the author used $p$ values of chi-squares and NFI to interpret one model and additionally used TLI and CFI to interpret the remaining four models. In another article containing 27 models, the authors used $p$ values of chi-squares, TLI, CFI, GFI, RMSEA, and SRMR to interpret the first two of their models, but did not use GFI and SRMR to interpret the remaining 25 models. These two articles seem to suggest that the authors had implicit rationales for using certain fit indices to interpret certain kinds of models, although such rationales were not made explicit in their articles.

**4.2.3.2 Frequency of Fit Indices Used**

As seen in Table 7, 50 articles using SEM reported a various number of fit indices from zero to 12. On average, 5.16 fit indices per study were reported overall; a smaller number of fit indices per study (3.90) were reported for studies on trait/test structure. These results are appropriate since it is recommended that two or more indices be reported. Exceptions include four articles reporting either zero or one fit index (i.e., zero in two articles, the chi-square statistic in one article, and CFI in one article).
We further examined whether the use of fit indices was appropriate in terms of their combinations because fit indices properly used in tandem can provide complementary information to one another. Table 5 shows that only one of the 50 articles reported SRMR and one of the following, CFI, TLI, or RMSEA, as recommended by Hu and Bentler (1998, 1999). Likewise, only three of the 50 articles reported all of the chi-square (and the Satorra-Bentler corrected chi-square), degree of freedom, \( p \) value, CFI, TLI, and GFI, as suggested by Hoyle and Panter (1995, for maximum likelihood estimation). Further, the reporting practice did not follow Bandalos and Finney (2010), Mueller and Hancock (2010), and Cheung and Rensvold (2002, when testing measurement invariance). In contrast, approximately a quarter of the 50 articles, although still low in percentage, were satisfactory with regard to Kashy et al.’s (2009) and Widaman’s (2010) guidelines. The results were similar when the top four content areas were analyzed individually. In sum, these researchers’ recommendations were not well reflected in actual practice, and this needs to be rectified.

4.2.4 Test of Data Normality

As shown in Table 8, overall, less than half of the articles reported univariate and multivariate normality tests (44% [22/50] and 32% [16/50] respectively). The results were
essentially the same regardless of the content areas. Although not shown in Table 8, we found that none of the six articles reporting 50 models in *Modern Language Journal* reported test of multivariate normality of data; nor did one article reporting two models in *Studies in Second Language Acquisition* and four articles reporting 12 models in *TESOL Quarterly*. Using nonnormal data would result in bias in calculating statistics, including fit indices. Since univariate and multivariate normality can be easily checked using statistics available in SEM programs, this should be routinely inspected and reported.

[Insert Table 8 about here]

### 4.2.5 Missing Data Treatment

The discussion on missing data treatment was often not reported (e.g., 66% among all studies; see Table 8), and it is not clear whether the data contained no missing responses, whether the missing data were deleted or imputed before analyzed, or whether the missing data were analyzed using the full information maximum likelihood method. Missing data were mostly handled using listwise deletion (20%). Listwise methods may not be desirable, however, since they only work on data missing completely at random although this assumption is often violated in practice (e.g., Muthén et al., 1987). When imputation was used, which was rare (6% in total), details were not clear. In case of two articles containing five models using EQS, it was not clear which of the implementation methods was used,
while EQS has five options (i.e., mean imputation, regression imputation, stochastic regression imputation, hot-deck imputation, and expectation maximization imputation; Bentler, 2006, pp. 278–279). Only one article (2%) reported to have no missing data. Researchers are encouraged to report how missing data, if any, were addressed. These frequencies and patterns of missing data treatment were also observed across the four topics.

### 4.2.6 Software

As seen in Table 8, overall, the most often used software for SEM analysis was LISREL (40%), followed by EQS (30%) and Amos (24%). This order of frequency seems natural since LISREL is the oldest of the three and Amos is the newest. Another reason would be that authors using LISREL tend to be especially prolific. Among the 20 articles using LISREL, 8 were by the same authors (4 articles by 1 group of author[s], and the other 4 by another). For studies investigating motivation and social milieu, Amos was most frequently used (70% and 50% respectively).

### 4.2.7 Sample Size

Table 9 shows a comparison of the actual median sample sizes of the 50 articles, using Kline’s (2005) guidelines. Six articles (12%) were found to be based on small sample sizes.
Table 10 shows the descriptive statistics of actual and minimum desired sample sizes calculated based on Raykov and Marcoulides (2006). We interpreted median rather than mean, standard deviation, and mode, because mean and standard deviation were misleading given the large skewed and kurtotic distribution of actual sample sizes, and because the mode was based only on two articles with the same sample size. The median sample size of the 50 articles was 258.5, a large sample size according to Kline (2005).

Table 10 also shows that the number of free parameters that needed to be estimated in each article had the median value of 35.5. Thus, the minimum desirable sample size based on the number of free parameters was 355 (35.5 x 10). A comparison of the actual (258.5) and minimum desirable sample size (355) across the 50 articles shows that, on average, the sample size used for structural equation models in the field of language testing and learning was insufficient. This result held across the top four content areas. For example, the actual sample size for strategy research was 281, although the desirable sample size was at least 580. Further, more details were obtained by examining the differences between the actual and minimum desirable sample sizes for each article, as seen in Figure 1. The differences ranged
from –690 to 550, with extreme values excluded. The differences peaked around zero, with more articles located in the minus area, suggesting that the actual sample sizes for each article were often smaller than the minimum desired sample sizes; in fact, 29 (58%) of the 50 articles had an insufficient sample size.

