A Review of Structural Equation Modeling Applications in Turkish Educational Science Literature, 2010-2015

Kubra Karakaya-Ozyer[1], Beyza Aksu-Dunya[2]
[1]Eskişehir Osmangazi University
[2]University of Illinois at Chicago

To cite this article:

A Review of Structural Equation Modeling Applications in Turkish Educational Science Literature, 2010-2015

Kubra Karakaya-Ozyer, Beyza Aksu-Dunya

Article Info	Abstract
	Structural equation modeling (SEM) is one of the most popular multivariate statistical techniques in Turkish educational research. This study elaborates the SEM procedures employed by 75 educational research articles which were published from 2010 to 2015 in Turkey. After documenting and coding 75 academic papers, categorical frequencies and percentages were calculated in the content analysis. Methodological issues related to use of SEM were identified. The results showed that usage and reporting SEM in Turkish educational research articles were mostly problematic. Content analysis revealed that educational researchers tend to avoid assumption testing of SEM and reporting estimation methods. General guidelines and suggestions on the use of SEM were provided.

Introduction

Structural Equation Modeling (SEM) is a technique to help examine cause and effect relationship among variables in various fields. SEM is also used to identify linear causation among latent and observed variables. These latent variables can be represented by multiple observed variables. In SEM, a hypothesized model is created displaying directional and non-directional relationships between latent and observed variables. In general, SEM is used to examine if the model is to account for variation and covariation of the observed or latent variables (MacCaluum & Austin, 2000). Furthermore, theory development and construct validation are two important topics that can be addressed using SEM when analyses are conducted properly.

SEM has two fundamental components: a measurement model and a structural model (Kline, 2005). The measurement model is basically a confirmatory factor analysis model which confirms if the data fit to the proposed model. The measurement model establishes relationship between latent variables and multiple observed variables. The structural model identifies interrelationships among latent variables in the hypothesized model. Although SEM was introduced to the educational sciences earlier, this method has started to become popular recently in Turkish educational research (Sumer, 2000).

There were approximately 150 studies using SEM between the years of 2010 and 2015. Even though SEM has many advantages for complex models, its benefits can only be obtained when properly applied. SEM requires several assumptions (e.g., linearity, normality) difficult to meet in practice. Mistakes and misuse of SEM can create invalid research results which may lead to incorrect information and knowledge produced in the literature. Hence, it is important to identify the problems in applying SEM in educational research literature.

The purpose of the study is to evaluate the use of SEM in 75 educational research articles published in Turkey between the years of 2010 and 2015. Because the use of SEM technique in educational science in Turkey has dramatically increased in recent years, Turkish educational research papers were chosen to review as the focus of this study. There is no prior study on the use of SEM technique in Turkish education literature. This review addresses disparities in the application of SEM across studies from different journals published in Turkey, some of them namely: *Education and Science, Hacettepe University Journal of Education, Educational Sciences: Theory and Practices, Contemporary Educational Technology, Eurasian Journal of Educational Research.* After a critical review of existing methodology studies, this paper provides guidelines for researchers on applying SEM.

Method

This study focused on 75 educational research studies that were published in Turkey. An assessment form was developed for data collection process. This form contained six categories such as data characteristics, assumptions and estimation methods. Methodological issues related to use of SEM were examined through content analysis method.

Data

Studies which used SEM as an analysis tool were considered in this study. We selected all the Turkish Educational Journals as the target study area. All issues from 2010 to 2015 were reviewed for identifying empirical SEM research and theoretical explanation of the SEM was not included in the current study.

There were several articles that used confirmatory factor analysis (CFA) as an SEM model; however, CFA is distinct from SEM. As a result, studies used confirmatory factor analysis models were excluded from the sample journal articles. Specifically, traditional SEM, multi-group SEM and path models were selected as the applications of SEM for data analysis. After the elimination and quick review, ten journal articles were excluded and a total of 75 articles were identified as the applications of SEM. Figure 1 shows number of studies using SEM from 2010 to 2015. While there were only seven studies using SEM in 2010, this number has increased gradually over time. In 2015, 21 education research studies were conducted which demonstrated that SEM has become a popular tool in Turkish education literature.
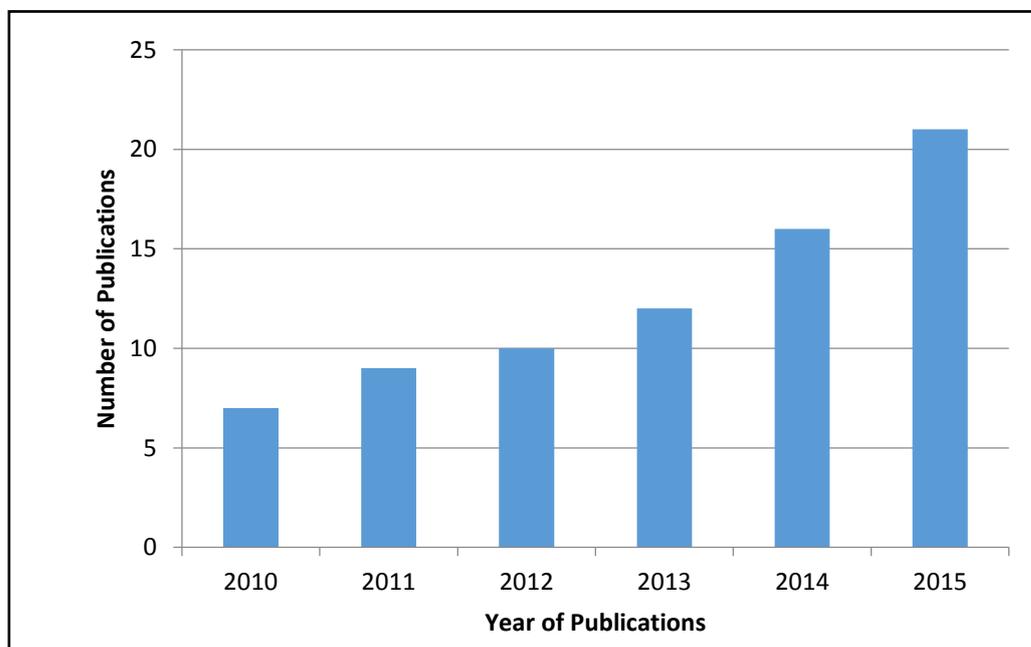

