Refereeing articles including SEM – what should referees look for?

Chris Ryan
China-New Zealand Tourism Research Unit,
University of Waikato Management School, Hamilton, New Zealand

Received 13 March 2020
Revised 20 March 2020
Accepted 20 March 2020

Abstract

Purpose – After expressing an initial disquiet about the nature of many studies that are published using structural equation modelling (SEM), a rationale for using the technique is provided. Given the advantages provided by the technique, the differences between covariance-based and partial least squares techniques are briefly described. The argument progresses by indicating assumptions behind the techniques and what it is that referees require before being able to properly referee the paper. Some issues are fundamental to survey-based materials and include the requirement to distinguish between importance and discriminatory power, and the over-dependency on cross-sectional analysis when making claims of generalisation. Other issues of scale creation and sample size are touched upon. This paper finishes by suggesting a checklist for referees who are asked to review papers using SEM.

Keywords SEM, Statistical methods, Structural equation modelling, Refereeing papers, Statistical practice

Paper type General review

Introduction

The writing of this paper has been prompted by my experiences as a long-term editor of a journal (*Tourism Management*) for approximately 25 years and being a referee for even more years for a number of other journals. The immediate catalyst for the paper was, however, my coming across a special issue of a journal that referred to the use of structural equation modelling (SEM) with applications to tourism. While, in my view, there were good papers in the issue, for the majority of them, I wondered how referees could really have assessed and judged the papers given the paucity of information in them. This simply reinforced what I have found on a number of occasions, and as Pearl and MacKenzie (2018) (amongst others) have commented, namely, the process of much of the reporting of SEM papers has become mechanistic, not transparent and at times exercises in data manipulation as authors seek to achieve the normally required indices of good fit. Therefore, this paper begins with a review of why SEM has developed, and it reviews the assumptions behind the algorithms before identifying those things that, in the view of this author, would constitute a more transparent and hence improved research paper. It is hoped that this paper will help those of us who are asked to act as referees to be diligent in reviewing papers for journals.

Tourism Critiques: Practice and Theory
Vol. 1 No. 1, 2020
pp. 47-61
Emerald Publishing Limited
p-ISSN: 2633-1225
e-ISSN: 2633-1233
DOI 10.1108/TRC-03-2020-0002

Why use structural equation modelling?

The use of SEM has become quite popular for a number of valid reasons. It addresses the requirement for better assessing patterns of causality by taking the process beyond the conventional multiple regression. While multiple regression is a robust statistical method for determining the degrees of variance found in a determined variable, it is directly a measure between determined and determining variables. It does not measure the relationship between each of the determining variables and fails to completely consider the contribution each of the determining factors or variables makes to the determined and determining variables. To summarize, SEM provides much more detail about the statistical relationships between all the variables included in a model. It is this explanatory power of the technique that accounts for its popularity.

Equally, however, the methods of SEM are constructed on a series of assumptions. As is generally known, there are primarily two main forms of SEM. The older of the two is covariance-based SEM (CB-SEM) and the newer format is partial least squares structural equation modelling (PLS-SEM). Commonly, both involve a process of identifying common dimensions that underlies a series of observed measures. These observed measures may be a proxy for that latent or unobserved dimension or factor. So, for example, one may have a series of measures that purport to be measures of intelligence such as those including scores on various scales and possibly other measures such as scholastic aptitude test scores or grade point averages. The common factor that underlies these scores (and which is not directly observed) can be hypothesised to be "intelligence." Both complement the process of confirmatory factor analysis (CFA) in that they either seek to confirm relationships between variables or explore to confirm an ability of a model to be predictive.

This approach compares with the older exploratory factor analysis (EFA), which is simply "exploratory" in that it seeks to show that a series of correlations exist between the observed and the latent variable. The underlying "factor" – it simply, it might be said, that a series of correlations exist. It explores a relationship that is not necessarily posited or hypothesised. If, however, one was wanting to confirm a hypotheses that a series of measures comprise different underlying factors or dimensions and that each of these factors help "explain" a determined variable, then a CFA is undertaken to confirm the integrity of the variables, followed by a SEM calculation that in showing relationships between variables may "prove" or "disprove" a supposed set of relationships. To summarize, the propositions move from the exploratory to an explanation of causality.

