

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 ¹ and Rex B. Kline ²

School for Policy Studies, University of Bristol, 10 Woodland Road, Bristol BS8 1SZ, UK, ¹Institute of Work Psychology, Sheffield University Management School, Conduit Road, Sheffield S10 1FL, UK, and ²Department 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

We would like to thank the anonymous reviewers and BJM associate editor Marc Goergen for their constructive comments of this paper. Numerous SEM users and colleagues also provided helpful feedback at the formative stages of this work for which we are grateful.

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 giv-

ing 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. Two techniques, both known as confirmatory composite analysis (CCA), provide other alternatives. The first of these is the measurement model assessment step of a partial least squares structural equation modelling (PLS-SEM) technique that aims to analyse formative or reflective measurement models, where factors in reflective models are viewed as empirical proxies to corresponding hypothetical constructs instead of as essentially identical to those concepts (Rigdon, 2012; Hair 2020), although its ability to do so is disputed by Schuberth (2020). A second method, also called CCA, was developed earlier by Henseler and colleagues (Henseler *et al.*, 2014; Schuberth, Henseler and Disjkstra, 2018). This method, which can evaluate composite models, is closer to CFA as conducted in CB-SEM, with good sensitivity to model misspecification (Schuberth *et al.*, 2018). 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). 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.

[Correction added on August 26, 2020 after online publication on June 25, 2020: The paragraph starting with “It is worth.....otherwise indicated” was modified.]

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

Coding criteria	Example 1	Example 2
(1) Does the study give specific reasons or justifications for using SEM or a specific form of it?	'... 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).	'... 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).
(2.1) Does the study specify the overall structural equation model(s) to be tested?	'The model we advance is shown in Figure 1' (Kirkman <i>et al.</i> , 2011, p. 1236).	'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).
(2.2) Does the study specify the relations between the variables or constructs?	All the studies reviewed met this criterion by specifying concrete research hypotheses to be tested.	
(3.1) Does the study justify the sample size?	'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 <i>et al.</i> , 2014, p. 48).	'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 <i>et al.</i> , 2015, p. 1646).
(3.2) Does the study test statistical power?	'The power of our analyses was found to be 1.0 for a test of close fit...' (McCarthy, Trougakos and Cheng, 2016, p. 284).	'To ensure that the data permitted a valid testing of our hypotheses, we conducted <i>a priori</i> 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).
(4) Are distributional assumptions of the method(s) respected in the data?	'... we ensured that the assumptions of normality... were met' (de Stobbeleir, Ashford and Buyens, 2011, p. 821).	'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).
(5.1) Does the study report incomplete data? (Note: respondents may provide incomplete data, which, however, is different from non-responses.)	'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).	'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).
(5.2) Does the study clearly discuss the ways of dealing with missing data, if presented?	'... <i>Mplus</i> uses full information maximum likelihood estimation that allows for missing data under the missing at random assumption' (Gielnik, Klemann and Consultancy, 2015, p. 1017).	'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 <i>et al.</i> , 2013, p. 746).
(5.3) Does the study deal with missing data in an appropriate way, if presented?		

(Continued)

Table 1. Continued

Coding criteria	Example 1	Example 2
(6) Does the study calculate score reliability coefficients in its own sample(s)?	'The mean of these ratings was then calculated to create a reliable ($\alpha = 0.87$) measure...' (Mortensen, 2014, p. 921).	'Reliability estimates for all measures exceeded 0.70' (Ragins <i>et al.</i> , 2012, p. 766).
(7) Does the study distinguish the measurement model from the structural model?	'The first step in analyzing our data was examining the adequacy of our measurement model' (Colquitt and Rodell, 2011, p. 1193).	'Prior to testing the hypothesized structural model, we tested to see if the measurement model had good fit' (Mayer <i>et al.</i> , 2012, p. 159).
(8.1) Does the study report RMSEA and its 90% or 95% CIs?	'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).	'The fit statistics for this model indicated acceptable fit, $\chi^2(365) = 557.56, p < 0.01$, RMSEA = 0.06 (90% CI [0.05, 0.07]), CFI = 0.91, TLI = 0.90, and SRMR = 0.07' (Cullen <i>et al.</i> , 2014, pp. 1770–1771).
(8.2) Does the study report SRMR?		
(8.3) Does the study report CFI or TLI?		
(8.4) Does the study report the result of chi-square test for the model?		
(9) Does the study report residuals, that is, quantitative measures of model–data discrepancy at the level of pairs of observed variables?	'Values shown are unstandardized parameter estimates, with standard errors in parentheses' (Ferguson <i>et al.</i> , 2016, p. 528).	'The middle panel of Table 3 presents factor loadings and error variances...' (Bagozzi <i>et al.</i> , 2012, p. 71).

