
Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
10.1111/1467-8551.12415

Link to publication record in Explore Bristol Research
PDF-document

This is the final published version of the article (version of record). It first appeared online via Wiley at https://onlinelibrary.wiley.com/doi/full/10.1111/1467-8551.12415. Please refer to any applicable terms of use of the publisher.

**University of Bristol - Explore Bristol Research**

**General rights**

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/
Evaluating the Use of Covariance-Based Structural Equation Modelling with Reflective Measurement in Organizational and Management Research: A Review and Recommendations for Best Practice

Mary F. Zhang, Jeremy F. Dawson and Rex B. Kline

School for Policy Studies, University of Bristol, 10 Woodland Road, Bristol BS8 1SZ, UK, Institute of Work Psychology, Sheffield University Management School, Conduit Road, Sheffield S10 1FL, UK, and Department of Psychology, Concordia University, 7141 Sherbrooke W. Montréal, Montreal, Quebec, Canada

Corresponding author email: mary.zhang@bristol.ac.uk

Covariance-based structural equation modelling (CB-SEM) with reflective measurement has been a popular data analysis tool in organizational and management research. Extensive studies and guidelines have been published on what constitutes its best practice. What is much less known is the extent to which CB-SEM users in organizational and management research comprehend and adhere to the standards and principles behind this advanced analytical technique. In this study, we first devised an evaluation scheme to assess the quality of CB-SEM performed in a study, and then utilized this scheme to examine 144 CB-SEM studies published in 12 top organizational and management journals between 2011 and 2016. The evaluation of the published studies revealed a pressing need for more systematic and standardized approaches to planning, conducting and reporting CB-SEM studies. We discussed the implication of the findings for future work.

Introduction

Covariance-based structural equation modelling (CB-SEM), especially with reflective measurement where hypothetical constructs are estimated as common factors that are assumed to cause their indicators (i.e. observed or manifest variables), is a flexible and compelling data analysis method. It has become widely used in organizational and management research (Williams, Vandenberg and Edwards, 2009). As other members in the SEM family, CB-SEM has several appealing features relative to some other frequently used analytical methods. First, it is an integration of several multivariate techniques – for example, regression analysis, path analysis and confirmatory factor analysis (Cheung, 2015). It can perform a simultaneous analysis of observed variables and latent structures, their relations and their impact on the corresponding outcomes (Cudeck, Jöreskog and Sörbom, 2001). Second, CB-SEM can account for measurement error in both the predictive and outcome variables (Grewal, Cote and Baumgartner, 2004), providing a more accurate estimate of the model parameters and effects and offering a better control for both the measured and the latent factors (Cheung and Lau, 2008; Hoyle and Smith, 1994). Third, CB-SEM allows a series...
of contrasting models to be tested, interpreted and compared quantitatively (Mitchell, 1992). In doing so, it can help researchers identify the best approximating models that are theoretically precise and parsimonious (Burnham and Anderson, 2013).

Given the widespread use of CB-SEM, extensive studies and guidelines have been published on what constitutes its best practice. What is much less known is the extent to which researchers adhere to these standards and principles, especially in the context of organizational and management research. Such knowledge is crucial, as it can help researchers, students, reviewers and editors to identify, clarify and explain critical issues in applying this advanced analytical technique (MacCallum and Austin, 2000). More importantly, it echoes the intensively debated replication crisis in social and behavioural sciences (Gelman, 2018; Simmons, Nelson and Simonsohn, 2011; Szucs and Ioannidis, 2017) and provides a timely instance of the endeavour to maximize research transparency and replicability (Haller and Krauss, 2002; Ioannidis, 2005; John et al., 2012; Kerr, 1998).

As said, it is not difficult to find textbooks or review papers on the recommendations for best CB-SEM practice. Worth further investigation is whether ‘what ought to be done’ matches ‘what actually has been done or reported’, and why and how CB-SEM can be (in)appropriately applied in examining the theories, hypotheses and data in organizational and management research. To the best of our knowledge, few reviews have been published to facilitate users of CB-SEM to understand the ‘what’ (the best practices are), ‘why’ (failure to meet these criteria can lead to impacted organizational and management scholarship) and ‘how’ (they can be achieved in empirical practices) questions simultaneously.

In this paper, we attempt to bridge this gap, by first identifying what researchers may reasonably consider as best practices in CB-SEM, then reviewing recent publications in top management and organizational journals in which CB-SEM was applied, and evaluating how closely they followed best practices. We also identify areas of best practice that need greater attention from researchers. We use our findings to give recommendations about steps that researchers using CB-SEM should follow. In doing so, our contribution is twofold: examining the state-of-the-art in management and organizational studies, and giving clear advice for what practices scholars should follow.