The first computer programme that made the technique popular was linear structural relations (Lisrel), which was devised in the early 1970s by Karl Jöreskog and Dag Sörbom. In turn, this was followed by programs such as EQS, MPlus and AMOS. Initially, the programs were text-based similar to the popular Software Package for the Social Sciences (SPSS) program and older researchers like myself will remember the sometimes frustrating processes of searching through the syntax to identify the typographical error that caused a failure to produce the required statistics. What specifically increased the popularity of the SEM programs was the introduction of a graphical interface. To our knowledge, *analysis of a moment structures* (AMOS) was among the first if not the first that introduced such an interface for CB-SEM programs. Wong (2013) indicates that the first of the PLS-SEM programs with such an interface was PLS-Graph that appeared in the 1980s. At much about the same time, Svante Wold created PLS-regression (PLS-R) that was reliant on principal components analysis that permitted maximisation of the explanation of variance in the dependent model, although based on linear relationships. Such programs were among the antecedents of Smart-PLS 2.0, which was released in 2005 with a diagrammatic interface.

The introduction of such interfaces certainly made the whole process of statistical calculations much easier in that it permitted the researcher or author to draw a series of squares to represent the observed measures, as well as link these to circles that represented the latent or unobserved or underlying factors with a series of arrows; however, other arrows linked the circles with each other and the final determined variable. The requirement to painfully write the syntax was no longer required and the software did the remainder by completing the calculations.

In theory, in the process of drawing, the diagrams should help the researcher by making them consider the relationships more carefully between the variables and what are known as the "inner" and "outer" models, as well as the direction of the arrows. The "inner" model relates to the relationships between the latent variables where one such variable is a dependent variable. The "outer" model shows relationships between the observed and latent variables. The direction of the arrows indicates whether the SEM relationship is either formative or reflective. A formative relationship is one where the latent construct is the consequence of the observed variables and the arrows point from the observed variables to the construct, which effectively is behavioural response to the changes in those observed variables.

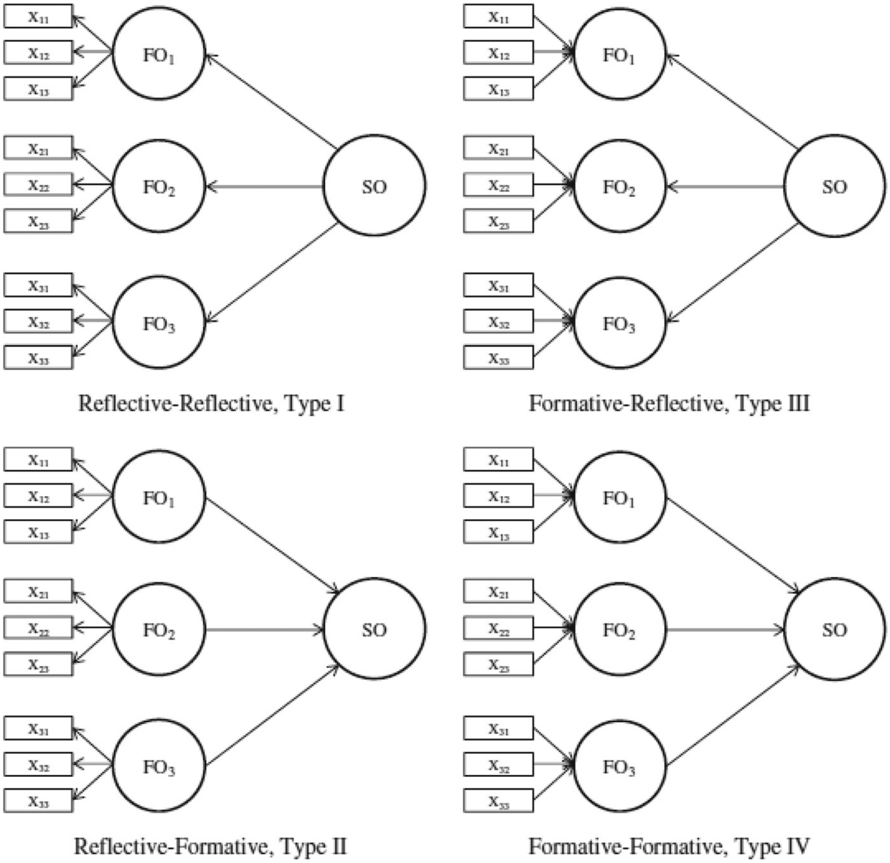
The reflective relationship is the situation where the latent variable is the underlying dimension that shapes the observed variables, and hence the direction of the arrows from the latent variable (conventionally drawn as a circle) to the squares that represent the observed measures. Figure 1 shows this relationship as illustrated by Becker *et al.* (2012, p. 363).