Figure 1. Number of publications that used SEM by year

Analyses

Methodological review studies were synthesized to conduct a coding table for the current study (see Appendix A). Combining categories from six journal articles, the coding table was developed to review the educational research studies published in Turkey (Baumgartner & Homburg, 1996; Gefen, Straub, & Boudreau, 2000; Holbert & Stephenson, 2002; MacCallum & Austin, 2000; Schreiber, Nora, Stage, Barlow, & King, 2006; Shook, Ketchen, Hult, & Kacmar, 2004). Then, each article was treated as a unit of analysis and was evaluated under six categories: data characteristics, reliability and validity, evaluating model fit, model estimation, model re-specification and reporting. The frequencies and percentages under each category were recorded and presented on Table 1 below. One researcher reviewed all articles and coded each category. Especially for data characteristics section, each research coded as "yes" and "no information". Table 1 demonstrated all the coding categories and Appendix A provide all important information for coding table.

Results

The review results demonstrated that there are six critical issues in the SEM applications: data characteristics, reliability and validity, evaluating model fit, model estimation, model re-specification and reporting. Table 1 summarizes the results drawn from the educational research articles published in Turkey.

Table 1. Frequencies concerning six categories

Category	Frequency (*n*)	Percentage (%)
Data Characteristics Issues		
Models designed as cross-sectional	75	100
Data that checked for outliers	7	9.3
Data that checked for missing values	11	14.6
Data that checked for normality	18	24
Sample size < 150	6	8
Sample size > 1000	11	14.6
Reliability and Validity Issues		
Models with reliability information	67	89.3
Models with validity information	51	68
Evaluating Model Fit Issues		
Models using chi-square test	44	58.6
Models providing p value of chi-square test	28	37.3
Models using chi-square/*df*	49	65.3
Models using Goodness of Fit Index	58	77.3
Models using Adjusted Goodness of Fit Index	48	64
Models using Comparative Fit Index	65	86.6
Models using Bollen's Incremental Fit Index	20	26.6
Models using Normed Fit Index	48	64
Models using Non- normed Fit Index	25	33.3
Models using Trucker- Lewis Index	15	20
Models using Relative Fit Index	8	10.6
Models using Root of Mean Square Error of Approximation	70	93.3
Models using Confidence Interval of RMSEA	6	8
Models using Root of Mean Square Residuals	11	14.6
Models using Standardized Root of Mean Square Residuals	32	42.6
Model Estimation Methods		
Models using Maximum Likelihood method	23	30.6
Models using Asymptotic Covariance Matrix method	2	2.6
Models using Robust Maximum Likelihood method	2	2.6
Model Re-specification		
Models that modified error covariance	10	13.3
Models that deleted items	2	2.6
Models that deleted paths	3	4
Reporting		
Models that provided correlation matrix	33	44
Models that provided degrees of freedom	44	58.6
Models that provided power analysis	2	2.6
Models that provided computer programs	63	84

Note: *N*= 75.

Data Characteristics

Cross-sectional versus Longitudinal

A cross sectional design is a type of study that collects data from a single point of time. All the variables in the study were examined and measured concurrently (MacCallum & Austin, 2000). On the other hand, in a longitudinal design study, data is collected from repeated observations of the same variables in a long-time period. Because SEM analysis has been described as a causal modeling in the literature, Kelloway (1995) suggested that temporal ordering of variables should be demonstrated before the inference of results. Nevertheless, many SEM articles designed as a cross sectional study, theoretical background of variables should

be emphasized before SEM analysis. We coded one time period studies as cross sectional studies, while more than one time period studies were coded as longitudinal studies. In Turkish Educational literature, all the SEM application articles were found cross sectional designs.