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.

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 favour 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).¹

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 resid-

uals 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(l)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.

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

Year	AMJ	ASQ	BJM	JAP	JoM	JoMS	MS	OS1	OS2	OBHDP	PP	SMJ	Total	%
2011	3	0	1	6	3	3	1	2	0	0	7	0	26	18.1
2012	5	0	2	3	6	2	0	1	0	0	2	1	22	15.3
2013	4	0	4	8	3	2	0	1	1	1	4	0	28	19.4
2014	3	1	3	5	6	1	0	2	0	1	3	1	26	18.1
2015	2	0	2	4	2	3	0	2	1	2	2	2	22	15.3
2016	4	0	2	7	3	1	0	0	0	0	3	0	20	13.9
Total	21	1	14	33	23	12	1	8	2	4	21	4	144	
%	14.6	0.7	9.7	22.9	16.0	8.3	0.7	5.6	1.4	2.8	14.6	2.8		

Notes: AMJ = *Academy of Management Journal*; ASQ = *Administrative Science Quarterly*; BJM = *British Journal of Management*; JAP = *Journal of Applied Psychology*; JoM = *Journal of Management*; JoMS = *Journal of Management Studies*; MS = *Management Science*; OS1 = *Organization Science*; OS2 = *Organization Studies*; OBHDP = *Organizational Behavior and Human Decision Processes*; PP = *Personnel Psychology*; SMJ = *Strategic Management Journal*.

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.² 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.³ 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 ($\kappa = 0.93$, $p < 0.001$; 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 models) have been met well (i.e. over 90% of reviewed publications met these standards), fol-

²The four papers include: Diestel and Schmidt (2011), which applied latent moderated SEM with non-normally distributed outcomes; Koppman (2016), which used SEM to examine interview data generated from 54 participants; Krasikova and LeBreton (2012), which used SEM to examine simulated data; and Maclean, Harvey and Kling (2014), which adopted SEM to test the issue of endogeneity bias.

³One paper was randomly selected from each journal ($N = 12$) and five additional papers were randomly selected from the remaining sample. If a journal only contained one SEM study, that article was selected.