Two immediate implications arise from this. The first is that the actual diagram reflects a series of hypothesised relationships; hence, the researcher builds the diagram on the software to reflect a proposed pattern of relationships. Hence the term, CFA – as noted above – the researcher is seeking to "confirm" (or alternatively is testing to fail) a hypothesis to show that a prima facie pattern of causality exists. Having noted this traditional nomenclature it must, however, be fully recognised that the compilers of Smart-PLS are firmly of the opinion that PLS-SEM is a predictive-oriented approach and hence exploratory while being grounded in causal relationships (Hair *et al.*, 2017a, 2017b; Sarstedt *et al.*, 2018). PLS-SEM therefore seeks to maximise the variance in the dependent variable. This is in comparison to CB-SEM that seeks to minimise differences between an observed covariance sample matrix and a subsequent estimated covariance after a model has been confirmed. The model is confirmed by the goodness of fit between the two calculations.

Becker *et al.* (2012) indicated four types of models (as shown in Figure 1). Of these, it might be noted that their comment that the reflective-reflective has been characterised by Lee and Cadogan (2013) as being generally inconceivable in management research; furthermore, such a model is "at worst, misleading, and at best meaningless" (Lee and Cadogan, 2013, p. 245). It is suspected that, in tourism studies, many models would be of the reflective-formative or indeed formative-formative type. This, it is thought, represents the general situation where PLS-SEM is being used.

PLS-SEM has become commonly used in tourism studies, especially with the introduction of the software program Smart-PLS. Many readers will be familiar with the blue and yellow diagrams that are found in such studies. The reasons for its popularity are generally stated as including:

- the method can be used with relatively small samples;
- the observed variables do not necessarily have to be independent of each other; and
- normally distributed data that does not need to be present (Hair *et al.*, 2017a, 2017b).



Reflective-Reflective, Type I

Formative-Reflective, Type III

Reflective-Formative, Type II

Formative-Formative, Type IV

Figure 1.
Reflective and formative relationships in SEM

The last two considerations can be important given much of tourism research. Generally, holiday-makers will not select holiday destinations that they will not enjoy. Indeed, it is suspected that, in the majority of surveys of holiday-makers, negative skew exists and indeed such will be the level of satisfaction that kurtosis may be relatively high. That is, if a seven-point satisfaction scale is used, the majority of responses will vary from five to seven where seven represents the highest score. Equally, when destination attributes are being discussed using measures such as beautiful landscape, cultural richness, architecture, crafts, heritage and history, in such a list it might be stated that each item is not wholly independent of the other. Landscape shapes buildings, buildings reflect a culture and culture helps inform heritage as well as history. In such, the multiple constructs of destination attractiveness the items and variables being used may not be entirely independent.

Assumptions relating to structural equation modelling

As a general statement of opinion, this author queries whether, in tourism studies, it is generally apposite to use CB-SEM given the theoretical assumptions that lie behind the technique. First, there is assumed to be a normality of distribution in the patterns of measures. In tourism, generally, people select destinations to visit that will satisfy their demands, and hence form the outset there is skew toward higher scores of perceived destination attractiveness, experiences and satisfaction even prior to processes of cognitive dissonance that will mitigate against disappointment. Second, CB-SEM assumes that the variables are independent of each other; however, as noted in the last paragraph, this in many situations may not be the case. This can even apply to the way in which socio-demographics might apply. From one perspective, gender, age, education and income might appear to be independent variables; however, in practice, the older, well-educated male possibly earns more than the female who leaves school with minimal qualifications. Gender income disparities exist, higher education does tend to permit higher incomes to be earned, and older people have generally had more time to acquire accumulated wealth. In short, the practical effect is that life stage and past family histories do not necessarily create independence between the variables.