Outliers and Missing Values

Data should be checked for possible coding errors and missing values before using SEM. After correcting the coding errors in data, it is important to handle missing values properly (Kaplan, 1990; Kline, 2005). There are several missing value analysis methods (e.g., pairwise deletion, listwise deletion or series mean). Research indicates importance of handling missing values for the further steps of SEM (Raykov, Tomer, & Nesselroade, 1991). If the missing values are not random values, then pairwise and listwise deletion methods are not recommended (Schreiber et al., 2006).

The percentage of studies that did not address missing values in their data was 85.4 % (*n*= 64). Only 11 articles indicated that they checked their data for missing values. Most of them indicated that they eliminated the cases that have missing values.

Besides coding errors and missing values, outliers are crucial issues to address in order to meet the assumptions of SEM (Çokluk, Şekercioğlu, & Büyüköztürk, 2014; Kline, 2005). Maximum likelihood estimation, a commonly used estimation method, assumes that data is drawn a multivariate normal population. Severe outliers can cause non-normal data distribution; hence, examining and handling outliers are important to meet the normality assumption for maximum likelihood estimation. Only 9.3 % (*n*=7) of the articles mentioned the outlier screening in their results. They checked skewness and kurtosis values, difference between trimmed mean and mean and Mahalanobis distance measure.

Normality

Traditional SEM is an advance statistical technique which requires meeting multivariate normality assumption for data. MacCallum, Roznowski and Necowitz (1992) suggested that non-normal data distribution may lead the invalid goodness of fit indices and results. According to Ullman (2001), estimation methods for SEM analysis are affected by normality of data.

Unfortunately, there was little evidence that the researchers tried to demonstrate data screening constraints for their data. We checked if the normality screening was described in the papers or not. Only 24% (*n*=18) of the selected articles indicated that they check the data for normality and demonstrated that the data were approximately normal. There are different type of normality screening methods such as Bartlet's test, skewness and kurtosis values, Lilliefor's test and Kolmogorov- Smirnov test. None of the articles handled non-normal data issue. The results indicated that maximum likelihood is the most commonly used estimation technique while there was not sufficient attention to check the assumption of multivariate normality. Moreover, it is clear that there were no common normality screening method in the Turkish Education literature. As a result, we recommend that multivariate normality should be checked at the pre-analysis period of SEM.

Sample Size

Sample size is another crucial aspect that should be considered prior to SEM analysis because it relates to the stability of the parameter estimates (Çokluk, et al., 2014; MacCallum & Austin, 2000; Schreiber et al., 2006). In general, SEM requires a larger sample relative to other complex and advanced statistical techniques. To reach the valid test statistics, sample size should be large enough. For SEM analysis, there is no certain rule of thumb for sample size; however, researchers suggested several ways for the number of participants needed. 10 per estimated parameter rule is suggested by Schreiber et al. (2006) as a general consensus. Having said that, Bentler and Chou (1987) recommended that the ratio of sample size to number of free parameters should be 5:1. Likewise, Tanaka (1987) suggested that trustworthy results and stable test statistics can be obtained with 4:1 sample size-to parameters ratio. Anderson and Gerbing (1988) suggested 150 participants for a minimum level but some other researchers claimed that 200 is an acceptable sample size level for SEM analysis (Chou & Bentler, 2000). For moderately reliable results, Hoyle and Kenny (1999) recommend to achieve at least 100 participants per model. MacCallum, Widaman, Zhang and Hong(1999), and MacCallum and Austin (2000) studies suggested that it is difficult to find a rule of thumb for minimum sample size because complexity of the

model can cause discrepancies in the parameter estimates and model fit indices. Hair, Black, Babin, Anderson, and Tatham (2010) claimed five considerations which can affect the sample size for SEM analysis. These are multivariate normality of data, estimation method, model complexity, amount of missing data and average error variance among the reflective indicators. Our findings showed that, 8% (*n*= 6) articles had 150 or below participants, 25.3 % (*n*=19) articles had between 150 and 300 participants, 52% (*n*=39) articles between 300 and 1000 participants and finally 14.6 % (*n*=11) articles had 1000 and above participants as sample population. Only three articles out of 75 mentioned which rule they used to determine the sample size adequacy for their complex models. Given the results, it can be concluded that researchers did not pay sufficient attention insufficient to determine sample size before using SEM.

Reliability and Validity Issues

Reliability and validity of instruments are crucial criteria which researchers need to address. Reliability is defined as consistency of measurement results. Validity is the accuracy of interpretations of results. Hence, reliability and validity analyses give researchers a hint that SEM results can be interpreted correctly.

Cronbach alpha coefficient is one of the most common reliability coefficients used in Turkish educational studies. Although Cronbach alpha coefficient is frequently used, it has several limitations. For instance, Cronbach alpha assumes that all items contribute equally to reliability (Bollen, 1989). However, according to Bollen (1989) composite reliability estimates use standardized loadings and measurement errors for each single item. As a result, composite reliability gives more accurate results than Cronbach alpha estimation method.

We checked if an article examined reliability and if yes, what type of reliability method was used. 89.3 % (*n*=67) of the studies reported reliability for measures they used for SEM. Forty-two studies out of 67 relied on Cronbach's alpha, five studies used composite reliability coefficient, 13 studies indicates test-retest reliability of the measures, three studies gave KR-20 coefficient, one study used maximal reliability, one study used Mc Donald's omega coefficient and one study relied on eight studies did not report any reliability coefficient.