It is worth mentioning some alternative methods that can analyse composites or weighted combinations of observed variables. For example, the CB-SEM technique of confirmatory factor analysis (CFA) estimates common factors as proxies for hypothetical constructs, and CFA can test a wide range of hypotheses about measurement from the perspective of classical test theory. The technique of confirmatory composite analysis (CCA) is for measurement models where composites instead of common factors approximate hypothetical constructs (Rigdon, 2012). The CCA method is basically a series of steps implemented within the framework of partial least squares (PLS) path modelling, also known as PLS-SEM (Hair, Howard and Nitzl, 2020). The CCA technique can be applied to analyse either reflective measurement models or formative measurement models where latent variables are assumed to be caused by their indicators. Although formative measurement models can also be tested in CB-SEM, doing so can be challenging, because: (1) there are special identification requirements that can be difficult to satisfy; (2) technical problems in the analysis, such as nonconvergence of iterative estimation, can be encountered; and (3) large sample sizes are needed (e.g. Bollen and Davis, 2009). Other potential advantages of CCA over CFA include the generation of more precise estimates in small samples and greater likelihood of convergence when analysing models with many observed or latent variables. The technique of CCA may also be preferred when the primary research goal is prediction, or to maximize variation in dependent variables rather than the confirmation of measurement theory (see Hair, Howard and Nitzl, 2020; Rigdon, 2014 for more information). In this paper, we restrict our attention to reflective measurement models as evaluated in CB-SEM. To save space, the term ‘SEM’ in the following refers to CB-SEM unless otherwise indicated.

Examining 144 studies published in 12 top organizational and management journals between 2011 and 2016, our review reveals a pressing need for more care and prudence in SEM applications. We call for organizational and management journals to establish a more explicit and standardized way of conducting and reporting SEM studies. This work may serve as one step towards this goal.
Devising an evaluation scheme

Framework

We view the SEM technique as falling within a wide context of data use in scientific research. Burnham and Anderson’s (2013) work on data reduction suggests that model development should follow four main steps: (i) model formulation, that is, building up a set of candidate models according to logic and scientific knowledge; (ii) model specification, that is, selecting plausible, testable and informative models from a wider range of candidate models for making detailed examinations; (iii) model estimation, that is, estimating model parameters; and (iv) model evaluation, that is, assessing the accuracy and validity of the tested models and their scientific implications in concrete research contexts. Extending this framework, we further argue that model formulation ought to be a comprehensive and strategic preparation stage. It should not only focus on building up the hypothesized models for testing, but also needs to embrace a careful consideration on sample size, statistical power, multivariate normality and other such issues central to the generalized estimating equations underlying the SEM technique. On the other hand, depending on the complexity of the datasets and models to be tested, there may not always be a clear distinction between the model formulation and specification stages.

A consensual approach

To identify important methodological issues at each of the four stages (model formulation, specification, estimation and evaluation), we adopted a consensual approach reviewing recent seminal work on best SEM practice, including but not limited to Appelbaum et al. (2018), Goodboy and Kline (2017), Hoyle and Isherwood (2013), MacCallum and Austin (2000), McDonald and Ho (2002), Mueller and Hancock (2008), Nunkoo, Ramkissoon and Gursoy (2013) and Shah and Goldstein (2006). Issues emphasized by approximately 80% of the early work were considered as critical and served as a foundation for the preliminary evaluation scheme. After piloting the initial scheme, discussing the ambiguities and redundancies in wording and the evaluation standards, and consulting with SEM experts and frequent users of SEM for their comments, the scheme was edited and refined again, leading to a total of nine major domains as the focus.

Evaluation criteria and examples

We now turn to the details of these nine evaluation dimensions and their corresponding criteria (in total 16 standards). Each criterion is presented in the format of a (set of) Yes/No question(s), followed by a detailed explanation of the meaning and importance. To help our readers understand how a criterion can be met in concrete studies, examples taken from previous work are presented in Table 1.

1. Justification: Does the study give specific reasons or justifications for using SEM or a specific form of it? To meet this criterion, a study should give one or more clear reasons about why SEM or any specific form of it (e.g. multilevel or cross-lagged SEM) is utilized. This could include but not be limited to the relative gain from the advantages of SEM, the consideration of methodological precision and so on.

Importance: Researchers should be specifically aware of the various advantages that SEM can bring and of whether the associated statistical assumptions and requirements are met in the concrete research context, in order to maximize the utility of this powerful technique.

2. Hypothesis: (2.1) Does the study specify the overall structural equation model(s) to be tested? To meet this criterion, a study must specify one or more hypothesized models to be tested. (2.2) Does the study specify the relations between the variables or constructs? To meet this criterion, a study must specify the relations between constructs included in the SEM.

Importance: SEM is essentially a confirmatory technique, although it can sometimes be used for exploratory purposes (McIntosh, Edwards and Antonakis, 2014). It is inappropriate to let SEM and its fitness indices guide the maintenance or deletion of correlations between different variables or their residuals, in order to ‘make poorly fitting models appear passable’ (Hermida et al., 2015, p. 25). It is important to have a solid theoretical framework – or at least strong precedents from which one or a set of candidate models can be generated, tested and compared (Burnham and Anderson, 2013).

3. Statistical power: (3.1) Does the study justify the sample size? To meet this criterion, a
Table 1. Examples of appropriate applications