After having examined the adequacy of sample sizes using the two above-mentioned criteria, we might be able to argue that the sample size of a study is insufficient if the sample size was judged to be insufficient by both the two above-mentioned criteria (i.e., a sample size of 99 or below according to Kline, 2005, and the actual sample size smaller than the minimum desirable sample size according to Raykov & Marcoulides, 2006). We found that there were six such articles (12%); thus, we conclude that six of the 50 articles had insufficient sample sizes.

Although researchers rarely justify the sample size of their study, we found two exceptions. In one article containing two models, the authors stated that their sample size was adequate since it was more than 200. In the other article containing one model, the authors defended their sample size by stating it was larger than 10 times the number of free model parameters. Justification for sample sizes should be encouraged more.
4.2.8 Others

Although not explicit from Tables 5 to 10, we found that the same author(s) tended to use the same analytical procedure for SEM. For example, if one does not check multivariate normality of the data in one’s early work, one does not do in one’s later work either. This suggests that authors, if guided appropriately in the early stages of their academic careers, would continue to follow good practice. Thus, for example, when an author submits an article for review for the first time, the editor(s) or reviewers can encourage the author to report on the normality of data distribution. That author is then more likely to report such information when writing a later article.

5 Summary and Implications

The use of SEM in language testing and learning research has recently increased. Thus, as Kunnan (1998) argued, its use needs to be periodically reviewed so that a new model can firmly build on existing models that were tested using a proper procedure. We therefore examined the characteristics of use of SEM in language testing and learning research. In the 20 internationally accessible published journals on language testing and learning we reviewed, 360 models from 50 articles were found in nine journals and inspected in terms of (a) content analysis, (b) parameter estimation methods, (c) the types and frequency of model fit indices used, (d) normality checks, (e) missing data treatment, (f) sample sizes, and (g) software used.
The content analysis showed that the most researched area using SEM was strategy, which was followed by trait/test structure. Regarding parameter estimation, maximum likelihood methods were most often used. With respect to model fit, chi-squares (and $p$ values and degrees of freedom associated with them), CFI, RMSEA, and TLI were often reported. SRMR was rarely used. Although multiple fit indices were usually reported, few studies reported overall exemplary combinations of fit indices.

Additionally, univariate and multivariate normality checks were infrequently reported. Missing data treatment was also infrequently reported. When it was reported, listwise deletion and full information maximum likelihood estimation were often used. According to Kline’s (2005) guidelines, a median sample size of 258.5, as observed in actual studies, was adequate for SEM analysis; however, the difference of actual (258.5) and minimum desired sample sizes (355) based on the number of parameters (e.g., Raykov & Marcoulides, 2006) suggested the need to have larger sample sizes. Combining these two criteria, six of the 50 articles were found to have insufficient sample sizes. Nevertheless, given that models with varying characteristics (e.g., the amount and patterns of missing data and the strength of the relationships among the indicators) have been used in the language testing and learning literature, and given that the sample size is influenced by these characteristics, our findings based on these two guidelines need to be interpreted carefully. Further, LISREL was more widely used as compared to EQS and Amos. These results were generally the same across the
50 studies and across the top four content areas of strategy, motivation, social milieu, and trait/test structure.

Two recommendations for better practice in terms of using and reporting SEM for language testing and learning research are discussed. First, all topics in the language testing and learning fields, parts of which were investigated in previous studies and are shown in Table 4, would merit further investigation using SEM. We particularly suggest three topics that may be well addressed applying SEM. To begin, we suggest conducting more research into the stability of a model across samples to examine whether the same constructs are measured across samples. This issue is especially relevant to studies involving learners of varying characteristics, especially for large-scale, high-stakes tests intended for heterogeneous test-takers. Data can be analyzed using multi-sample analysis, and results can provide insight into the fairness of test score interpretation across learners (see Purpura, 1998; Shin, 2005). Moreover, we suggest conducting studies on longitudinal change in language proficiency. Although they are as important as cross-sectional studies, it is only recently that longitudinal studies have received growing attention as a means for advancing our knowledge of language acquisition processes (Ortega & Byrnes, 2008). Insights that are gained from longitudinal research may be useful in developing proficiency descriptors such as Can-do statements and forming sound educational practices. For this purpose, latent growth modeling allows for the evaluation of change over time on group as well as individual levels (see
Matsumura, 2003; Ross, 2005). Finally, we suggest more investigations into sources of variability in performance assessment. The data in performance assessment often have a hierarchical structure where observations are nested within higher levels of classification (e.g., ratings are nested within raters). There has been little empirical research that considers such nested structure, despite the sustained interest in how rater attributes are related to systematic variance in ratings of speaking or writing performance. In this case, multilevel modeling is useful in examining variables of interest at each hierarchical level (see Barkaoui, 2010).