In Marsh and Hau's (2001) study, researchers need to be extra cautious while assessing the validity of results (as cited in Green, 2015, p. 3; Shook et al., 2004). First of all, based on acceptable fit indices and factor loadings, an acceptable measurement model should be obtained (Anderson & Gerbing, 1988). After the identification of a measurement model, researchers need to provide detailed information about justification of their proposed model (McDonald & Ho, 2002). Other validity methods (e.g., convergent validity, discriminant validity, content validity) are also crucial for SEM analysis. Bagozzi, Yi and Singh (1991) introduced discriminant validity to determine if constructs are significantly different to improve the model fit. The average variance extracted should be more than 50% to achieve an acceptable convergent validity (Shook et al., 2004). On the other hand, measurement model with acceptable fit gives evidence of convergent and discrimination validity (Anderson & Gerbin, 1988).

We looked at if researchers addressed validity and reported which validity method they used. In total, 68% (*n*=51) of studies examined validity of their results while 32% (*n*=24) of articles did not mention anything about the validity in their results. Moreover, convergent validity was reported in 11 studies and discriminant validity was reported in 6 studies. 15 articles mentioned about the construct validity for SEM analysis. Four studies reported language validity while translating the measures to Turkish language. Content validity was mentioned in 3 studies. There are six studies which indicated that the measures were valid but they did not mention any specific method to achieve validity. Two of the studies used CFA and one study used EFA to address validity.

Evaluating Model Fit

Evaluating model fit is the most controversial issue in SEM literature (Kline, 2005; Özdamar, 2016). Even though there are several attempts to obtain "rule of thumps" for assessment of SEM, there is no consensus on fit indices which show an acceptable fit (Davcik, 2014). After the SEM analysis was conducted, researchers need to take important steps for the assessment of the model. At first, the data is assessed if it fits with a prior model. After the overall model fit, individual parameters are assessed. To assess overall model fit, researchers use several goodness-of-fit indices. The most commonly used fit index is chi-square statistics which is not recommended for large sample size (Çokluk et al., 2014; Karadag, 2012; Shook et al., 2004). The other common fit indices are the Normed Fit Index (NFI), Goodness Fit Index (GFI), Comparative Fit Index (CFI), Root Mean

Square Error of Approximation (RMSEA), Standardized Root Mean Residual (SRMR) (Gefen et al., 2000; Hair et al., 2010). Table 2 summarizes some of the commonly used fit indices and decision criteria.

Table 2. Evaluation of fit indices

Fit Indices	Criteria	Rule of Thumb	Source
χ^2	$p > 0.05$	Not significant	McDonald & Ho (2002)
χ^2/df		$2 < \chi^2/df < 3$ = acceptable fit	Schermelleh-Engel, Moosbrugger, & Müller (2003)
	Small sample size	≤ 2.5 = perfect fit	Kline (2005)
	Large sample size	≤ 3 = perfect fit	Kline (2005)
	Small and moderate sample size	< 3	Hair et al. (2010)
GFI	0 = not fit 1= perfect fit	≥ 0.90 = good fit	Schumacker & Lomax (1996); Hair et al., 2010
		≥ 0.95 = perfect fit	Hair et al. (2010); Schermelleh-Engel, Moosbrugger & Müller, (2003)
CFI	0 = not fit 1= perfect fit	≥ 0.90 = acceptable fit	Hair et al. (2010); McDonald & Ho, (2002)
		≥ 0.95 = perfect fit	Hu & Bentler (1999); Kline (2005); Schermelleh-Engel, Moosbrugger & Müller (2003); Thompson (2004)
NFI	0 = not fit 1= perfect fit	≥ 0.90 = acceptable fit	Schumacker & Lomax (1996); Thompson (2004)
		≥ 0.95 = perfect fit	Hu & Bentler (1999); Schermelleh-Engel, Moosbrugger & Müller, (2003)
RMSEA	0 = perfect fit 1= not fit	≤ 0.05 = perfect fit	Kline (2005); McDonald & Ho (2002); Schumacker & Lomax (1996)
		≤ 0.06 = acceptable fit	Hu & Bentler (1999); Thompson (2004)
RMR/ SRMR	0 = perfect fit 1= not fit	≤ 0.05 = perfect fit	Kline (2005); McDonald & Ho (2002); Schermelleh-Engel, Moosbrugger & Müller (2003)
		≤ 0.08 = acceptable fit	Hair et al. (2010); Hu & Bentler (1999)
		≤ 0.10 = poor fit	Hair et al. (2010); Kline (2005)

Note. χ^2 = chi-square statistic; χ^2/df = ratio of chi-square statistics and degrees of freedom; GFI = goodness of fit index; CFI = comparative fit index; NFI = normed fit index; RMSEA = root mean square error of approximation; RMR = root mean square residual; SRMR = standardized root mean residual.

In chi-square statistics test, the null hypothesis that the estimated covariance matrix derivate from the sample covariance matrix. The goodness of fit index (GFI), Adjusted goodness of fit index (AGFI) and Root mean square residual (RMR) assess model fit in an absolute sense. NFI and NNFI are incremental fit indices that compare the hypothesized model to the fit of a baseline model (Baumgartner & Homburg, 1996).