Again, CB-SEM requires representative samples and some adherents argue that samples should be properly random. Much research in tourism is based on convenience sampling rather than random sampling. Indeed, in many instances, the characteristics of any given population might not be accurately known or being easily accessible. For example, in much academic research, while surveys are taken of visitors, what is known of non-visitors who perceive a destination as being part of their inert or inept sets of considered holiday locations?

It might be philosophically asked why are assumptions for CB-SEM so rigorous? One answer is that a key function of CB-SEM is to validate data as confirming a set of hypotheses. It can be suggested that the rigour of proof required to support a theoretical model should indeed be tighter or higher than that required to explore possible measures. Therefore, it would follow that sample size is important when using CB-SEM.

Consequently, similar to many others, Astrachan *et al.* (2014, p. 118) noted that CB-SEM has an inherent tendency to require relatively large sample sizes because "[. . .] because relationships between all variables must be assessed (i.e. a full information approach), while with PLS-SEM the model is separated into different smaller components (a component for each construct in the model; hence the name partial least squares)". What constitutes a large sample, however, is not often defined and many authors take the path of citing some "rule of thumb" about the numbers of respondents as some ratio of the number of items on a questionnaire. There seems to be comparatively little statistical justification for such rules, and Westland (2010) noted this is suggesting that many papers had far too small a sample size to justify the claims being made by authors. Although he subsequently revised his proposed algorithms for calculating sample size (Westland, 2012), this point remains valid. Indeed, easily available calculators exist online (such as www.danielsoper.com) and various statistical programs such as PASS15, which will provide estimates for sample size in SEM procedures at various levels of power and size effects alternatives. Having suggested these to various authors at differing times, there does appear to exist some misapprehension about what to do and often authors simply select the smallest size that they feel they can get away with. As one of the "rules of thumb," I often ask graduate research students at what level of sub-sample do they wish to work. For example, if wishing to apply analysis of variance (ANOVA) to differing age groups, given a sample divided into perhaps six different age groups, then assume 80 per age group, which alone would signify 480 respondents. Allow for non-responses when asking for gender adds a further number, and a requirement to distinguish between males and females at each age group makes again for a larger sample size. If a technique such as hierarchical logistic multinomial regression is to be used or some derivative thereof, then to ensure that the various cells established by the routine contain sufficiently numerous numbers, the sample must be large.

While it is often stated that PLS-SEM calculations do not require large samples, that does not amount to argument that sample size is of little importance. Indeed, researchers may often accompany their SEM calculations with other statistical tests, some of which are discussed below. While PLS-SEM is about building potential models rather than confirming models, the predictive effort does seek to achieve generalisation and, as suggested in the preceding paragraph, samples should remain substantive in nature when applied in tourism research.

Another issue seemingly forgotten by many authors is a requirement for some simple transparency. All too often one finds studies in which missing data abounds in all manners of ways. First, there is (at least for me) the annoying practice of referring to items by acronyms. One finds list of statistics for variables or items such as SAT1, SAT2, SAT3 [. . .] SATn. Reading of the text indicates that SAT means "Satisfaction" and that *n* measures of satisfaction were used, but one looks in vain for a questionnaire or any accompanying clue as to what the actual were.

Without the presence of a questionnaire, other issues arise: Did the sample answer every question? Was there the provision of a non-response option? If so, how many selected that option? If that did arise, how was the missing data handled? Was the item included in calculations? These days SPSS does provide help in identifying patterns of non-response, and the methods of imputation of values, in addition to simply using the overall mean score for an item. If non-response options were not included, then a requirement to carefully examine the questionnaire arises. If, for example, a sample is asked to assess their evaluation of every service facet, how does a respondent respond when they had not used a specific service? One suspects they select the mid-point option, often labelled "neither agree nor disagree" – but surely there is a qualitative difference between "I am genuinely indifferent" and "I cannot express an opinion because I have not tried the service." Had the author of the questionnaire considered that possibility in the design of the questionnaire?