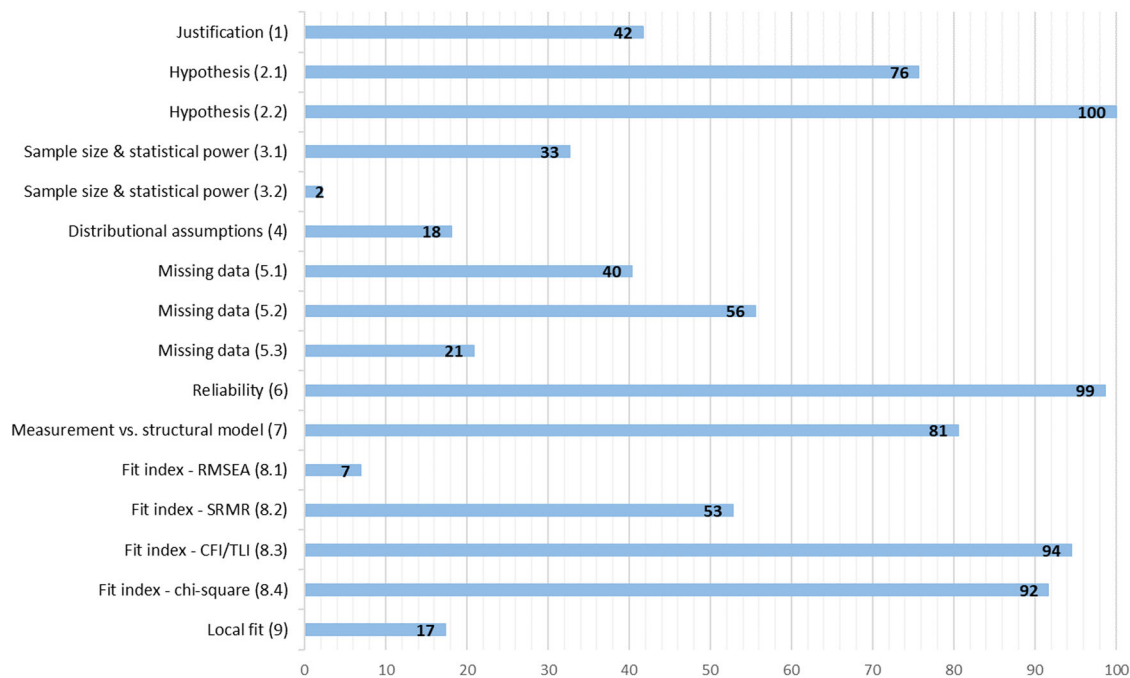


Figure 1. Percentage (%) of publications ($N = 144$) satisfying each evaluation criterion [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1111/467-8851.12415)]

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](https://onlinelibrary.wiley.com/terms-and-conditions)]

lowed by criteria 7 (testing and distinguishing the measurement vs. structural model), 2.1 (hypothesizing overall model), 5.2 (discussing the ways of dealing with missing data) and 8.2 (reporting SRMR), which had a middling degree of consideration (i.e. over 50% of reviewed publications met these requirements), while the remaining criteria received a low level of attention (i.e. only about 40% or less of reviewed studies met these criteria).

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,

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 hap-

pen 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 con-

structs 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

literature, due in part to the requirement for large samples in SEM, but also to our collective failure in the behavioural sciences to properly value replication (Porte, 2012). External replication – where new data are collected in different settings by other researchers – is the strongest form, but internal replication would do in a pinch. In very large samples, the same model could be evaluated over random subsets of the original sample – such as in cross-validation, where the whole sample is split at random into two halves, which may be called the validation set and the test set, respectively. The failure to replicate SEM results across random splits of the original sample would indicate a serious problem, and yet the opposite outcome – stability of the solution – is actually weak evidence for replication, because there is a single sample (i.e. it is not external replication over independent samples). In any event, evidence for replication signals that the original results are not just a statistical fluke.