In literature, different fit indices worked well with different sample sizes, data types and ranges of acceptable scores (Hu & Bentler, 1999; MacCallum, Browne, & Sugawara, 1996). While the chi-square statistic was provided by 58.6% (*n*=44) of the models cited in literature, only 28 models provided significance for the chi-square statistic. However, only 22 of the models produced a non-significant chi-square value. Since chi-square statistic is influenced by sample size, most of the researchers preferred to report X^2/df value to assess the model fit (65.3% , *n*= 49). The RMSEA was provided for most of the models (93.3%). Only six studies provided confidence internals of RMSEA as a measure of overall fit. Based on the most recent research, the other popular model fit indices are found as CFI and GFI which were reported by 86.6% and 77.3% of the models, respectively. Despite the popularity of GFI usage, GFI was criticized as a poor measure of overall model fit (Hu & Bentler, 1999). AGFI and NFI was reported by 64% (*n*= 48) of models while NNFI was provided by only 33.3% of the models.42.6% of the models provided SRMR and 14.6 % of the models reported RMR while evaluating the model fit. IFI was provided by only 26.6% of models and only 8 studies reported RFI. Finally, TLI was reported by 15 models. Most of the common computer programs for SEM report numerous fit indices; hence, researchers need to make decisions about which model fit index is the most suitable for their research. It is recommended that multiple fit indices (such as chi-square, RMSEA, CFI and TLI) should be reported to show the data-model fit in a research (Hu & Bentler, 1999; Kline, 2005; Shook et al., 2004).

Model Estimation

After the model specification, estimation method should be indicated. Maximum likelihood (ML), robust maximum likelihood (RML), unweighted least squares (ULS) and weighted least squares (WLS) are some approaches to estimate item parameters in a SEM. Even though there are several available options for a SEM solution, maximum likelihood (ML) estimation is one of the most commonly used SEM estimation procedures (Davcik, 2014; Green, 2015). Researchers should consider the complexity of their model, sample size, and distribution of their data to decide which estimation method is most suitable. Research showed that under some assumptions (i.e., *n* > 50) ML estimation method produce valid results (Holbert & Stephenson, 2002). In addition, to perform ML estimation for a SEM analysis, researchers need to collect large sample size and assess multivariate normality with normally distributed errors (Jöreskog & Sörbom, 1988). According to Hair et al. (2010), ML gives more efficient and stable results when the data is normally distributed. Slightly departure from normality can be handled by ML or Generalized Least Squares procedures; however, interpretation of chi-square statistics and standard errors can be problematic (Hair et al., 2010). If the data is non-normal, than other forms of estimation method should be used for a SEM study. Weighted Least Square (WLS), Robust Maximum Likelihood (RML) and Asymptotically Distribution Free (ADF) estimation methods are popular for data deviating from normality (Jöreskog & Sörbom, 1996; Muthen, 1984). It is important to report the estimation method used and the distribution of the data (Holbert & Stephenson, 2002). Besides, Loehlin (1998) suggested reporting multiple estimators for SEM analysis. Similar results of different estimators demonstrate that ML gives acceptable solution for SEM. In contrast, researchers should decide which estimation method gives the accurate solution based on the sample size and model type. According to Holbert and Stephenson (2002), ADF can be preferable under major violation of normality. Generally, some computer programs (e.g., LISREL, AMOS, Mplus) use ML as a default option in every multivariate analysis. As a result, researchers need to assess multivariate normality of the data, display descriptive statistics (especially skewness and kurtosis) and pay attention to apply SEM procedure in statistical software packages.

Recent studies showed that more than half of the researchers (57.3%, *n* = 39) did not mention which estimation method they used specifically. On the other hand, 30.6% (*n* = 23) of the recent articles used ML estimator, two of them used ACM and two of them used RML. Eleven of the 23 studies which used ML estimator checked normality of the data. The study results indicated that almost half of the researchers did not report which estimation method they used that is because they did not know the impact of the estimation method used for SEM analysis. Researchers probably do not see a need to consider the estimation method because they have a lack of concern for the validity of the normality assumptions. Moreover, the reason for preferring ML estimation procedure over the other methods seems to be that it is the default method in the most statistical programs.

Model Re-Specification

In educational science, it is common to re-specify the initial model because most of the models do not fit the data well. However, respecification of a SEM model is controversial issue in the literature (Davcik, 2014). Finn and Kayande (2004) claimed that there is little attention to evaluate consequences of model modifications and they have pointed out that a modified model can be reliable for one specific purpose but not for another.

Jöreskog (1993) demonstrated a way to get final model via the LaGrange multiplier test when the initial model does not fit the data well. There are research claiming that modifying a model should be based on the theory and the content (Anderson & Gerbing, 1988). On the other hand, Chin (1998) and Kelloway (1995) recommended that not only theory-based modifications but also new sample is needed to justify a re-specified model. Even though the model is validated by a new sample, Brannick (1995) claimed that respecification cannot be done in any way. If the modification indices show a new model and this model can be justified by the theory, then this new model can be used as an alternative model rather than a modified model.