<table>
<thead>
<tr>
<th>Coding criteria</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Does the study give specific reasons or justifications for using SEM or a specific form of it?</td>
<td>‘… besides controlling for measurement errors, an important strength of SEM is its capability to test all hypothesized relationships simultaneously’ (Nifadkar, Tsui and Ashforth, 2012, p. 1158).</td>
<td>‘… we tested all hypotheses using multilevel structural equation modelling… (which) is able to capture the nested nature of the data, examine multiple mediated and moderated relationships simultaneously, and… provide more accurate estimations of the proposed relationships’ (Hu and Liden, 2015, p. 1109).</td>
</tr>
<tr>
<td>(2.1) Does the study specify the overall structural equation model(s) to be tested?</td>
<td>‘The model we advance is shown in Figure 1’ (Kirkman et al., 2011, p. 1236).</td>
<td>To deepen our understanding of the relationships between these predictors and of the reasons why they predict job performance, we used structural equation modelling (via EQS) to test the model depicted in Figure 1’ (Lievens and Patterson, 2011, p. 933).</td>
</tr>
<tr>
<td>(2.2) Does the study specify the relations between the variables or constructs?</td>
<td>All the studies reviewed met this criterion by specifying concrete research hypotheses to be tested.</td>
<td>‘Because the ratio of sample size to number of estimated parameters is an important concern in structural equation modelling… we used parcels as indicators of feeling trusted and emotional exhaustion’ (Baer et al., 2015, p. 1646).</td>
</tr>
<tr>
<td>(3.1) Does the study justify the sample size?</td>
<td>‘Because we tested relations among latent variables, we created indicators from dimensional scores or item parcels using the item-to-construct-balance method to reduce the number of parameters to be estimated…’ (Ou et al., 2014, p. 48).</td>
<td>‘To ensure that the data permitted a valid testing of our hypotheses, we conducted a priori power analyses, using the procedures and conventional effect sizes suggested by Cohen (1988)... As our actual sample of 72 for each measurement point was only slightly smaller, this was of minor concern… Acknowledging that SEM imposes higher sample requirements, multiple analyses supported the stability of our results, demonstrating that they are not artifacts of any particular analytic approach’ (Kim, Hornung and Rousseau, 2011, p. 1687).</td>
</tr>
<tr>
<td>(3.2) Does the study test statistical power?</td>
<td>The power of our analyses was found to be 1.0 for a test of close fit…’ (McCarthy, Trougakos and Cheng, 2016, p. 284).</td>
<td>‘Models were estimated using the maximum likelihood estimation with robust standard errors due to non-normality in the indicators’ (Kaltiainen, Lipponen and Holtz, 2016, p. 640).</td>
</tr>
<tr>
<td>(4) Are distributional assumptions of the method(s) respected in the data?</td>
<td>‘… we ensured that the assumptions of normality… were met’ (de Stobbeleir, Ashford and Buyens, 2011, p. 821).</td>
<td>‘Of the 223 firms that we visited in wave one (T1), 133 firms (including 133 CEOs, 133 CFOs and 469 other senior managers) provided complete information (have answered each question) for all the wave one variables…’ (Wei and Wu, 2013, p. 396).</td>
</tr>
<tr>
<td>(5.1) Does the study report incomplete data? (Note: respondents may provide incomplete data, which, however, is different from non-responses.)</td>
<td>‘After two reminders, a total of 207 firms had responded to the survey, a response rate of 21%. However, because of missing answers, only 169 responses were usable for statistical analysis’ (Foss, Laursen and Pedersen, 2011, p. 989).</td>
<td>‘We had missing data for some teams… We tested the degree to which missing data were random or systematic by examining means and standard deviations for measures of teams with complete data with the means and standard deviations of teams that had missing data’ (Lanaj et al., 2013, p. 746).</td>
</tr>
<tr>
<td>(5.2) Does the study clearly discuss the ways of dealing with missing data, if presented?</td>
<td>‘… Mplus uses full information maximum likelihood estimation that allows for missing data under the missing at random assumption’ (Gielen, Klemann and Consultancy, 2015, p. 1017).</td>
<td>(Continued)</td>
</tr>
<tr>
<td>Coding criteria</td>
<td>Example 1</td>
<td>Example 2</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>(6) Does the study calculate score reliability coefficients in its own sample(s)?</td>
<td>‘The mean of these ratings was then calculated to create a reliable ($\alpha = 0.87$) measure...’ (Mortensen, 2014, p. 921).</td>
<td>‘Reliability estimates for all measures exceeded 0.70’ (Ragins et al., 2012, p. 766).</td>
</tr>
<tr>
<td>(7) Does the study distinguish the measurement model from the structural model?</td>
<td>‘The first step in analyzing our data was examining the adequacy of our measurement model’ (Colquitt and Rodell, 2011, p. 1193).</td>
<td>‘Prior to testing the hypothesized structural model, we tested to see if the measurement model had good fit’ (Mayer et al., 2012, p. 159).</td>
</tr>
<tr>
<td>(8.1) Does the study report RMSEA and its 90% or 95% CIs?</td>
<td>‘The study reported chi-square test, CFI, TLI, SRMR and RMSEA with 90% CI in Tables 2 and 4’ (Stanhope, Pond and Surface, 2013, pp. 824, 828).</td>
<td>‘The fit statistics for this model indicated acceptable fit, $\chi^2(365) = 557.56, p &lt; 0.01$, RMSEA = 0.06 (90% CI [0.05, 0.07]), CFI = 0.91, TLI = 0.90, and SRMR = 0.07’ (Cullen et al., 2014, pp. 1770–1771).</td>
</tr>
<tr>
<td>(8.2) Does the study report SRMR?</td>
<td>‘Values shown are unstandardized parameter estimates, with standard errors in parentheses’ (Ferguson et al., 2016, p. 528).</td>
<td>‘The middle panel of Table 3 presents factor loadings and error variances...’ (Bagozzi et al., 2012, p. 71).</td>
</tr>
<tr>
<td>(8.3) Does the study report CFI or TLI?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8.4) Does the study report the result of chi-square test for the model?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(9) Does the study report residuals, that is, quantitative measures of model–data discrepancy at the level of pairs of observed variables?</td>
<td>‘The middle panel of Table 3 presents factor loadings and error variances...’ (Bagozzi et al., 2012, p. 71).</td>
<td></td>
</tr>
</tbody>
</table>

Notes: RMSEA = root mean square error of approximation; CIs = confidence intervals; SRMR = standardized root mean square residual; CFI = comparative fit index; TLI = Tucker–Lewis index.