Second, we suggest that authors use appropriate procedures and provide sufficient information about and justification for five perspectives: (a) parameter estimation methods, (b) model fit indices used, (c) normality checks, (d) missing data treatment, and (e) sample sizes. The first perspective is regarding the type of estimation method; our study showed that 30% of the articles were unclear regarding the estimation method used. This needs to be rectified, since choice of estimation affects subsequent analyses (for details, see Ullman, 2007) and is easy to report. We also strongly encourage researchers using SEM to pay closer attention to Finney and DiStefano (2006) for nonnormal or categorical data issues. Although these authors recommend the robust weighted least squares estimator if the variables have four or less categories, this estimator is available only in Mplus. One needs to be familiar with Mplus or needs to reconsider increasing the number of ordered scale categories to five or more.
The second perspective is as regards fit indices; we showed that some authors used different fit indices depending on the models, even when using the same data within a series of analyses. However, such practice cannot be recommended unless a clear explanation is provided as to why some fit indices are more appropriate for evaluating certain kinds of models. Additionally, more attention needs to be paid to SRMR. We suggest routine reporting of SRMR in evaluating model fit. As can be clearly seen from our results, SRMR is rarely used in language testing and learning, in contrast to Hu and Bentler’s (1998, 1999) recommendation to report the index. Since intuitive interpretability of SRMR is appealing to researchers unfamiliar with fit indices and since SRMR can be calculated using SEM programs, this index needs to be reported more often. SRMR appears in the output if one uses Amos, CALIS, EQS, LISREL, or Mplus (see section 4.2.3.1). Finally, we encourage researchers to follow reporting guidelines for combinations of fit indices recommended by researchers such as Hu and Bentler (1998, 1999) and others (see section 2.2). These combinations consist of fit indices of different nature, thereby assessing data-model fit from various aspects.

Third, we suggest that authors examine data normality before the main data analysis is run. Univariate normality can be checked by calculating the skewness and kurtosis of each variable. Multivariate normality, especially multivariate kurtosis, can be checked by calculating Mardia’s estimate using SEM programs. Skewed and/or kurtotic distribution of a
variable can be transformed if normal distribution is expected in the population (see, for example, Kline, 2005). If nonnormal distribution is expected in the population, skewed and/or kurtotic data can be analyzed using estimation methods such as the robust maximum likelihood estimation available in EQS, LISREL, and Mplus.

Regarding the fourth perspective, we suggest that authors check if there are any missing cases in their data. Perhaps the easiest way is to invoke Excel’s “find” function (i.e., the control key + f), enter nothing (just leave the search window blank), and press the “find” button. The activated cell will be moved to any blank cell showing missing responses. Although listwise and pairwise methods were commonly used, a more appropriate way is to use full information maximum likelihood or expectation maximization methods (see, for example, Enders, 2001). These can be carried out in SEM programs.

The fifth perspective is about sample sizes. We suggest that authors refer to guidelines on sample sizes, such as in Kline (2005) and Raykov and Marcoulides (2006), when reporting on sample sizes for their studies. Although there is no universally agreed-upon sample size needed for analysis, mentioning these guidelines would help readers understand authors’ decisions about sample sizes.

Of course, not all language testers and language learning researchers are versed in SEM, nor are they expected to keep up with the latest SEM methodological developments. Further, space limitations may prevent complete reporting of analysis details. However, at least the
latter recommendation, which includes the aforementioned five perspectives, would be essential to take complete advantage of SEM. Further, relatively little space is required for reporting results related to these five perspectives. More details on these recommendations and other relevant issues are found in introductory books such as Byrne (2006) and Kline (2005), which are highly readable. In conclusion, these changes would not happen overnight, but we believe that paying close attention to the five perspectives would promote the proper use and reporting of SEM among researchers in the fields of language testing and learning, and would eventually lead to the accumulation and advancement of knowledge in the field.

Acknowledgement

An earlier version of this paper was presented at the American Association for Applied Linguistics 2011 Conference, Chicago, Illinois, USA. We would like to thank Hiroaki Maeda and the three anonymous reviewers for their valuable comments on earlier versions of this paper.

References

References marked with an asterisk indicate articles included in the review.


*Csizér, K., & Kormos, J. (2009). Modelling the role of inter-cultural contact in the


Lincolnwood, IL: Scientific Software International.


Re-examining the models of L2 reading and listening abilities and their relations to lexico-grammatical knowledge. (UMI No. 3348356)


equation analyses. *Psychological Methods*, 7, 64–82.


*Turner, C. E. (1989). The underlying factor structure of L2 close test performance in


806–838.


Table 1

<table>
<thead>
<tr>
<th>Fit Indices Recommended in the Literature</th>
<th>Basic</th>
<th>Incremental</th>
<th>Absolute</th>
<th>Residual-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>CFI</td>
<td>Gamma hat</td>
<td>IFI</td>
</tr>
<tr>
<td>Simulation articles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hu &amp; Bentler (1998, 1999)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Review articles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoyle &amp; Panter (1995)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Boomsma (2000)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>MacCallum &amp; Austin (2000)</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McDonald &amp; Ho (2002)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Russell (2002)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Martens (2005)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Kahn (2006)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Worthington &amp; Whittaker (2006)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Jackson et al. (2009)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Kashy et al. (2009)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introductory textbooks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kline (2005)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brown (2006)</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ullman (2007)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mueller &amp; Hancock (2008)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Bandalos &amp; Finney (2010)</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lomax (2010)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mueller &amp; Hancock (2010)</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Widaman (2010)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Total number</td>
<td>11</td>
<td>17</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Our recommendation</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

*Note. URFI = Unbiased Relative Fit Index. CI = Confidence interval. Predictive fit indices are not listed because they are different in nature in that they are used to compare models. IFI, also called Bollen’s fit index (BL89) and Bollen’s delta 2, is less variable in small samples and more consistent across estimators than TLI (Bollen, 1989a). MC transforms the population noncentrality index into a range from 0 to 1 (McDonald, 1989). RNI and URFI are equivalent to CFI, except they can be below 0 or over 1, showing the extent of overfit (see Hoyle & Panter, 1995, for RNI; McDonald & Ho, 2002, for URFI).