Respecification of the initial model is still a vague issue but the modification indices suggest important modifications on paths and relationships. Even though respecification was not recommended by many researchers, the researchers should report the modification test results (Lagrange, Wald or chi-square) with a theoretical support. After a researcher modified the initial model, chi-square difference statistics should be reported to show that new model is better than the prior model. As a result, researchers should indicate why and how they use modification indices and justify the new model with the theory.

After analyzing the structural model, model misspecification can be detected by various tools. Modification indices are available in the output of the statistical software (e.g., LISREL, EQS, AMOS…). Researches can look at these indices and improve the model fit level (Baumgartner & Homburg, 1996). Some researchers described the modification of the initial model and but some others did not mention any modification issue in their analyses. Hence, it is difficult to identify how many initial models were modified. In 80% ($n = 60$) of all the articles researchers did not mention any form of modification or respecification of the initial models. 10 studies provided information that they used error covariance modification indices and respecified initial models to obtain adequate model fit. Three these studies deleted the paths between latent variables. Finally, only one researcher acknowledged that two items are highly correlated and decided to delete one of them. As a result, item numbers reduced based on the modification indices.

Reporting

An ideal SEM research should provide key details to make it replicable (Shook et al., 2004). Replication can be possible only if the details about input matrix, degrees of freedom, power analysis results, criteria for evaluating values of fit indexes and the computer software and version are sufficiently described (Chin, 1998; MacCallum & Austin, 2000).

Covariance versus Correlation Matrix

SEM was developed based on covariance matrix over time (Cudeck, 1989). In the past, researchers should have decided which matrix they were going to use to conduct SEM analysis; however, nowadays SEM programs can compute the parameter estimates and fit indices from the raw data. Even though statistical programs make every step easier, researchers should provide either correlation or covariance matrix for interpretations of the results (Hair et al., 2010; Holbert & Stephenson, 2002). About 44% ($n=33$) of the articles fit models to correlation matrix. Nevertheless, most of the researchers (53%, $n= 40$) seem unaware of potentially serious problems that can be caused by failing to indicate the matrix used for SEM. LISREL program can correct estimation of a correlation structure while AMOS program cannot accept correlation matrix for SEM analysis. Only 2.6% ($n=2$) SEM applications indicated that they used covariance matrix and explained why they used it.

Degrees of Freedom

Degrees of freedom (*df*) demonstrate the number of parameter that vary independently (Hair et al., 2010). There is a formula to calculate degrees of freedom. However, most of the researchers prefer using software to calculate *df*. It is important to provide *df* in a study to ensure statistical validity of a chi-square tests. Consequently, degrees of freedom gives more information for validity of outcomes. Despite these issues, the findings showed that 41.3 % of the studies ($n=31$) did not provide degrees of freedom. To that end, the many researchers notably did not consider *df* as a key output to validate their solutions.

Power Analysis

Another key aspect to emphasize is power analysis. To get valid and stable results from model fit indices, an adequate statistical power is essential. The findings in the recent research demonstrated that most of the articles did not provide statistical power. Only two of them gives Cohen'd value to report the power. MacCallum, Browne, and Sugawara (1996) gave a list to decide minimum sample size for adequate power. Moreover, there are several power analysis software programs to check the power of the model fit. Unfortunately, 41.3% of the articles did not provide degrees of freedom; hence, it is impossible to check their studies' power by hand.

In this research, we found that there are some articles that demonstrated significant model fit results with small sample size. This may due to the lack of statistical power. As a result, researchers should consider power analysis before interpreting their results.

Computer Programs

There are several statistical programs to perform SEM analysis. The most widely and the first used computer program is LISREL (Linear Structural Relations) (Hair et al., 2010; Özdamar, 2016). The latest version of LISREL gives opportunity to use graphical interface as well as syntax option. Moreover, EQS and AMOS (Analysis of Moment Structure) are other widely used statistical programs for SEM analysis (Özdamar, 2016). Mplus is a flexible syntax based program and it allows multi-group SEM analysis in the same model.

Based on the literature, 53.3% (n=40) of the models were analyzed by LISREL program while 29.3% of the articles used AMOS for SEM analysis. On the other hand, 12 of the studies did not mention any software name or version and 10 of the research provided a program name but did not indicate the version number. One article claimed that they used both LISREL and AMOS for the analysis part. It is clear that Turkish educational researchers prefer LISREL and AMOS statistical programs more than Mplus or EQS. This is because LISREL and AMOS have graphical interface which can be easy to draw path diagram and set measurement and structural models all together.

Criteria for Evaluating Values

Assessing a model fit is another controversial issue in SEM literature (Özdamar, 2016). Thus, providing a criterion for evaluation fit indices is very important to permit replication of studies by other researchers. To accept a measurement model, researchers should obtain low chi-square values with p values greater than .05. Furthermore, high goodness of fit indices (e.g., CFI, GFI, NFI, NNFI, or TLI) and low RMSEA, SMRS and RMR are expected to be obtained from acceptable models. In the literature, there is no consensus about which index gives accurate results in specific conditions (Bentler, 1990; Raykov et al., 1991).