There remains another issue and that is the distinction between items being important as measured by their high mean score, and items possessing strong discriminatory power as often measured by a relatively high standard deviation. Many statistical tests are based on variance and an ability to discriminate. This means that researchers using regression techniques and SEM is such an associated technique, which will often overlook the importance of including descriptive statistics in their studies; therefore, they fail to report what is important and unimportant when describing tourist's perceptions. Hence, if studying a group of hikers who respond to an item such as "To what do you agree that hiking creates moments of special bonding with friends or family?" using a seven-point scale where "1" equals "very strongly disagree" and "7" represents "very strong agree" and "0" represents a non-response option – with a mean score of say 6.3 and a standard deviation of say 0.6 – it can be said that this is a very important consideration, there is very little deviation from this viewpoint and this is not a particularly good discriminatory item because most respondents are agreeing. Almost certainly there would be a strong negative skew with a high kurtosis score as would be seen from a histogram of scores and there would be a great deviation from a normal distribution of scores. In any analysis of discriminatory power, this item would not show as possessing importance. The problem is that many researchers writing a SEM paper tend to omit the descriptive statistics, and thus failing to note the importance of such an item. In studies of a collective-oriented culture, such as China, such a result confirms, for example, the social aspects of hiking in that context (Li *et al.*, 2020).

Equally, very low scores with a high positive skew and again a high kurtosis score would provide little discriminatory power; however, surely from a managerial perspective, it is possibly just as important to know what has no importance to tourists to avoid unnecessary infrastructure or promotional effort.

In the literature, given this observation and noting that many studies reported that report SEM calculations omit these types of considerations, one can only conclude that many studies being published fail to be transparent and equally many referees are failing to ask the right questions of authors.

In examining the special issue of the journal and reading the various papers, it became evident that various examples of these practices were easily found in the papers, yet all had been reviewed and obviously the papers had been supported by the referees.

Proposed check list for referees

At the 2018 APAC-Chrie Conference, Ryan (2018) noted that many tourism journals are members of Committee on Publishing Ethics (COPE) and that COPE is fully congruent with best practice as indicated on web sites such as UK EQUATOR Centre. For those involved in medical research, questions of transparency in research are paramount, especially in light of notorious examples such as the Tuskegee Syphilis experiments that gave rise to the 1979 Belmont Report in the USA and the controversial nature of the Stanford and Stanley Milgram experiments in social psychology. A large number of those recommendations apply to tourism research; however, one finds that if tourism research was to be judged by those standards, it would often be found to be lacking. If anything, as increasingly sophisticated methods of analysis are emerging in these days of big data, web scrapping, panel data and the like, researchers and referees are seemingly caught up in the specific mechanics of process and reporting but fail to provide the wider context. This tendency is being reinforced, in my opinion, by a metric-driven system that encourages an almost slavish adherence to complex analytical methods when often, as noted by Dolnicar (2015), simpler methods will suffice to answer the research question. So, the following criteria are proposed for the refereeing (and writing) of SEM papers. In saying this, it is recognised that many of these criteria are generic in nature and can apply to other research outputs. Examples of good practice would therefore involve:

- publishing a copy of the questionnaire;
- justifying the items being used on the scales;
- providing dates as to when the research was carried out;
- providing details as to the sample and justifying the selection of the sample and its size;
- reporting descriptive statistics with the mean and standard deviations and commenting on the item scores without recourse to acronyms;
- checking for normality of distribution and commenting as to why a lack of normality is not an issue; and
- justifying the use of either CB-SEM or PLS-SEM – do not just automatically provide results.

Why is the questionnaire essential?

The validity of the research depends in the integrity of the questions being asked as much as on transparency when reporting results. Statistical software will always report a number, but (as yet) the various statistical software programs cannot assess the content of the question or the mode of asking. Even today, it is not uncommon to see reports published that say that tourists value [. . .] followed by series of items such as “clean environment,” “congestion,” and “history.” What is meant by such a list? What was the actual question set to the tourist? Language matters! And the contents of a questionnaire matter.