*aShe argued SRMR may be more preferable than RMSEA and TLI with a small sample size. *bShe also recommended reporting AIC and CAIC when comparing nonnested models.
Table 2

Number of Articles and Models by Journal

<table>
<thead>
<tr>
<th></th>
<th>AL</th>
<th>LAQ</th>
<th>LL</th>
<th>LT</th>
<th>MLJ</th>
<th>RELC</th>
<th>SSLA</th>
<th>System</th>
<th>TQ</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>21</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>Model</td>
<td>19</td>
<td>1</td>
<td>48</td>
<td>224</td>
<td>50</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>12</td>
<td>360</td>
</tr>
</tbody>
</table>

*Note.* The total number also applies to Tables 5, 7, and 8.
Table 3

Number of Articles and Models by Year

<table>
<thead>
<tr>
<th>Year</th>
<th>Article</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>81</td>
<td>82</td>
</tr>
<tr>
<td>81</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>82</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>86</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>00</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>02</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>03</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>04</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>05</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>06</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>08</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>360</td>
</tr>
</tbody>
</table>
Table 4

*Frequency of Articles by Content Area*

<table>
<thead>
<tr>
<th>Article (%)</th>
<th>Article (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test-taker variables</strong></td>
<td><strong>Test-taker variables</strong></td>
</tr>
<tr>
<td>Strategy</td>
<td>13 (26%)</td>
</tr>
<tr>
<td>Motivation</td>
<td>8 (16%)</td>
</tr>
<tr>
<td>Social milieu or cultural background</td>
<td>6 (12%)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>5 (10%)</td>
</tr>
<tr>
<td>Reading</td>
<td>5 (10%)</td>
</tr>
<tr>
<td>Speaking</td>
<td>5 (10%)</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>5 (10%)</td>
</tr>
<tr>
<td>Listening</td>
<td>4 (8%)</td>
</tr>
<tr>
<td>Writing</td>
<td>4 (8%)</td>
</tr>
<tr>
<td>Longitudinal growth</td>
<td>3 (6%)</td>
</tr>
<tr>
<td>Metacognitive knowledge</td>
<td>3 (6%)</td>
</tr>
<tr>
<td>Grammar</td>
<td>2 (4%)</td>
</tr>
<tr>
<td>Willingness to communicate</td>
<td>2 (4%)</td>
</tr>
<tr>
<td>Aptitude</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Explicit &amp; implicit knowledge</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Intelligence</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Pragmatic development</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Teacher perception on communicative activities</td>
<td>1 (2%)</td>
</tr>
<tr>
<td><strong>Test-/task-related variables</strong></td>
<td><strong>Test-/task-related variables</strong></td>
</tr>
<tr>
<td>Trait/test structure</td>
<td>10 (20%)</td>
</tr>
<tr>
<td>Cloze test</td>
<td>2 (4%)</td>
</tr>
<tr>
<td>Differential item (or test) functioning</td>
<td>2 (4%)</td>
</tr>
<tr>
<td>Rater effect</td>
<td>2 (4%)</td>
</tr>
<tr>
<td>C-test</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Classroom assessment &amp; standardized tests</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Overall task characteristics</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Paper-/computer-based test comparison</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Self-rating</td>
<td>1 (2%)</td>
</tr>
</tbody>
</table>

*Note.* Articles investigating more than one variable were classified into more than one category.
### Table 5