The study needs to explicitly state at least one of the following issues: (a) information about the appropriateness and sustainability of the ratio between sample size and the number of estimated parameters, or (b) concerns about the relatively small sample size of the study and the corresponding strategies to handle this potential problem (e.g. justifications for using parcels). (3.2) Does the study test statistical power? To meet this criterion, a study needs to explicitly state a numerical estimate of statistical power for tests of the model(s) or individual effects. Importance: SEM is a ‘power-hungry’ technique that generally requires the ratio between the number of observations and the number of estimated parameters to be large (e.g. the often-quoted 20:1; see Jackson, 2003; Kline, 2016). A study with insufficient sample size and statistical power may fail to reject an incorrectly or inadequately hypothesized model, due to a non-significant chi-square test of the difference between the data and the model (Kim, 2009). Another consequence of low statistical power is that the detection of close-fitting models in the population may fail even if such models exist. Thus, researchers applying SEM should consider whether their research has a sufficient sample to test the hypothesized model(s) or individual effects.

4. Distributional assumptions: Are distributional assumptions of the method(s) respected in the data? To meet this criterion, a study needs to examine and specify whether the data used in the SEM meet the assumption of multivariate normality or whether appropriate methods (e.g. bootstrap, permutation, maximum likelihood estimation) are used to correct the fiducial estimates when the distributions for continuous outcome variables are non-normal (Anderson and Braak, 2003; Cheung, 2009). Importance: The assumption of multivariate normality is critical to SEM, especially when, for instance, the methods of default maximum likelihood estimation or generalized least squares assumptions are used (McDonald and Ho, 2002; Mueller, 1997). The violation of this assumption may lead to incorrect standard errors for individual effects or an inflated estimate of the model chi-square (Curran, West and Finch, 1996; Fabrigar et al., 1999), and thus a wrong
rejection of the hypothesized model (i.e. Type I error).

5. **Missing data**: (5.1) Does the study report incomplete data? To meet this criterion, a study should report the number or percentage of cases for which some variables are known but some are unknown (i.e. missingness), or the study should at least report the number or percentage of cases that can provide complete data to each variable (respondents may provide incomplete data, which, however, is different from non-responses). (5.2) Does the study clearly discuss the ways of dealing with missing data, if presented? Methods of dealing with missing data include – but are not limited to – listwise deletion, pairwise deletion, multiple imputation, full information maximum likelihood estimation for incomplete datasets and so on (see Allison, 2003; Brown, 1994; Kline, 2016; Larsen, 2011; McDonald and Ho, 2002).

(5.3) Does the study deal with missing data in an appropriate way, if presented? To meet this criterion, a study should adopt appropriate methods to deal with missing data. First, a study should examine the pattern of the missingness, that is, whether the circumstances of missing data are ignorable (non-systematic and less than 5% missing) or not (systematic or more than 5% missing; see Kline, 2016). Within the ‘non-systematic missing’ category, researchers next need to further examine whether the data are missing completely at random or missing at random (see Allison, 2003; Rubin, 1976). Thereafter, researchers should specify the ways that they adopt to deal with missing data, ideally, with the justification of one method over another. **Importance**: The ways of dealing with missing data in SEM are critical to the estimates of standard errors, model parameters and test statistics (Allison, 2003; Larsen, 2011), and yet many studies are not clear about this important step in their analysis (Kline, 2016). To increase the generalizability and reproducibility of their findings, researchers should report details of the approach(es) to dealing with missing data.

6. **Reliability**: Does the study calculate score reliability coefficients in its own sample(s)? A study meets this criterion if it examines the internal consistency (e.g. alpha coefficient), temporal stability (i.e. test–retest reliability) or interrater reliability of the observed measures. **Importance**: The reliability of scores in a particular sample generally estimates the proportion of observed variation not due to random measurement error (Raines-Eudy, 2000). Score reliability is critical in many, if not most, types of statistical methods for behavioural data, because the analysis of imprecise scores can severely bias the results. Through the specification of manifest variables with error terms as indicators of hypothetical latent variables, score unreliability in SEM can be explicitly estimated in the analysis. Nevertheless, high levels of imprecision can seriously distort results (Cole and Preacher, 2014). A consequence of such distortion is unstable or poor fit of a theoretically feasible model to the data (Brannick, 1995). This criterion is consistent with the appeal in general reporting standards for quantitative studies to estimate and report reliability coefficients for the scores analysed (e.g. Appelbaum et al., 2018).

7. **Measurement vs. structural model**: Does the study distinguish the measurement model (i.e. hypotheses about relations between factors and indicators) from the structural model (i.e. hypotheses about causal effects between factors)? To meet this criterion, a study needs to test the general adequacy of the measurement model before examining the overall fit and statistical properties of the whole model with both its measurement and structural components; otherwise, there is a potential confound in the basic sources for poor model fit. **Importance**: A well-appreciated advantage of SEM is its ability to display and assess the structural model and the measurement model simultaneously (Anderson and Gerbing, 1988; Landis, Beal and Tesluk, 2000). However, this feature may sometimes become a limit, as the failure to ‘distinguish between the measures of a construct and the construct itself’ (Williams, Gavin and Williams, 1996, p. 89) can lead to a vague understanding and potentially misleading interpretation of the results. Imagine that a study reports a model with poor fit to the data. Without a test of the properties of the measures in advance, it is hard to distinguish if this poor fit is due to misspecification about causal relations or the inappropriateness of the measures (e.g. low reliability or validity). Therefore, it is best to first assess the psychometric properties of the measure of each variable before inspecting the overall model fit.
8. **Global fit**: Does the study report a series of goodness-of-fit indices, including (8.1) root mean square error of approximation and its 90% or 95% confidence intervals, (8.2) standardized root mean square residual, (8.3) comparative fit index or Tucker–Lewis index and (8.4) chi-square? To meet this criterion, a study needs to report these goodness-of-fit indices.