Discussion

Educational researchers have been employing SEM technique for more than a quarter century for theory testing and causal modeling (Raykov et al., 1991). Due to growing number of applications of SEM in educational research in Turkey, it is crucial to address methodological flows that existing studies carry for the sake of future SEM applications. Therefore, this study aimed to fill an existing gap in the SEM literature by examining methodological issues in the existing studies and providing guidelines on how they can be handled. The guidelines will also be helpful for researchers who want to run SEM analysis and new to the field. More specifically, these guidelines were prepared to help researchers in providing clear data description, model interpretation and evaluation and explicit reporting of results.

As emphasized by other studies (Davcik, 2014; Guo, Perron, & Gillespie, 2009; Holbert & Stephenson, 2002), clear description and evaluation of SEM analysis are needed to interpret causation in a structural equation model. Several important considerations associated with data description and evaluation were focus of this study. First of all, the evaluated studies had problems in checking data characteristics. Checking research design, missing values and outliers, normality of variables and sample size of are crucial to decide parameter estimation method. Even small number of missing values and outliers can lead to the non-normal distribution of variables (Kline, 2005). Majority of the articles used cross-sectional design without missing value and outlier explanation and most of them did not provide normality of data. Hence, researchers should carefully screen data

before analyzing it. The significance of outliers can be checked by analyzing the model with and without those cases. As a result, the effect of outliers on the actual results can be displayed and the researchers can decide which analysis is better for the collected data. Estimating a complex model with small sample is another error which researchers tend to commit. Even though there were no consensus about the adequate sample size, the ratio of cases to the number of parameters less 10:1 or less than 150 number of observation are considered as inadequate number to run a SEM analysis (Holbert & Stephenson, 2002; Kline, 2005). Almost all of the investigated articles claimed that their sample size is more than 150. However, these assumptions are made for maximum likelihood (ML) estimation method. It is important to remember that other estimation methods (e.g., ADF) require very large sample.

Secondly, lack of reliability and validity analyses were concerning. A lot of measurement errors in the scores can lead to unreliable and inaccurate results; thus, estimates about latent variables may be too high and causal effects may be underestimated (Kline, 2005). Even though most of the articles provided reliability of their results; there were arguments that which reliability coefficient is the best for SEM analysis. More than third of the SEM studies provided validity information but there were no consensus on the type of validity aspect reported. Discriminant validity, convergent validity, content validity, construct validity, language validity were some of them. Education researchers must recognize the limitation of unreliable and invalid results and measures. Hence, we encourage researchers to display reliability coefficients and validity analysis along with their rationales.

While McDonald and Ho (2002) claimed that CFI, GFI, NFI and the NNFI are the most commonly reported indices, Hu and Bentler (1999) recommended reporting SRMR, NNFI or TLI, RMSEA and CFI all together. However, there were some researchers who tended to report overall model fit indices and avoided to assess other types of fit indices. Specifically, it is possible to manipulate the results from SEM analysis and researchers should provide detailed information about fit indices to achieve validity. It is important to say that fit indices are a useful guide but without strong substantive theory these numbers do not provide valid results (Hooper, Coughlan, & Mullen, 2008).

The most commonly used estimation method for SEM is ML which assumes multivariate normality for endogenous variables (Holbert & Stephenson, 2002; Kline, 2005). Other estimation methods such as WLS and RML can be used if the distributions of the continuous variables are severely non-normal (Davcik, 2014). Moreover, researchers need to use a different and more appropriate estimation method if categorical endogenous variables are in their models (Kline, 2005).

According to Kline (2005), any complex models may tend to fit the data perfectly even though reasons for causality are not provided. Hence, by simply adding error disturbance or measurement error correlations without solid reasons may improve fit indices but this may mask the fact (Davcik, 2014). Researchers who want to modify their models should provide substantive evidence to add measurement error correlations when they aim to obtain acceptable model fit.

Recommendations

While this study aimed to shed light on some common mistakes observed on the recent studies using SEM approach, there are still issues that need to be cleared on conducting methodologically strong SEM analyses. The future studies may provide more guidelines on acceptable fit indices, selection of estimation method and reporting of SEM results. Power analysis for each model, mediator and moderator effects, and the interpretation of the SEM results are other issues that need to be examined. It is also important to conduct simulation studies to address controversial issues in SEM analysis since simulation gives researchers more authority to manipulate variables.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411-423.

Bagozzi, R. P., Yi, Y., & Singh, S. (1991). On the use of structural equation models in experimental designs: Two extensions. *International Journal of Research in Marketing, 8*(2), 125-140.

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13*(2), 139-161.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*(1), 78-117.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17*(3), 303-316.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16*(3), 201-213.

Chou, C. P., Bentler, P. M., & Pentz, M. A. (2000). A two-stage approach to multilevel structural equation models: Application to longitudinal data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data, practical issues, applied approaches, and specific examples* (pp. 33-49). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Chin, W. (1998). Commentary: Issues and Opinion on Structural Equation Modeling. *MIS Quarterly, 22*(1), Vii Xvi. Retrieved from http://www.jstor.org/stable/249674

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları.* Ankara, Turkey: Pegem Akademi.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*(2), 317-327.