**Characteristics of Parameter Estimation Methods and Model Fit Indices Used**

<table>
<thead>
<tr>
<th>Parameter estimation</th>
<th>Article</th>
<th>Strategy</th>
<th>Motivation</th>
<th>Social milieu</th>
<th>Trait structure</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalized least squares</td>
<td>1 (2%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20 (6%)</td>
</tr>
<tr>
<td>Maximum likelihood</td>
<td>26 (52%)</td>
<td>6 (46%)</td>
<td>2 (25%)</td>
<td>5 (83%)</td>
<td>6 (60%)</td>
<td>220 (61%)</td>
</tr>
<tr>
<td>Robust maximum likelihood</td>
<td>4 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3 (30%)</td>
<td>43 (9%)</td>
</tr>
<tr>
<td>Unweighted least squares</td>
<td>1 (2%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4 (1%)</td>
</tr>
<tr>
<td>Weighted least squares (asymptotically distribution free)</td>
<td>1 (2%)</td>
<td>1 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>27 (7%)</td>
</tr>
<tr>
<td>Bootstrap</td>
<td>1 (2%)</td>
<td>1 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (0.3%)</td>
</tr>
<tr>
<td>Not reported</td>
<td>19 (30%)</td>
<td>6 (46%)</td>
<td>4 (50%)</td>
<td>1 (17%)</td>
<td>3 (30%)</td>
<td>45 (13%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model fit</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic fit indices</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chi-square ($\chi^2$)</td>
<td>42 (84%)</td>
<td>12 (92%)</td>
<td>6 (75%)</td>
<td>4 (67%)</td>
<td>8 (80%)</td>
<td>311 (86%)</td>
</tr>
<tr>
<td>$p$ value ($p$)</td>
<td>35 (70%)</td>
<td>9 (69%)</td>
<td>5 (63%)</td>
<td>4 (67%)</td>
<td>7 (70%)</td>
<td>201 (56%)</td>
</tr>
<tr>
<td>Degree of freedom ($df$)</td>
<td>42 (84%)</td>
<td>11 (85%)</td>
<td>5 (63%)</td>
<td>4 (67%)</td>
<td>9 (90%)</td>
<td>303 (84%)</td>
</tr>
<tr>
<td>$\chi^2/df$</td>
<td>15 (30%)</td>
<td>5 (38%)</td>
<td>5 (63%)</td>
<td>3 (30%)</td>
<td>3 (30%)</td>
<td>119 (33%)</td>
</tr>
<tr>
<td>Satorra-Bentler corrected chi-square (SB$\chi^2$)</td>
<td>4 (8%)</td>
<td>0</td>
<td>0</td>
<td>1 (17%)</td>
<td>2 (20%)</td>
<td>28 (8%)</td>
</tr>
<tr>
<td>SB$\chi^2/df$</td>
<td>2 (4%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (10%)</td>
<td>8 (2%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Incremental or comparative fit indices</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CFI (Comparative Fit index)</td>
<td>32 (64%)</td>
<td>10 (77%)</td>
<td>6 (75%)</td>
<td>6 (100%)</td>
<td>5 (50%)</td>
<td>220 (61%)</td>
</tr>
<tr>
<td>IFI (Incremental Fit Index)</td>
<td>5 (10%)</td>
<td>3 (23%)</td>
<td>3 (38%)</td>
<td>1 (17%)</td>
<td>0</td>
<td>11 (3%)</td>
</tr>
<tr>
<td>NFI (Normed Fit Index)</td>
<td>15 (30%)</td>
<td>6 (46%)</td>
<td>4 (50%)</td>
<td>3 (30%)</td>
<td>2 (20%)</td>
<td>65 (18%)</td>
</tr>
<tr>
<td>PCFI ( Parsimony-adjusted CFI)</td>
<td>2 (4%)</td>
<td>0</td>
<td>2 (25%)</td>
<td>2 (34%)</td>
<td>0</td>
<td>26 (7%)</td>
</tr>
<tr>
<td>PNFI (Parsimony-adjusted NFI)</td>
<td>2 (4%)</td>
<td>1 (8%)</td>
<td>1 (13%)</td>
<td>1 (17%)</td>
<td>0</td>
<td>4 (1%)</td>
</tr>
<tr>
<td>RFI (Relative Fit Index)</td>
<td>2 (4%)</td>
<td>0</td>
<td>1 (13%)</td>
<td>1 (17%)</td>
<td>0</td>
<td>7 (2%)</td>
</tr>
<tr>
<td>PRATIO ( Parsimony Ratio)</td>
<td>1 (2%)</td>
<td>0</td>
<td>1 (13%)</td>
<td>1 (17%)</td>
<td>0</td>
<td>1 (0.3%)</td>
</tr>
<tr>
<td>TLI (Tucker-Lewis Index)</td>
<td>22 (44%)</td>
<td>11 (85%)</td>
<td>4 (50%)</td>
<td>2 (34%)</td>
<td>1 (10%)</td>
<td>137 (38%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Absolute fit indices</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GFI (Goodness of Fit Index)</td>
<td>15 (30%)</td>
<td>5 (38%)</td>
<td>5 (63%)</td>
<td>1 (17%)</td>
<td>1 (10%)</td>
<td>70 (19%)</td>
</tr>
<tr>
<td>AGFI (Adjusted GFI)</td>
<td>10 (20%)</td>
<td>5 (38%)</td>
<td>5 (63%)</td>
<td>1 (17%)</td>
<td>0</td>
<td>34 (9%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Residual-based fit indices</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RMR (Root Mean Square Residual)</td>
<td>3 (6%)</td>
<td>1 (8%)</td>
<td>1 (13%)</td>
<td>0</td>
<td>1 (10%)</td>
<td>13 (4%)</td>
</tr>
<tr>
<td>RMSEA (Root Mean Square Error of Approximation)</td>
<td>25 (50%)</td>
<td>8 (62%)</td>
<td>5 (63%)</td>
<td>4 (67%)</td>
<td>4 (40%)</td>
<td>201 (56%)</td>
</tr>
<tr>
<td>RMSEA confidence interval</td>
<td>2 (4%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (10%)</td>
<td>16 (4%)</td>
</tr>
<tr>
<td>SRMR (Standardized RMR)</td>
<td>1 (2%)</td>
<td>1 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3 (1%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predictive fit indices</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC (Akaike Information Criterion)</td>
<td>1 (2%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (10%)</td>
<td>4 (1%)</td>
</tr>
<tr>
<td>CAIC (Consistent AIC)</td>
<td>1 (2%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (10%)</td>
<td>4 (1%)</td>
</tr>
<tr>
<td>Degree of likelihood ratio</td>
<td>1 (2%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20 (6%)</td>
</tr>
<tr>
<td>ECVI ( Expected Cross-Validation Index)</td>
<td>2 (4%)</td>
<td>1 (8%)</td>
<td>0</td>
<td>1 (17%)</td>
<td>0</td>
<td>17 (5%)</td>
</tr>
<tr>
<td>ECVI confidence interval</td>
<td>1 (2%)</td>
<td>0</td>
<td>0</td>
<td>1 (17%)</td>
<td>0</td>
<td>12 (3%)</td>
</tr>
</tbody>
</table>