**Importance**: A variety of goodness-of-fit indices are developed based on different assessing assumptions of what comprises a good model (see Kaplan, 2009) and are able to provide a continuous rather than a coarse and dichotomous evaluation of the match between the proposed structural model and the data (Mulaik et al., 1989). It is thus recommended to report a full range of goodness-of-fit indices of the model and to avoid only presenting indices that may particularly favor the hypothesized model on any arbitrary basis. To meet this criterion, a study needs to report the values of goodness-of-fit indices that reflect different aspects of model quality (Kaplan, 2009).

9. **Local fit**: Does the study report residuals, that is, quantitative measures of model–data discrepancy at the level of pairs of observed variables? To meet this criterion, a study needs to report on the standardized, normalized, covariance or correlation residuals. An alternative is to report conditional independences or empirical values of partial correlations expected to equal zero after controlling for all causal effects or non-causal associations between a pair of observed variables (Pearl, 2009).

**Importance**: The failure to report residuals or conditional independences is a serious shortcoming in SEM studies. It can happen that values of global fit statistics look reasonable, while evidence of grossly poor fit is clear in the residuals. For simpler models, it may be possible to present a whole residual matrix in a table. In more complicated models with many observed variables, though, the residuals should at least be described in the main text, and tables or appendices of the residuals should be available in the supplemental materials (Goodboy and Kline, 2017).

### Utilizing the evaluation scheme

#### Sampling

We used the above scheme to assess the quality of SEM application in studies published between 2011 and 2016 in 12 top organizational and management journals, including *Academy of Management Journal*, *Administrative Science Quarterly* and so on (see the comprehensive list in Table 2). These journals were selected as they are acknowledged as prominent in organizational and management research, covering a variety of timely and important issues in these fields (Conlon et al., 2006; Molina-Azorin, 2012). We chose 2011–2016 as the timeframe, considering the number of journals, studies and criteria focused, and the recent computational and statistical advances in SEM. To appraise earlier studies is admittedly more comprehensive, but may bias our judgement on the status quo of current SEM application in organizational and management research.

Among all the manuscripts published between 2011 and 2016 in these selected journals, keywords were used to further search publications that might adopt SEM. They included components of the term ‘SEM’ and their combinations (e.g. ‘structure’, ‘structural equation’, ‘model’, ‘model(ing)’), commonly used goodness-of-fit indices (e.g. ‘RMSEA’, ‘SRMR’, ‘CFI’, ‘TLI’ – see the meaning of these abbreviations in the footnote to Table 1) and frequently used software packages for conducting SEM (e.g. ‘MPlus’, ‘AMOS’, ‘LISREL’, ‘EQS’, ‘lavaan’). This keyword searching returned 365 academic papers that might have applied SEM.

We excluded 100 studies in which SEM was only used in the form of CFA to evaluate purely measurement models. Examples included the application of CFA to test construct validity (i.e. convergent and discriminant validity) or to evaluate common method variance. Such models generally feature covariances between pairs of factors without presuming direct causal effects, and do not usually raise many of the issues related to the overarching principles of the SEM technique. Likewise, studies without latent variables (i.e. path analysis) were excluded (N = 77). In addition,

---

1. If all outcome variables are continuous, then SEM computer tools usually print values for the model chi-square, RMSEA, CFI and SRMR in the output. However, if some outcomes are categorical or the model includes interactive effects of continuous latent variables, then not all of the aforementioned global fit statistics will be calculated, and thus cannot be reported.

© 2020 The Authors. British Journal of Management published by John Wiley & Sons Ltd on behalf of British Academy of Management.
studies using meta-analytic (N = 8) or Bayesian structural modelling (N = 3), latent change or growth modelling (N = 19) or partial least squares SEM (N = 9) were excluded, as these modelling methods have specific statistical assumptions and approaches for handling the data and analyses (Hair, Ringle and Sarstedt, 2012; Hoch and Kozlowski, 2014; Jak, 2015; Lee, 2007; Nunkoo, Ramkissoon and Gursoy, 2013; Ployhart, Van Iddekinge and Mackenzie, 2011). Four studies applied SEM in creative but uncommon research designs. One study did not adopt SEM but contained the keyword ‘structure equation’—all these were excluded from further analyses. In total, 144 papers were included in the final sample, among which 130 were cross-sectional, 11 longitudinal and 3 experimental or quasi-experimental (see Table 2). The unit in this evaluation was each individual publication; in a few cases where the researchers used more than one SEM in a single publication, their ways of dealing with different structural models were assessed and graded as a whole.