Without the questionnaire being provided, how is the referee supposed to assess the research article and the veracity of the results? If, for example, the findings of an EFA are being reported but no questionnaire is being provided, how is a referee to tell if all of the items were included in the EFA? The author is possibly hiding the fact that some items “did not work.” Some authors seem to engage in what is little more than a tautology by stating the “x” numbers of items were drawn from one scale and “y” number of items from another scale. Given the two scales measure two separate dimensions, it is not surprising that a two-factor solution might result. This, however, can be justified as indicating that the sample was responding in a consistent and logical manner, and it confirms the validity of progressing to assess relationships between two factors for model building. This nonetheless reinforces the importance of the theoretical model building that lies behind research purpose; thus, clear linkages need to be made between any literature review, its implications for the research question, the selection of the items and the modes of questioning, as shown by the questionnaire. Literature review, theoretical propositions and the actual questionnaire are all part of a seamless whole; if the questionnaire is not provided, the reporting can be held to be deficient. The questionnaire can be easily added as an appendix to the main article.

Justifying the items

Justification of the items can be undertaken at two levels. First, has the author indicated why specific items have been selected? As noted above, an author may write that they have “taken 3 items from the well-researched scale by XYZ”; however, sometimes, as a referee, looking up the scale, one might find the scale contain many more than three items. So, why these three? Perhaps, the author used more than three items on the questionnaire; however, if one does not have access to the questionnaire, one cannot judge. Possibly when conducting an EFA, the communalities for the items were too low – if so – then the author should clearly state this is the fact.

The second mode of justification is from the statistical perspective. Indeed, today most authors do create and report reliability measures such as Cronbach’s alpha coefficient and eigenvalues; unfortunately, they do not always the focus on the abovementioned commonalities. These can add to an understanding of the rigour of an item because they indicate the degree of variance in the individual items that is explained by the factors revealed by the EFA calculation. Reporting such data does add to the confidence a referee might have in the data.

It might be objected that adding such information within the main text of a study may mean that the author is in danger of exceeding the numbers of words being permitted by the editors – a total many editors seem to adhere too with almost fierce determination. However, the major publishers today provide means by which additional data can be added to an article – something which is certainly not uncommon in many fields of research – but which traditionally has not been commonly used in tourism research. Unfortunately, I suspect many referees and readers remain unaware of such additional data when it is supplied, and in the final published article possibly few actually click on the final footnote that states additional information might be found on a specific webpage.

Finally, it is recommended that authors should be using composite reliability measures to supplement the more traditional Cronbach alpha coefficients because, as shown by Chin (1998), this approach adopts the use of the indicators differential weights, whereas the alpha coefficient assigns equal weights to the indicators (Dijkstra and Henseler, 2015; Hair *et al.* (2017a, 2017b).

Dating the research

The argument relating to the dating of research is, at one level, obvious, and at another level irrelevant. The relevant argument is that much of tourism research is clearly context- and time-related. Butler (1980) clearly stated in the concept of the tourist area life cycle that destinations change over time. Land usage changes, the scale of tourism changes and so too do the type of visitors being attracted (Ryan, 2020). From this perspective, the dates when the research was being undertaken significantly can matter because work done, say a decade earlier, may no longer be wholly relevant to the contemporary position of the destination. This may be particularly true of areas that have undergone significant change in short periods of change, a situation not uncommon in the People's Republic of China.

However, if the emphasis of the research relates more to conceptual or more generic aspects of tourist behaviour or destination development, it might be that the timing of the research is less crucial. Destinations pass through the stages of the tourist life cycle at different times; hence, research that has been written up several years after the actual study was undertaken may still have relevance in terms of lessons for future places or students. Perhaps, the main moral to be drawn is that referees (and editors) should ask for the dates when the research was done, and the readers can then judge for themselves as to whether research retains validity.

On this point, one thing that can irk some referees is when authors write that "little research has been undertaken on XYZ" – and then proceed to support the contention with a reference taken from two or more decades earlier. This does raise questions as to whether the contention is correct. Perhaps, it might be better to state that a search through sources such as Google Scholar revealed only a small number of studies having been undertaken in recent years.