**Others**
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average standardized residuals</td>
<td>4 (8%)</td>
<td>1 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13 (4%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average off-diagonal standardized residuals</td>
<td>4 (8%)</td>
<td>2 (15%)</td>
<td>0</td>
<td>0</td>
<td>1 (10%)</td>
<td>12 (3%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bollen and Stein bootstrap</td>
<td>1 (2%)</td>
<td>1 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (0.3%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\chi^2$ difference test</td>
<td>15 (30%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>140 (39%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hoelmer's Critical N</td>
<td>1 (2%)</td>
<td>1 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (0.3%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null (independent) model $\chi^2$</td>
<td>4 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7 (2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Null (independent) model df</td>
<td>4 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7 (2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRMR, CFI</td>
<td>1 (2%)</td>
<td>1 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3 (1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRMR, TLI</td>
<td>1 (2%)</td>
<td>1 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3 (1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRMR, RMSEA</td>
<td>1 (2%)</td>
<td>1 (8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3 (1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRMR, IFI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russell (2002)</td>
<td>18 (32%)</td>
<td>6 (46%)</td>
<td>4 (50%)</td>
<td>3 (50%)</td>
<td>3 (30%)</td>
<td>132 (37%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TLI, $\chi^2$, RMSEA</td>
<td>12 (24%)</td>
<td>7 (54%)</td>
<td>2 (25%)</td>
<td>0</td>
<td>1 (10%)</td>
<td>67 (19%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMS, CFI, TLI, RMSEA with CI, SRMR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMS, CFI, TLI, RMSEA with CI, SRMR, CFI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMS, CFI, TLI, RMSEA with CI, SRMR, NFI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMS, CFI, TLI, RMSEA with CI, SRMR, TLI</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CMS, Gamma Hat, McDonald's Noncentrality Index</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. CI = Confidence interval. $^a$k = 50, $^b$k = 13, $^c$k = 8, $^d$k = 6, $^e$k = 10. $^f$Although these three statistics are presented separately for clarity, we reiterate that they are used together to examine model fit. $^g$Three articles containing a total of 31 models reported two estimation results (i.e., maximum likelihood and maximum likelihood robust estimations). Both estimation methods were separately coded. $^h$When one article reported multiple fit indices and thereby fell into multiple categories, coding was conducted as such. $^i$Including the Satorra-Bentler corrected chi-square.
### Table 6
*Data Type, Distribution, and Parameter Estimation*

<table>
<thead>
<tr>
<th>Category</th>
<th>Maximum likelihood</th>
<th>Robust maximum likelihood</th>
<th>Robust weighted least squares</th>
<th>least squares</th>
<th>Bootstrapping</th>
<th>Weighted least squares</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. At least five categories &amp; normally distributed</td>
<td>6 (12%)</td>
<td>[30 (8%)]</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2. At least five categories &amp; nonnormally distributed</td>
<td>--</td>
<td>1 (2%)</td>
<td>1 (0.2%)</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3. Fewer than five categories</td>
<td>3 (0.6%)</td>
<td>2 (0.4%)</td>
<td>0</td>
<td>--</td>
<td>1 (0.2%)</td>
<td>[1 (0.003%)]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strategy</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>2 (15%)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2.</td>
<td>--</td>
<td>0</td>
<td>--</td>
<td>1 (8%)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3.</td>
<td>--</td>
<td>--</td>
<td>0</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Motivation</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>0</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2.</td>
<td>--</td>
<td>0</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3.</td>
<td>--</td>
<td>--</td>
<td>0</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Social milieu</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>0</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2.</td>
<td>--</td>
<td>0</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3.</td>
<td>--</td>
<td>--</td>
<td>2 (34%)</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trait structure</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>0</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2.</td>
<td>--</td>
<td>1 (17%)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>3.</td>
<td>2 (34%)</td>
<td>2 (34%)</td>
<td>0</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

*Note.* This table included only studies that reported all three kinds of information (i.e., data type, distribution, and parameter estimation). A counted frequency outside the square bracket [ ] is based on an article as a unit of analysis (N = 50), whereas that inside the square bracket [ ] is based on a model as a unit of analysis (N = 360).
Table 7