Davcik, N. (2014). The use and misuse of structural equation modeling in management research: A review and critique. *Journal of Advances in Management Research, 11*(1), 47-81.

Finn, A., & Kayande, U. (2004). Scale modification: alternative approaches and their consequences. *Journal of Retailing, 80*(1), 37-52.

Gefen, D., Straub, D., & Boudreau, M. C. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Systems, 4*(1), 7-21.

Green, T. (2016). A methodological review of structural equation modelling in higher education research. *Studies in Higher Education, 41*(12), 2125-2155.

Guo, B., Perron, B. E., & Gillespie, D. F. (2008). A systematic review of structural equation modelling in social work research. *British Journal of Social Work, 39*(8), 1556-1574.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *Multivariate data analysis.* Upper Saddle River, NJ: Pearson.

Holbert, R. L., & Stephenson, M. T. (2002). Structural equation modeling in the communication sciences, 1995 2000. *Human Communication Research, 28*(4), 531-551.

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60.

Hoyle, R. H., & Kenny, D. A. (1999). *Statistical strategies for small sample research.* Thousand Oaks, CA: Sage

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

Jöreskog, K. G. (1993). Testing structural equation models. *Sage Focus Editions, 154,* 294.

Joreskog, K. G., & Sorbom, D. (1988). *Prelis. A program for multivariate data screening and data summarization. User's guide.* Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software International.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research, 25*(2), 137-155.

Karadag, E. (2012). Basic features of structural equation modeling and path analysis with its place and importance in educational research methodology. *Bulgarian Journal of Science and Education Policy, 6*(1), 194-212.

Kelloway, E. K. (1995). Structural equation modelling in perspective. *Journal of Organizational Behavior, 16*(3), 215-224.

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

Loehlin, J. C. (1998). Latent variable models: *An introduction to factor, path, and structural analysis.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*(1), 201-226.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 1-30.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99.

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64-82.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132.

Özdamar, K. (2016). *Eğitim, sağlık ve davranış bilimlerinde ölçek ve test geliştirme yapısal eşitlik modellemesi.* Eskişehir, Turkey: Nisan.

Raykov, T., Tomer, A., & Nesselroade, J. R. (1991). Reporting structural equation modeling results in Psychology and Aging: Some proposed guidelines. *Psychology and Aging, 6*(4), 499-503.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323-338.

Schumacker, R. E., & Lomax, R. G. (1996). *A beginner's guide to structural equation modeling.* Mahwah, NJ: Laurence Erlbaum.

Shook, C. L., Ketchen, D. J., Hult, G. T. M., & Kacmar, K. M. (2004). An assessment of the use of structural equation modeling in strategic management research. *Strategic Management Journal, 25*(4), 397-404.

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar. *Türk Psikoloji Yazıları, 3*(6), 49-74.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development, 58,* 134-146.

Thompson, B. (2004*). Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Ullman, J. B. & Bentler, P. M. (2003). Structural equation modeling. *Handbook of Psychology, 4,* 607–634.

Author Information

Kubra Karakaya-Ozyer
Eskişehir Osmangazi University
Eskişehir, Turkey
Contact e-mail: *kozyer@ogu.edu.tr*

Beyza Aksu-Dunya
University of Illinois at Chicago
Chicago, Illinois, USA

Appendix A

Coding table

Evaluating Model Fit

p value of chi-square test provided (Yes:1, No:2)	Chi-square/df provided (Yes:1, No:2)	Goodness of Fit Index provided (Yes:1, No:2)	Adjusted Goodness of Fit Index provided (Yes:1, No:2)	Comparative Fit Index provided (Yes:1, No:2)	Bollen's Incremental Fit Index provided (Yes:1, No:2)	Normed Fit Index provided (Yes:1, No:2)	Non-normed Fit Index provided (Yes:1, No:2)	Trucker-Lewis Index provided (Yes:1, No:2)	Relative Fit Index provided (Yes:1, No:2)	RMSEA provided (Yes:1, No:2)	Confidence interval of RMSEA provided (Yes:1, No:2)	RMR provided (Yes:1, No:2)	SRMR provided (Yes:1, No:2)

Data Characteristics					Reliability and Validity		Model Estimation			Model Re-specification			Reporting			
Cross-sectional design (Yes:1 No:2)	Missing value Provided (Yes:1, No:2)	Outliers provided (Yes:1, No:2)	Normality check (normal data:1, non-normal data:2, NA:3)	Sample size	Reliability info provided (Yes:1, No:2)	Validity info provided (Yes:1, No:2)	Maximum Likelihood (Yes:1, No:2)	Asymptotic Covariance Matrix (Yes:1, No:2)	Robust Maximum Likelihood (Yes:1, No:2)	Modified error covariance (Yes:1, No:2)	Deleted items (Yes:1, No:2)	Deleted paths (Yes:1, No:2)	Correlation matrix provided (Yes:1, No:2)	Degrees of freedom provided (Yes:1, No:2)	Power analysis provided (Yes:1, No:2)	Computer programs provided (Yes:1, No:2)