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

- Allison, P. D. (2003). 'Missing data techniques for structural equation modeling', *Journal of Abnormal Psychology*, **112**, pp. 545–557.
- Anderson, J. C. and D. W. Gerbing (1988). 'Structural equation modeling in practice: a review and recommended two-step approach', *Psychological Bulletin*, **103**, pp. 411–423.
- Anderson, M. and C. T. Braak (2003). 'Permutation tests for multi-factorial analysis of variance', *Journal of Statistical Computation and Simulation*, **73**, pp. 85–113.
- Appelbaum, M., H. Cooper, R. B. Kline, E. Mayo-Wilson, A. M. Nezu and S. M. Rao (2018). 'Journal article reporting standards for quantitative research in psychology: the APA Publications and Communications Board task force report', *American Psychologist*, **73**, pp. 3–25.
- Arbuckle, J. L. (2014). Amos (v23.0) [Computer Program]. Chicago, IL: IBM SPSS. <https://www.ibm.com/support/pages/what-correct-format-citing-amos#:~:text=Here%20is%20the%20citation%20for,Chicago%3A%20IBM%20SPSS>.
- *Baer, M. D., R. K. Dhensa-Kahlon, J. A. Colquitt, J. B. Rodell, R. Outlaw and D. M. Long (2015). 'Uneasy lies the head that bears the trust: the effects of feeling trusted on emotional exhaustion', *Academy of Management Journal*, **58**, pp. 1637–1657.
- *Bagozzi, R. P., M. Bergami, G. L. Marzocchi and G. Morandini (2012). 'Customer–organisation relationships: development and test of a theory of extended identities', *Journal of Applied Psychology*, **97**, pp. 63–76.
- *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.
- Bollen, K. A. and W. R. Davis (2009). 'Causal indicator models: identification, estimation, and testing', *Structural Equation Modeling*, **16**, pp. 498–522.
- Brannick, M. T. (1995). 'Critical comments on applying covariance structure modeling', *Journal of Organizational Behavior*, **16**, pp. 201–213.
- Brown, R. L. (1994). 'Efficacy of the indirect approach for estimating structural equation models with missing data: a comparison of five methods', *Structural Equation Modeling*, **1**, pp. 287–316.
- Burnham, K. P. and D. R. Anderson (2013). *Model Selection and Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag. https://play.google.com/store/books/details?id=W63hBwAAQBAJ&pcampaignid=books_web_aboutlink.
- Cheung, M. W.-L. (2009). 'Constructing approximate confidence intervals for parameters with structural equation models', *Structural Equation Modeling*, **16**, pp. 267–294.
- Cheung, M. W. L. (2015). *Meta-Analysis: A Structural Equation Modelling Approach*. Chichester: Wiley. https://play.google.com/store/books/details/Mike_W_L_Cheung_Meta-Analysis?id=VHFuCAAQBAJ.
- Cheung, G. W. and R. S. Lau (2008). 'Testing mediation and suppression effects of latent variables: bootstrapping with structural equation models', *Organisational Research Methods*, **11**, pp. 296–325.
- Cheung, G. W. and R. B. Rensvold (2001). 'The effects of model parsimony and sampling error on the fit of structural equation models', *Organisational Research Methods*, **4**, pp. 236–264.
- Chin, W. W. and J. Dibbern (2010). 'A permutation based procedure for multi-group PLS analysis: results of tests of differences on simulated data and a cross cultural analysis of the sourcing of information system services between Germany and the USA'. In V. E. Vinzi, W. W. Chin, J. Henseler and H. Wang (eds), *Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields*, pp. 171–193. Berlin: Springer-Verlag.
- Cohen, J. (1960). 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement*, **20**, pp. 37–46.
- Cole, D. A. and K. J. Preacher (2014). 'Manifest variable path analysis: potentially serious and misleading consequences due to uncorrected measurement error', *Psychological Methods*, **19**, pp. 300–315.
- *Colquitt, J. A. and J. B. Rodell (2011). 'Justice, trust, and trustworthiness: a longitudinal analysis integrating three theoretical perspectives', *Academy of Management Journal*, **54**, pp. 1183–1206.