*Number of Fit Indices Reported*

<table>
<thead>
<tr>
<th>Article</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Median</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>5.16</td>
<td>2.91</td>
<td>0</td>
<td>12</td>
<td>0.38</td>
<td>-0.41</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Strategy</td>
<td>7.00</td>
<td>2.35</td>
<td>2</td>
<td>10</td>
<td>-0.60</td>
<td>0.04</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Motivation</td>
<td>6.63</td>
<td>3.50</td>
<td>1</td>
<td>12</td>
<td>0.07</td>
<td>-0.03</td>
<td>5.5</td>
<td>5</td>
</tr>
<tr>
<td>Social milieu</td>
<td>6.33</td>
<td>3.56</td>
<td>1</td>
<td>12</td>
<td>0.19</td>
<td>1.79</td>
<td>6.5</td>
<td>--</td>
</tr>
<tr>
<td>Trait structure</td>
<td>3.90</td>
<td>2.69</td>
<td>0</td>
<td>10</td>
<td>1.12</td>
<td>2.58</td>
<td>3.5</td>
<td>5</td>
</tr>
<tr>
<td>Model</td>
<td>4.47</td>
<td>2.44</td>
<td>0</td>
<td>14</td>
<td>0.65</td>
<td>1.27</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

*Note.* Chi-squares, *p* values, and/or degrees of freedom are often reported together, so they are all coded as one fit index.
### Table 8
**Characteristics of Measurement Issues**

<table>
<thead>
<tr>
<th></th>
<th>Article</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test of data normality</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate normality check (skewness &amp; kurtosis)</td>
<td>22 (44%)</td>
<td>124 (34%)</td>
</tr>
<tr>
<td>Multivariate normality test</td>
<td>16 (32%)</td>
<td>105 (29%)</td>
</tr>
<tr>
<td>Not reported (univariate)</td>
<td>28 (56%)</td>
<td>236 (66%)</td>
</tr>
<tr>
<td>Not reported (multivariate)</td>
<td>34 (66%)</td>
<td>255 (71%)</td>
</tr>
<tr>
<td><strong>Missing data treatment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listwise</td>
<td>10 (20%)</td>
<td>98 (27%)</td>
</tr>
<tr>
<td>Pairwise</td>
<td>1 (2%)</td>
<td>4 (1%)</td>
</tr>
<tr>
<td>Mean replacement</td>
<td>1 (2%)</td>
<td>1 (0.3%)</td>
</tr>
<tr>
<td>Full information maximum likelihood</td>
<td>1 (2%)</td>
<td>24 (7%)</td>
</tr>
<tr>
<td>Imputation (multiple imputation)</td>
<td>1 (2%)</td>
<td>4 (1%)</td>
</tr>
<tr>
<td>Imputation (not specified)</td>
<td>2 (4%)</td>
<td>5 (1%)</td>
</tr>
<tr>
<td>No missing data</td>
<td>1 (2%)</td>
<td>1 (0.3%)</td>
</tr>
<tr>
<td>Not reported</td>
<td>33 (66%)</td>
<td>223 (62%)</td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amos</td>
<td>12 (24%)</td>
<td>51 (14%)</td>
</tr>
<tr>
<td>CALIS</td>
<td>1 (2%)</td>
<td>15 (4%)</td>
</tr>
<tr>
<td>EQS</td>
<td>15 (30%)</td>
<td>96 (27%)</td>
</tr>
<tr>
<td>LISREL</td>
<td>20 (40%)</td>
<td>187 (52%)</td>
</tr>
<tr>
<td>Mplus</td>
<td>1 (2%)</td>
<td>4 (1%)</td>
</tr>
<tr>
<td>Not reported</td>
<td>1 (2%)</td>
<td>7 (2%)</td>
</tr>
</tbody>
</table>
Table 9

Sample Sizes Among 50 Articles by Kline’s (2005) Definition

<table>
<thead>
<tr>
<th>Category</th>
<th>Small: 99 or less</th>
<th>Medium: 100–199</th>
<th>Large: 200 or above</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>6 (12%)</td>
<td>13 (26%)</td>
<td>31 (62%)</td>
</tr>
<tr>
<td>Strategy</td>
<td>0</td>
<td>3 (23%)</td>
<td>10 (77%)</td>
</tr>
<tr>
<td>Motivation</td>
<td>1 (13%)</td>
<td>2 (25%)</td>
<td>5 (63%)</td>
</tr>
<tr>
<td>Social milieu</td>
<td>0</td>
<td>1 (17%)</td>
<td>5 (83%)</td>
</tr>
<tr>
<td>Trait/test structure</td>
<td>0</td>
<td>4 (40%)</td>
<td>6 (60%)</td>
</tr>
</tbody>
</table>

*Note. To save space, we only reported results based on the article level (50 models) and did not report those based on the model level (360 models). This also applies to Table 10 and Figure 1.*
### Table 10