Describing the sample

This is important, and to be fair to the authors of the journal issue that prompted this paper, the great majority of contributors did describe the socio-demographics of their samples, but often the tables given are not commented upon. Samples do need to be justified as being, ideally, representative of populations, or if not representative, to have sufficiently large enough sub-sample sizes relating to socio-demographic groupings that permit comparative analysis and hence permit generalisation.

In qualitative research again, it is necessary to describe and justify the sample, albeit qualitative research may well not be targeted at achieving generalisation but rather the seeking of more detailed detail, often related to the experiences of respondents. Nonetheless, the researcher should be transparent as to why the sample was thought pertinent to the objectives of the research. This issue arises where the SEM component is part of a mixed methods approach to a research question.

As noted above, the size of a sample must be discussed with reference to the demands of any statistical technique being applied and the level of sub-sample analysis being undertaken. Reference to measures such as Cohen's D or similar statistics are recommended to deal with power and size effects.

Another element of note with reference to sample size arises in discussions pertaining to sample size and the slow adoption of Bayesian techniques. It is not the function of this paper to cover the latter. The aforementioned book by Pearl and MacKenzie (2018) provides a history of the thinking; however, in tourism, Arch Woodside has argued to case for escaping from the “tyranny” of the null hypothesis in various blogs to Trinet and in various other published paper such as Woodside (2017). The debate about bootstrapping draws upon an important distinction. Friedman *et al.* (2013) summarises this by stating bootstrapping is concerned with “assigning measures of accuracy to statistics estimates and performing statistical inference” (p. 196). The usual way of attempting to do this is to achieve the highest possible measures of confidence that the likelihood of an assertion is true; hence, the use of testing a null hypothesis. However, Friedman *et al.* (2013) point to a second interpretation, namely, “to an assessment of the degree of support of a particular technique towards a given feature” (p. 196). They then commented that this latter notion neatly separates the variation within the data from any shortcomings in the analytical algorithm. In the remainder of their article, they test a series of test-retest calculations using various bootstrapping approaches to conclude that bootstrapping is conservative, has value and that non-parametric Boolean bootstrapping can generate results in which one can have confidence. Finally, it is noted that application of these techniques to small sample sizes is permissible.

Dealing with non-normality

As noted above, tourism data is often characterised by skew and items about which there are strong levels of agreement or disagreement are of interest but may lack distinct discriminatory power. It is recommended that authors explicitly deal with this aspect of their data and use measures of multi-collinearity and independence of variables. Referees should look for statistics such as variance inflation factors (VIF) and ideally values certainly should not be above 10 (which indicates significant correlation between supposedly independent items or factors). Just what is acceptable is a contentious issue, but a conservative approach would be not to accept VIF readings of above 2.5 – and many authorities would suggest values above 5.0 are too high for inclusion in analysis. However, note that the use of dummy variables of a nominal nature with three or even more classifications can produce high VIF scores but generally it is possible to ignore these (Hair *et al.*, 2017a, 2017b). Of interest here is the work of Jannoo *et al.* (2014) who found that CB-SEM was not able to calculate paths when non-normality was present, whereas it was possible when using PLS-SEM. However, the deficiency of CB-SEM could possibly be remedied if a sample was sufficiently large.

Standard tests of non-normality that may well be cited include the Kolmogorov–Smirnov (K–S) Goodness of Fit test (please do not confuse this with the similarly named Kolmogorov–Smirnov two sample tests, which creates measure of distribution differences between two samples). Some authors may prefer citing the Anderson–Darling test and/or Cramer–Von Mises test, which are thought to possess some advantages, are variants of the K-S Goodness of Fit statistic. These three tests are generally applicable to continuous data but some have argued that they may be applied to discrete data with caveats (Razali and Wah, 2011). Note that the K-S test requires the mean and the variance in its calculation and works by comparing the expected and actual distribution and hence is a general test of normality (Razali and Wah, 2011). The Shapiro–Wilk test is a more specific and hence more powerful test of normality, but it appears that most authorities tend to feel it is more applicable to smaller sample sizes (Mendes and Pala, 2003; Razali and Wah, 2011). Because it is recommended that samples in surveys of tourist behaviours should have a minimum size of 500 usable respondents as argued above, it follows that the K–S test statistic would be the one most cited. However, given the ease of calculation using SPSS, many authors cover themselves by citing both K–S and Shapiro–Wilk, which tend to agree when testing normality.