- Conlon, D., F. P. Morgeson, G. McNamara, R. M. Wiseman and P. Skilton (2006). 'Examining the impact and role of special issue and regular journal articles in the field of management', *Academy of Management Journal*, **49**, pp. 857–872.
- Cudeck, R., S. du Toit and D. Sörbom (2001). *Structural Equation Modelling: Present and Future: A festschrift in honor of Karl Jöreskog*. Lincolnwood, IL: Scientific Software International. https://books.google.co.uk/books/about/Structural_Equation_Modeling.html?id=AMLN-oKmC_IC&redir_esc=y.
- *Cullen, K. L., J. Fan and C. Liu (2014). 'Employee popularity mediates the relationship between political skill and workplace interpersonal mistreatment', *Journal of Management*, **40**, pp. 1760–1778.
- Curran, P. J., S. G. West and J. F. Finch (1996). 'The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis', *Psychological Methods*, **1**, pp. 16–29.
- de Stobbeleir, K. E. M., S. J. Ashford and D. Buyens (2011). 'Self-regulation of creativity at work: the role of feedback-seeking behavior in creative performance', *Academy of Management Journal*, **54**, pp. 811–831.
- Diestel, S. and K.-H. Schmidt (2011). 'Costs of simultaneous coping with emotional dissonance and self-control demands at work: results from two German samples', *Journal of Applied Psychology*, **96**, pp. 643–653.
- Fabrigar, L. R., D. T. Wegener, R. C. MacCallum and E. J. Strahan (1999). 'Evaluating the use of exploratory factor analysis in psychological research', *Psychological Methods*, **4**, pp. 272–299.
- *Ferguson, M., D. Carlson, W. Boswell, D. Whitten, M. M. Butts and K. M. Kacmar (2016). 'Tethered to work: a family systems approach linking mobile device use to turnover intentions', *Journal of Applied Psychology*, **101**, pp. 520–534.
- Floyd, K. (2014). 'Taking stock of research practices: a call for self-reflection', *Communication Monographs*, **81**, pp. 1–3.
- *Foss, N. J., K. Laursen and T. Pedersen (2011). 'Linking customer interaction and innovation: the mediating role of new organisational practices', *Organisation Science*, **22**, pp. 980–999.
- Gelman, A. (2018). 'The failure of null hypothesis significance testing when studying incremental changes and what to do about it', *Personality and Social Psychology Bulletin*, **44**, pp. 16–23.
- Gielnik, M. M., D. K. Klemann and K. Consultancy (2015). 'I put in effort, therefore I am passionate: investigating the path from effort to passion in entrepreneurship', *Academy of Management Journal*, **58**, pp. 1012–1031.
- Goodboy, A. K. and R. B. Kline (2017). 'Statistical and practical concerns with published communication research featuring structural equation modeling', *Communication Research Reports*, **34**, pp. 68–77.
- Grewal, R., J. A. Cote and H. Baumgartner (2004). 'Multicollinearity and measurement error in structural equation models: implications for theory testing', *Marketing Science*, **23**, pp. 519–529.
- Hair, J. F., M. C. Howard and C. Nitzl (2020). 'Assessing measurement model quality in PLS-SEM using confirmatory composite analysis', *Journal of Business Research*, **109**, pp. 101–110.
- Hair, J. F., C. M. Ringle and M. Sarstedt (2012). 'Partial least squares: the better approach to structural equation modeling?', *Long Range Planning*, **45**, pp. 312–319.
- Haller, H. and S. Krauss (2002). 'Misinterpretations of significance: a problem students share with their teachers?', *Methods of Psychological Research Online*, **7**, pp. 1–17.
- Hayduk, L., G. Cummings, K. Boadu, H. Pazderka-Robinson and S. Boulianne (2007). 'Testing! testing! one, two, three—testing the theory in structural equation models!', *Personality and Individual Differences*, **42**, pp. 841–850.
- Hayes, A. F. (1996). 'Permutation test is not distribution-free: testing $H_0: \rho = 0$ ', *Psychological Methods*, **1**, pp. 184–198.
- Henley, A. B., C. L. Shook and M. Peterson (2006). 'The presence of equivalent models in strategic management research