Evaluation procedure and reliability

On each criterion, the publications received a ‘Yes’ for satisfying it or a ‘No’ for not. Based on the evaluation criteria, two coders (the first and second authors) evaluated a randomly selected 17 papers from the sample together. The remaining 127 publications were coded by the first author. We calculated Cohen’s kappa (Cohen, 1960) for testing the level of consistency in the two coders’ ratings of the 17 randomly selected publications and found that the inter-rater agreement reached a high level (κ = 0.93, p < 0.001; McHugh, 2012). A careful check of the nine instances in the coding (about 3% out of the 272 pairs of coding scores) revealed that the inconsistencies were mainly due to one coder failing to spot the relevant information in the articles. After discussing each of these inconsistencies, the two coders by the end reached 100% agreement on the coding.

Results

Figure 1 illustrates the percentage of studies that have satisfied each evaluation criterion. For example, about 42% of the examined studies provided explicit justifications for why SEM was adopted in their research. It is apparent that some non-negotiable standards were met almost without exception (e.g. 100% of reviewed studies specified research hypotheses), whereas other criteria remained largely unsatisfied. Overall, criteria 2.2 (hypothesizing specific relations within the model), 6 (calculating score reliability), 8.3 and 8.4 (reporting CFI, TLI and chi-square of the structural model) have been met well (i.e. over 90% of reviewed publications met these standards), fol-

Table 2. Number of publications using SEM in selected journals between 2011 and 2016

<table>
<thead>
<tr>
<th>Year</th>
<th>AMJ</th>
<th>ASQ</th>
<th>BJM</th>
<th>JAP</th>
<th>JoM</th>
<th>JoMS</th>
<th>MS</th>
<th>OS1</th>
<th>OS2</th>
<th>OBHDP</th>
<th>PP</th>
<th>SMJ</th>
<th>Total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>26</td>
<td>18.1</td>
</tr>
<tr>
<td>2012</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>22</td>
<td>15.3</td>
</tr>
<tr>
<td>2013</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>28</td>
<td>19.4</td>
</tr>
<tr>
<td>2014</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>26</td>
<td>18.1</td>
</tr>
<tr>
<td>2015</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>22</td>
<td>15.3</td>
</tr>
<tr>
<td>2016</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>20</td>
<td>13.9</td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>1</td>
<td>14</td>
<td>33</td>
<td>23</td>
<td>12</td>
<td>1</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>21</td>
<td>4</td>
<td>144</td>
<td>14.6</td>
</tr>
</tbody>
</table>


© 2020 The Authors. British Journal of Management published by John Wiley & Sons Ltd on behalf of British Academy of Management.
Scientific Application of SEM

![Figure 1. Percentage (%) of publications (N = 144) satisfying each evaluation criterion [Colour figure can be viewed at wileyonlinelibrary.com]](image)

**Notes:**
1. Only about 7% (N = 10) of the reviewed publications reported RMSEA together with its 90% or 95% confidence intervals (CIs); 76% (N = 109) of the publications reported RMSEA without the 90% or 95% CIs; and 17% (N = 24) of the publications did not report RMSEA. In addition, Boh and Wong (2015) discussed why RMSEA was not reported and was coded as ‘not applicable’. These led to a low score of meeting this criterion.
2. The coding and evaluation of each individual publication is available upon request. [Color figure can be viewed at wileyonlinelibrary.com]

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Justification (1)</td>
<td>42</td>
</tr>
<tr>
<td>Hypothesis (2.1)</td>
<td>76</td>
</tr>
<tr>
<td>Hypothesis (2.2)</td>
<td>100</td>
</tr>
<tr>
<td>Sample size &amp; statistical power (3.1)</td>
<td>33</td>
</tr>
<tr>
<td>Sample size &amp; statistical power (3.2)</td>
<td>2</td>
</tr>
<tr>
<td>Distributional assumptions (4)</td>
<td>18</td>
</tr>
<tr>
<td>Missing data (5.1)</td>
<td>40</td>
</tr>
<tr>
<td>Missing data (5.2)</td>
<td>56</td>
</tr>
<tr>
<td>Missing data (5.3)</td>
<td>21</td>
</tr>
<tr>
<td>Reliability (6)</td>
<td></td>
</tr>
<tr>
<td>Measurement vs. structural model (7)</td>
<td>81</td>
</tr>
<tr>
<td>Fit index - RMSEA (8.1)</td>
<td>7</td>
</tr>
<tr>
<td>Fit index - SRMR (8.2)</td>
<td>53</td>
</tr>
<tr>
<td>Fit index - CFI/TLI (8.3)</td>
<td>94</td>
</tr>
<tr>
<td>Fit index - chi-square (8.4)</td>
<td>92</td>
</tr>
<tr>
<td>Local fit (9)</td>
<td>17</td>
</tr>
</tbody>
</table>

Looking more closely at the less-attended standards, we found that there were high proportions of studies lacking the justification for using SEM (58%), screening of missing data (60%), consideration of sample size or statistical power (67%), or examination of distributional assumptions such as multivariate normality (82%). Moreover, despite that most studies apparently managed to report ‘response rates’ (i.e. number of participants accepting to participate or returning the questionnaires); only 40% of them further presented the percentage missing of each research variable – or at least, the percentage of cases that provided a complete response to each question. A much smaller number of studies (i.e. 21%) reported the reasons for a method (e.g. listwise deletion or imputation) being used to deal with missing data and the consequences (e.g. possible selection biases) that may be attributed to using such method. Another striking finding was that only 17% of studies reported local fit indices such as residuals, 7% reported RMSEA with its 90% or 95% confidence intervals, and 2% explicitly provided a numerical estimate or examination of statistical power of the structural model(s) or individual effects. We discuss the implication of these findings in the next section.