*Actual and Desired Sample Sizes by Raykov and Marcoulides (2006)*

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Median</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>851.96</td>
<td>2,266.44</td>
<td>75</td>
<td>15,000</td>
<td>5.40</td>
<td>32.31</td>
<td>258.5b</td>
<td>561</td>
</tr>
<tr>
<td>1. Actual sample size</td>
<td>46.24</td>
<td>31.41</td>
<td>10</td>
<td>132</td>
<td>1.09</td>
<td>0.48</td>
<td>35.5</td>
<td>10</td>
</tr>
<tr>
<td>2. No. of free parameters</td>
<td>462.40</td>
<td>314.06</td>
<td>100</td>
<td>1,320</td>
<td>1.09</td>
<td>0.48</td>
<td>355</td>
<td>100</td>
</tr>
<tr>
<td>3. Desired minimum sample size</td>
<td>561</td>
<td>561</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Median</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>490.54</td>
<td>426.82</td>
<td>102</td>
<td>1382</td>
<td>1.56</td>
<td>1.57</td>
<td>281</td>
<td>561</td>
</tr>
<tr>
<td>2.</td>
<td>59.62</td>
<td>28.49</td>
<td>10</td>
<td>124</td>
<td>0.67</td>
<td>1.47</td>
<td>58</td>
<td>63</td>
</tr>
<tr>
<td>3.</td>
<td>596.15</td>
<td>284.91</td>
<td>100</td>
<td>1240</td>
<td>0.67</td>
<td>1.47</td>
<td>580</td>
<td>630</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Median</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Motivation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>804.25</td>
<td>1607.61</td>
<td>75</td>
<td>4765</td>
<td>2.78</td>
<td>7.78</td>
<td>248</td>
<td>--</td>
</tr>
<tr>
<td>2.</td>
<td>53.88</td>
<td>20.52</td>
<td>30</td>
<td>93</td>
<td>0.76</td>
<td>0.78</td>
<td>57</td>
<td>57</td>
</tr>
<tr>
<td>3.</td>
<td>538.80</td>
<td>205.20</td>
<td>300</td>
<td>930</td>
<td>0.76</td>
<td>0.78</td>
<td>570</td>
<td>570</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Median</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social milieu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>1054.17</td>
<td>1824.04</td>
<td>137</td>
<td>4765</td>
<td>2.41</td>
<td>5.86</td>
<td>309.5</td>
<td>--</td>
</tr>
<tr>
<td>2.</td>
<td>55.67</td>
<td>31.87</td>
<td>24</td>
<td>95</td>
<td>0.48</td>
<td>–2.10</td>
<td>47.5</td>
<td>--</td>
</tr>
<tr>
<td>3.</td>
<td>556.67</td>
<td>318.73</td>
<td>240</td>
<td>950</td>
<td>0.48</td>
<td>–2.10</td>
<td>475</td>
<td>--</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Median</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trait structure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>452.80</td>
<td>708.81</td>
<td>110</td>
<td>2450</td>
<td>3.05</td>
<td>9.46</td>
<td>222</td>
<td>110</td>
</tr>
<tr>
<td>2.</td>
<td>28.80</td>
<td>21.98</td>
<td>13</td>
<td>90</td>
<td>2.90</td>
<td>8.86</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>3.</td>
<td>288</td>
<td>219.84</td>
<td>130</td>
<td>900</td>
<td>2.90</td>
<td>8.86</td>
<td>240</td>
<td>240</td>
</tr>
</tbody>
</table>

*Note.* aCalculated using a formula of the number of free parameters times 10 (Raykov & Marcoulides, 2006). bWhen four articles with the largest sample sizes of 15,000, 5,000, 4,765, and 2,450 were excluded from analysis, the mean, median, and mode were 334.41, 251, and 561. When seven articles with the largest sample sizes of 15,000, 5,000, 4,765, 2,450, 1,382, 1,175, and 1,113 were excluded from analysis, the mean, median, and mode were 272.4, 237, and 561. These results show that the median was stable regardless of the extreme values.
<table>
<thead>
<tr>
<th>Frequency (%)</th>
<th>Stem and Leaf</th>
<th>Extremes (Value: )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (2%)</td>
<td>Extremes (Value: –959)</td>
<td></td>
</tr>
<tr>
<td>4 (8%)</td>
<td>-6.3499</td>
<td></td>
</tr>
<tr>
<td>1 (2%)</td>
<td>-5.7</td>
<td></td>
</tr>
<tr>
<td>1 (2%)</td>
<td>-4.6</td>
<td></td>
</tr>
<tr>
<td>5 (10%)</td>
<td>-3.02588</td>
<td></td>
</tr>
<tr>
<td>6 (12%)</td>
<td>-2.013478</td>
<td></td>
</tr>
<tr>
<td>3 (6%)</td>
<td>-1.046</td>
<td></td>
</tr>
<tr>
<td>8 (16%)</td>
<td>-0.01124557</td>
<td></td>
</tr>
<tr>
<td>10 (20%)</td>
<td>0.0013446779</td>
<td></td>
</tr>
<tr>
<td>3 (6%)</td>
<td>1.238</td>
<td></td>
</tr>
<tr>
<td>0 (0%)</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>0 (0%)</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>1 (2%)</td>
<td>4.9</td>
<td></td>
</tr>
<tr>
<td>1 (2%)</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>6 (12%)</td>
<td>Extremes (Value: 752, 953, 2,240, 4,195, 4,240, and 13,680)</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Stem-and-Leaf Plot of Differences Between Actual and Desired Sample Sizes ($N = 50$). The differences have been calculated by the actual sample size minus the minimum desired sample size using Raykov and Marcoulides (2006). Stem width is 100, and each leaf is 1 case, for example, “–6.3, –6.4, –6.9, –6.9” in the second row means that differences between the actual and minimum desired sample sizes are “–630, –640, –690, –690.” This suggests that actual sample sizes were less than minimum desired and that these four articles should have extra sample sizes of 630, 640, 690, and 690 in addition to the original sample sizes.