- using structural equation modeling', *Organizational Research Methods*, **9**, pp. 516–535.
- Henseler, J., T. K. Dijkstra, M. Sarstedt, C. M. Ringle, A. Diamantopoulos, D. W. Straub, D. J. Ketchen, J. F. Hair, G. T. M. Hult and R. J. Calantone (2014). 'Common beliefs and reality about PLS: Comments on Rönkkö and Evermann (2013)', *Organizational Research Methods*, **17**, pp. 182–209.
- Hermida, R., J. N. Luchman, V. Nicolaides and C. Wilcox (2015). 'The issue of statistical power for overall model fit in evaluating structural equation models', *Computational Methods in Social Sciences*, **25**, pp. 25–42.
- Hoch, J. E. and S. W. J. Kozlowski (2014). 'Leading virtual teams: hierarchical leadership, structural supports, and shared team leadership', *Journal of Applied Psychology*, **99**, pp. 390–403.
- Hoyle, R. H. and J. C. Iserwood (2013). 'Reporting results from structural equation modeling analyses in archives of scientific psychology', *Archives of Scientific Psychology*, **1**, pp. 14–22.
- Hoyle, R. H. and G. T. Smith (1994). 'Formulating clinical research hypotheses as structural equation models: a conceptual overview', *Journal of Consulting and Clinical Psychology*, **62**, pp. 429–440.
- Hu, J. and R. C. Liden (2015). 'Making a difference in the teamwork: linking team prosocial motivation to team processes and effectiveness', *Academy of Management Journal*, **58**, pp. 1102–1127.
- Iacobucci, D., N. Saldanha and X. Deng (2007). 'A meditation on mediation: evidence that structural equations models perform better than regressions', *Journal of Consumer Psychology*, **17**, pp. 139–153.
- Ioannidis, J. P. A. (2005). 'Why most published research findings are false', *PLoS Medicine*, **2**, pp. 696–701.
- Jackson, D. L. (2003). 'Revisiting sample size and number of parameter estimates: some support for the N:q hypothesis', *Structural Equation Modeling*, **10**, pp. 128–141.
- Jak, S. (2015). *Meta-Analytic Structural Equation Modelling*. Cham: Springer. <https://link.springer.com/book/10.1007/978-3-319-27174-3>.
- John, L. K., G. Loewenstein, D. Prelec, L. K. John, G. Loewenstein and D. Prelec (2012). 'Measuring the prevalence of questionable research practices with incentives for truth telling', *Psychological Science*, **23**, pp. 524–532.
- Jorgensen, T. D. (2017). 'Applying permutation tests and multivariate modification indices to configurally invariant models that need respecification', *Frontiers in Psychology*, **8**, pp. 1–9.
- Jorgensen, T. D., B. A. Kite, P. Y. Chen and S. D. Short (2018). 'Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis', *Psychological Methods*, **23**, pp. 708–728.
- *Kaltiainen, J., J. Lipponen and B. C. Holtz (2016). 'Dynamic interplay between merger process justice and cognitive trust in top management: a longitudinal study', *Journal of Applied Psychology*, **102**, pp. 636–647.
- Kaplan, D. (2009). *Structural Equation Modeling: Foundations and Extensions*, 2nd edn. Thousand Oaks, CA: Sage. <https://methods.sagepub.com/book/structural-equation-modeling>.
- Kerr, N. L. (1998). 'HARKing: hypothesizing after the results are known', *Personality and Social Psychology Review*, **2**, pp. 196–217.