**Conclusion and discussion**

Our review is in line with early observations in communication (Goodboy and Kline, 2017), tourism (Nunkoo, Ramkissoon and Gursoy, 2013) and operations management (Shah and Goldstein, 2020). The Authors. British Journal of Management published by John Wiley & Sons Ltd on behalf of British Academy of Management.
2006) that there still lacks transparency in reporting critical steps in data preparation and analysis (e.g. how the study dealt with missing data). The reviewed studies unfortunately failed to convey that a strategic research plan with appropriate analysis at its core was in place before the study was conducted, and can hardly be replicated by future studies in similar settings. However, as discussed at the beginning, such guidelines and reporting standards are not scarce. A more interesting question then becomes: why are pitfalls in writing, reporting and potentially conducting SEM studies widespread, especially in the face of plenty of published best-practice recommendations and reporting standards?

We infer that the discrepancy between ‘what has been commanded’ and ‘what has been followed’ is probably due to two reasons. First, perhaps sometimes researchers are pressured to use a state-of-the-art technique that is more complicated than essential. As noted by Floyd (2014), many of the existing publications are now filled with convoluted SEMs that are simply unnecessary to test the claims of the studies. It seems that this modelling technique has become an end unto itself. Researchers encouraged or pressured to apply SEM may do so with insufficient preparation or training in psychometrics (Lambert, 1991).

A closely related misconception in SEM studies is that an ultimate structural model must ‘fit’ the data. However, nothing could be further from the truth. This is because any model, even one that is grossly wrong, can be made to fit the data simply by making it more complicated or adding free parameters (Cheung and Rensvold, 2001). If all possible free parameters are estimated (i.e. \( df = 0 \)), then model fit is likely to be perfect. It can also happen that models with very few degrees of freedom (e.g. \( df = 1 \)) have near-perfect fit, but such models may have so many free parameters relative to the number of observations that they can hardly fail to explain the data substantially. One of the main goals of SEM is to test a theory (Hayduk et al., 2007). This means that it is perfectly acceptable to retain no model at the end of a SEM analysis. Indeed, this outcome is preferred over demonstrating that the data are explained by a scientifically meaningless model (Millsap, 2007). Perhaps due to the misconception that an ultimate SEM must be ‘successful’, the failure to report critical information became striking in the reviewed studies. These shortcomings are serious, because it can often happen that values of global fit statistics (e.g. CFI, TLI) look reasonable, while evidence of grossly poor fit is clear in the residuals. Without reporting such critical information, a study may claim or endorse a structural model seemingly fitting the data whilst lacking reliability and validity.

This study has several implications and contributions. First, consolidating and expanding earlier seminal work on best SEM practice, it devises a scheme for evaluating the quality of SEM application across the stages of model formulation, specification, estimation and evaluation. In comparison with previous work, which often enumerated the issues and problems of utilizing SEM all at once, this sequential approach can enable our readers to appreciate the essential practices step by step. Second, it provides concrete examples taken from existing high-quality publications to illustrate the ways to achieve those recommended analysing and reporting standards, with detailed explanations on the necessity and importance of each requirement. Future SEM studies can take the evaluation scheme together with the suggestions provided below as a practical guideline, and journal editors and reviewers can also adopt the scheme to create an objective assessment about the status quo of utilizing SEM in a particular study. Finally, it evaluates the status of applying SEM in various realms of organizational and management research, and reveals a pressing need for organizational and management journals to establish more explicit and standardized ways of conducting and reporting SEM studies.

There are two critical limitations of this work. One limitation is that we did not explicitly examine the reasons that some published studies failed to demonstrate that they followed best-practice standards. It is possible that in the reviewed literature researchers used SEM without sufficient knowledge of what the technique is for and what they should (not) do in a particular instance. It is also possible that a study was unable to report its every step. To address this limitation, we will investigate the reasons behind such ‘failure’ in future work. We will survey and interview researchers, students, reviewers and editors, in order to explore, for instance, whether studies not providing statistical power information are more likely to have certain features, whether those observed problematic practices are more prevalent in particular types of domains, whether studies engaged in non-desirable practices are more likely
to report ‘successful’ models, whether studies reporting ‘successful’ models are more likely to get published and so forth. Nevertheless, our recommendation remains that there is a pressing need to establish a more systemic and standardized analytical and reporting system of SEM.

A second critical limitation is that we did not code some other analytical issues that are frequently mentioned as crucial, such as the test of common method variance (Podsakoff *et al.*, 2003), specification of alternative or equivalent models (Henley, Shook and Peterson, 2006), non-independence of nested cases in multilevel data (Appelbaum *et al.*, 2018), measurement invariance in cross-group analysis (Kite, Jorgensen and Chen, 2018) and so on. They were not included due to the fact that these issues were not applicable to all SEM studies. In other words, our scheme intends to cover the necessary conditions constituting a good SEM application, and thus does not claim to be sufficiently comprehensive. To address this limitation, we will expand our investigation in the future by examining the status of satisfying these standards in relevant studies using a wider timeframe.