- Kim, K. H. (2009). 'The relation among fit indexes, power, and sample size in structural equation modelling', *Structural Equation Modeling*, **12**, pp. 368–390.
- *Kim, T. G., S. Hornung and D. M. Rousseau (2011). 'Change-supportive employee behavior: antecedents and the moderating role of time', *Journal of Management*, **37**, pp. 1664–1693.
- *Kirkman, B. L., J. E. Mathieu, J. L. Cordery, B. Rosen and M. Kukenberger (2011). 'Managing a new collaborative entity in business organisations: understanding organisational communities of practice effectiveness', *Journal of Applied Psychology*, **96**, pp. 1234–1245.
- Kite, B. A., T. D. Jorgensen and P.-Y. Chen (2018). 'Random permutation testing applied to measurement invariance testing with ordered-categorical indicators', *Structural Equation Modeling*, **25**, pp. 573–587.
- Kline, R. B. (2015). 'The mediation myth', *Basic and Applied Social Psychology*, **37**, pp. 202–213.
- Kline, R. B. (2016). *Principles and Practice of Structural Equation Modelling*, 4th edn. New York: Guilford Press. <https://www.guilford.com/books/Principles-and-Practice-of-Structural-Equation-Modeling/Rex-Kline/9781462523344>.
- Koppman, S. (2016). 'Different like me', *Administrative Science Quarterly*, **61**(2), pp. 291–331. <https://doi.org/10.1177/0001839215616840>.
- Krasikova, D. V. and J. M. LeBreton (2012). 'Just the two of us: misalignment of theory and methods in examining dyadic phenomena', *Journal of Applied Psychology*, **97**, pp. 739–757.
- Lambert, N. M. (1991). 'The crisis in measurement literacy in psychology and education', *Educational Psychologist*, **26**, pp. 23–35.
- Lanaj, K., J. R. Hollenbeck, D. R. Ilgen, C. M. Barnes and S. J. Harmon (2013). 'The double-edged sword of decentralized planning in multitask systems', *Academy of Management Journal*, **56**, pp. 735–757.
- Landis, R. S., D. J. Beal and P. E. Tesluk (2000). 'A comparison of approaches to forming composite measures in structural equation models', *Organizational Research Methods*, **3**, pp. 186–207.
- Larsen, R. (2011). 'Missing data imputation versus full information maximum likelihood with second-level dependencies', *Structural Equation Modeling*, **18**, pp. 649–662.
- Lee, S.-Y. (2007). *Structural Equation Modelling: A Bayesian Approach*. Chichester: Wiley. <https://www.wiley.com/engb/Structural+Equation+Modeling%3A+A+Bayesian+Approach-p-9780470024232>.
- *Lievens, F. and F. Patterson (2011). 'The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection', *Journal of Applied Psychology*, **96**, pp. 927–940.
- Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York: Guilford Press. <https://www.guilford.com/books/Longitudinal-Structural-Equation-Modeling/Todd-Little/9781462510160>.
- Loehlin, J. C. (2004). *Latent Variable Models: An Introduction to Factor, Path and Structural Equation Analysis*, 4th edn. Mahwah, NJ: Erlbaum. <https://psycnet.apa.org/record/2004-00126-000>.
- MacCallum, R. C. and J. T. Austin (2000). 'Applications of structural equation modeling in psychological research', *Annual Review of Psychology*, **51**, pp. 201–226.
- Maclean, M., C. Harvey and G. Kling (2014). 'Pathways to power: class, hyper-agency and the French corporate elite', *Organisation Studies*, **35**, pp. 825–855.

- *Mayer, D. M., K. Aquino, R. L. Greenbaum and M. Kuenzi (2012). 'Who displays ethical leadership, and why does it matter? An examination of antecedents and consequences of ethical leadership', *Academy of Management Journal*, **55**, pp. 151–171.
- *McCarthy, J. M., J. P. Trougakos and B. H. Cheng (2016). 'Are anxious workers less productive workers? It depends on the quality of social exchange', *Journal of Applied Psychology*, **101**, pp. 279–291.
- McDonald, R. P. and M.-H. R. Ho (2002). 'Principles and practice in reporting structural equation analyses', *Psychological Methods*, **7**, pp. 64–82.
- McHugh, M. L. (2012). 'Interrater reliability: the kappa statistic', *Biochemia Medica*, **22**, pp. 276–282.
- McIntosh, C. N., J. R. Edwards and J. Antonakis (2014). 'Reflections on partial least squares path modeling', *Organizational Research Methods*, **17**, pp. 210–251.
- Millsap, R. E. (2007). 'Structural equation modelling made difficult', *Personality and Individual Differences*, **42**, pp. 875–881.
- Mitchell, R. J. (1992). 'Testing evolutionary and ecological hypotheses using path analysis and structural equation modelling', *Functional Ecology*, **6**, pp. 123–129.
- Molina-Azorin, J. F. (2012). 'Mixed methods research in strategic management: impact and applications', *Organizational Research Methods*, **15**, pp. 33–56.
- *Mortensen, M. (2014). 'Constructing the team: the antecedents and effects of membership model divergence', *Organisation Science*, **25**, pp. 909–931.
- Mueller, R. O. (1997). 'Structural equation modeling: back to basics', *Structural Equation Modeling*, **4**, pp. 353–369.
- Mueller, R. O. and G. R. Hancock (2008). 'Best practices in structural equation modeling'. In J. W. Osborne (ed.), *Best Practices in Quantitative Methods*, pp. 488–508. Thousand Oaks, CA: Sage.
- Mulaik, S. A., L. R. James, J. van Alstine, N. Bennett, S. Lind and C. D. Stilwell (1989). 'Evaluation of goodness-of-fit indices for structural equation models', *Psychological Bulletin*, **105**, pp. 430–445.
- *Nifadkar, S., A. S. Tsui and B. E. Ashforth (2012). 'The way you make me feel and behave: supervisor-triggered newcomer affect and approach-avoidance behavior', *Academy of Management Journal*, **55**, pp. 1146–1168.
- Nunkoo, R., H. Ramkissoon and D. Gursoy (2013). 'Use of structural equation modeling in tourism research: past, present, and future', *Journal of Travel Research*, **52**, pp. 759–771.
- *Ou, A. Y., A. S. Tsui, A. J. Kinicki, D. A. Waldman, Z. Xiao and L. J. Song (2014). 'Humble chief executive officers' connections to top management team integration and middle managers' responses', *Administrative Science Quarterly*, **59**, pp. 34–72.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd edn. New York: Cambridge University Press. <https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B>.
- Pek, J. and R. H. Hoyle (2016). 'On the (in)validity of tests of simple mediation: threats and solutions', *Social and Personality Psychology Compass*, **10**, pp. 150–163.
- Ployhart, R. E., C. H. Van Iddekinge and W. I. MacKenzie (2011). 'Acquiring and developing human capital in service contexts: the interconnectedness of human capital resources', *Academy of Management Journal*, **54**, pp. 353–368.
- Podsakoff, P. M., S. B. MacKenzie, J. Y. Lee and N. P. Podsakoff (2003). 'Common method biases in behavioral research: a critical review of the literature and recommended remedies', *Journal of Applied Psychology*, **88**, pp. 879–903.
- Porte, G. (ed.) (2012). *Replication Research in Applied Linguistics*. New York: Cambridge University Press. <https://www.worldcat.org/title/replication-research-in-applied-linguistics/oclc/762135194>.
- *Ragins, B. R., J. A. Gonzalez, K. Ehrhardt and R. Singh (2012). 'Crossing the threshold: the spillover of community racial diversity and diversity climate to the workplace', *Personnel Psychology*, **65**, pp. 755–787.
- Raines-Eudy, R. (2000). 'Using structural equation modeling to test for differential reliability and validity: an empirical demonstration', *Structural Equation Modeling*, **7**, pp. 124–141.
- Rigdon, E. E. (2012). 'Rethinking partial least squares path modeling: in praise of simple methods', *Long Range Planning*, **45**, pp. 341–358.
- Rigdon, E. E. (2014). 'Rethinking partial least squares path modeling: breaking chains and forging ahead', *Long Range Planning*, **47**, pp. 161–167.
- Rubin, D. B. (1976). 'Inference and missing data', *Biometrika*, **63**, pp. 581–532.
- Schuberth, F. (2020) 'Confirmatory composite analysis using partial least squares: setting the record straight', *Review of Managerial Science*. <https://doi.org/10.3389/fpsyg.2018.02541>.
- Schuberth, F., J. Henseler and T. K. Dijkstra (2018) 'Confirmatory Composite Analysis' *Frontiers in Psychology*, **9**, p. 2541.
- Shah, R. and S. M. Goldstein (2006). 'Use of structural equation modeling in operations management research: looking back and forward', *Journal of Operations Management*, **24**, pp. 148–169.
- Simmons, J. P., L. D. Nelson and U. Simonsohn (2011). 'False-positive psychology', *Psychological Science*, **22**, pp. 1359–1366.
- *Stanhope, D. S., S. B. Pond and E. A. Surface (2013). 'Core self-evaluations and training effectiveness: prediction through motivational intervening mechanisms', *Journal of Applied Psychology*, **98**, pp. 820–831.
- Szucs, D. and J. P. A. Ioannidis (2017). 'When null hypothesis significance testing is unsuitable for research: a reassessment', *Frontiers in Human Neuroscience*, **11**, pp. 1–21.
- *Wei, L.-Q. and L. Wu (2013). 'What a diverse top management team means: testing an integrated model', *Journal of Management Studies*, **50**, pp. 389–412.
- Williams, L. J., M. B. Gavin and M. L. Williams (1996). 'Measurement and nonmeasurement processes with negative affectivity and employee attitudes', *Journal of Applied Psychology*, **81**, pp. 88–101.
- Williams, L. J., R. J. Vandenberg and J. R. Edwards (2009). '12 structural equation modelling in management research: a guide for improved analysis', *The Academy of Management Annals*, **3**, pp. 543–604.
- [Correction added on August 26, 2020 after online publication on June 25, 2020: The references list was updated to reflect the prior text additions.]

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.