**Implications and suggestions for future work**

We end the review with a brief case study based on lessons learned from the results of this investigation. The example concerns mediation analysis, for which there are thousands of empirical studies in management, psychology, education and other disciplines (i.e. this is a ‘popular’ topic). The basic rationale is that changes in one variable cause changes in another (i.e. the mediator), which in turn leads to changes in an outcome (Little, 2013). There are many good reasons to estimate mediation effects using SEM compared with traditional statistical methods, such as multiple regression. These advantages include: (a) generally lower standard errors due to the simultaneous estimation of all model parameters in SEM compared with the separate application of regression techniques to each dependent variable; (b) the capability to explicitly model measurement errors in SEM (regression assumes perfect reliability for all predictors); (c) the option to analyse multiple indicators of the same construct in a latent variable model for mediation; and (d) the flexibility to add constructs to an extant nomological network that involves trivariate mediation (Iacobucci, Saldanha and Deng, 2007) for computer simulation results about these points. However, there are problems with many, if not most, published mediation studies that raise doubts about whether the results have any meaningful interpretation as ‘mediation’ (Kline, 2015; Pek and Hoyle, 2016). These problems include the failure to state all assumptions in the analysis, the misuse of statistical significance tests, lack of complete reporting about model fit and the failure to appreciate the critical role of research design in mediation analysis, among other shortcomings. Some of these deficits correspond directly to criteria applied in this study (e.g. criteria 1, 4, 8 and 9 in Table 1). If SEM is poorly applied, potential benefits of using it in studies of mediation will be nullified.

To sum up, there are several practical suggestions to help our readers prepare and conduct future SEM studies with enhanced transparency and replicability.

**Prepare a rational research and analytic plan**

This includes the considerations about: (a) why SEM is an appropriate method given the research aims; (b) the rationale for the sample size, for example, demonstrating that power is adequate if significance testing plays a critical role in the analysis; and (c) the justification for directional specifications in the initial model, namely, why we assume that X causes Y instead of the reverse.

**Document re-specification of the initial model**

That is, explain the rationale for changes to the original model and outline the bases for doing so. Model changes should more reflect theories and results from prior empirical studies in the same area than results from significance testing in the present sample. It is poor practice to drop paths with coefficients that are not significant, just as it is to add paths that would reduce the model chi-square by the greatest amount, if there is no theoretical justification for these changes (Kline, 2016; Loehlin, 2004).

**Replicate the analysis**

This would represent a type of nirvana for SEM: replication is extraordinarily rare in the SEM...
Do not retain a model at any cost

Models that are re-specified solely according to empirical considerations, such as modification indexes, are unlikely to be replicable. It would be better in this case to (a) retain no model, (b) consider why and how predictions based on theory are wrong and (c) offer guidance about how to move forward in future studies. In such circumstances, a permutation test may be useful as a technique for coping with situations where the assumptions of multivariate normality or measurement (in)variance are violated (Anderson and Braak, 2003; Jorgensen, 2017; Jorgensen et al., 2018). It can also be used to determine whether models other than the researchers’ targets but with even better fit to the data might exist and are worthy of further examination (Anderson and Braak, 2003). Briefly, permutation tests examine the likelihood of obtaining a certain outcome, if the data for the dependent variable are randomly distributed across the levels of the independent variables (Hayes, 1996). The $p$-value in this circumstance refers to the proportion in the permuted samples that have a parameter value equal to or higher than the one obtained from the real sample (Chin and Dibbern, 2010). Some computer tools for SEM, such as AMOS (Arbuckle, 2014), support the permutation of models by considering fit in large numbers of model variations (Chin and Dibbern, 2010). Therefore, even if a model drawn from the real sample may not have absolute satisfactory goodness-of-fit indices or parameter values, comparatively, the model could still be considered a nearest approximation of the data (Burnham and Anderson, 2013), if its targeted indicators are greater than (or, in some cases, lower than) most of those generated by other permuted models (Chin and Dibbern, 2010).

In sum, SEM should be used with careful plans and rigorous strategies. Currently, the top-level SEM studies in organizational and management science still suffer from deficiencies in demonstrating that they adhered to some of the core principles, assumptions and recommended procedures of this powerful analytical tool. More efforts are needed to enhance the clarity, transparency and completeness of SEM studies in organizational and management research.

Conflict of interest

The authors declare that they have no conflict of interest.

References


*Boh, W. F. and S.-S. Wong (2015). ‘Managers versus co-workers as referents: comparing social influence effects on within- and
outside-subsidiary knowledge sharing’, Organisational Behavior and Human Decision Processes, 126, pp. 1–17.


Henley, A. B., C. L. Shook and M. Peterson (2006). ‘The presence of equivalent models in strategic management research

© 2020 The Authors. British Journal of Management published by John Wiley & Sons Ltd on behalf of British Academy of Management.
using structural equation modeling’, Organizational Research Methods, 9, pp. 516–535.


Mary F. Zhang is a Senior Research Associate in the School for Policy Studies at the University of Bristol. Her main research and publication interests are in the areas of poverty and social exclusion, gender equality and child rights, development and well-being. She adopts qualitative and quantitative approaches in her research, among which structural equation modelling is frequently used.

Jeremy F. Dawson is Professor of Health Management, working jointly between the Institute of Work Psychology and the School of Health and Related Research at the University of Sheffield. He is a statistician by training and has worked in the fields of work psychology, team working, human resource management, evaluation of interventions to improve staff well-being, and diversity and discrimination in the workplace.

Rex B. Kline is Professor of Psychology at Concordia University in Montreal. Much of his research concerns measurement, psychological assessment and child psychology. He was trained as a child clinical psychologist and a methodologist. His current areas of work include structural equation modelling, psychometrics and reform of methods for statistical inference in the social and behavioural science.