As a note, and reverting back to the K–S two sample test, and assuming non-normality exists for in scores generated by both genders, it would make sense to use the non-parametric K–S two sample test rather than the more conventional *t*-test. However, parametric tests are very rigorous and in the author's experience a calculation that uses the parametric test and passes will usually pass the less rigorous non-parametric test.

Justify the use of the version of structural equation modelling being used

As an author and reviewer, it has been noticed that several authors do use AMOS without commenting on why they believe it to be appropriate for their data. It is generally recognised that CB-SEM is rigorous and pertinent for the testing of previously stated hypotheses (Razali and Wah, 2011); however, as previously noted, it is based on various assumptions about datasets – assumptions not always present when dealing with the type of data common in research into tourism sites, their perceived attractiveness and tourist perceptions and behaviour. It would therefore seem incumbent on researchers to state why it is felt appropriate to use CB-SEM based software.

Given this, there seems to be some logic in assuming that PLS-SEM would be the normally adopted software when seeking to examine causal relationships in tourism. Normal justifications are that it is suitable for smaller samples and skewed data. As Chin (1998, p. 295) notes the technique is certainly appropriate for suggesting "where relationships might or might not exist and to suggest propositions for testing later." Chin (1998, p. 316) goes on to write:

> Because PLS makes no distributional assumption, other than predictor specification in its procedure for estimating parameters, traditional parametric-based techniques for significance testing/evaluation would not be appropriate. Instead, Wold (1980, 1982b) argued for tests consistent with the distribution-free/predictive approach of PLS. In other words, rather than based on covariance fit, evaluation of PLS models should apply prediction-oriented measures that are also nonparametric. To that extent, the R-square for dependent LVs, the Stone-Geisser (Geisser, 1975; Stone, 1974) test for predictive relevance, and Fornell and Larcker's (1981) average variance extracted measure are used to assess predictiveness, whereas resampling procedures such as jackknifing and bootstrapping are used to examine the stability of estimates.

Again, to make an observation, it does appear that while several authors do specifically use the Fornell and Larcker Average Variance Extracted (AVE) test to ensure statistics exceed the conventional 0.5 criterion, often the r-squared statistic is ignored. Note that AVE is a test of convergent reliability while the R-squared statistic is "a measure of the variance explained in each of the endogenous constructs and is thus a measure of the model's predictive accuracy (in terms of in-sample prediction)." The R^2 ranges from 0 to 1, with higher levels indicting a greater degree of predictive accuracy (Sarstedt *et al.*, 2014, p. 110). Sarstedt and his co-authors go on to suggest that "R^2 values of 0.75, 0.50 and 0.25 may be considered substantial, moderate and weak, respectively (Hair *et al.*, 2017a, 2017b; Henseler *et al.*, 2009)" (Sarstedt *et al.*, 2014, p. 110) and conclude that the validity of the R^2 needs to be evaluated in the context of the current and related studies.

It also seems to the present author that it is only fairly recently in the tourism literature that the techniques of bootstrapping and jackknifing are being used, probably because of improvements in the nature of the software.

In this respect, it is suggested that authors keep abreast of developments in PLS-SEM software and theoretical advances. For a long time, it was thought that comparative analysis of three or more groups was not possible. Indeed, Hair *et al.* (2017a, 2017b) indicated this even in the second edition of their popular book, *A Primer on partial least squares structural* *equation modelling (PLS-SEM),* but subsequently the technique has become increasingly powerful with the launch of Smart-PLS 3.0.

This variant of Smart-PLS now includes additional (and desirable tests) that previously had to be independently done of the software. Examples include f^2 effect size and tests for multi-collinearity. More notable additions include confirmatory tetrad analysis, a heterotrait–monotrait ratio correlation tests of discriminant validity, prediction-oriented segmentation, moderator analysis, multi-group analysis and invariance testing. These improvements met some of the previous criticisms made of the software such as limited group comparison testing; however, incorporating these into the software creates more ease for researchers wishing to text their results. It is suggested that referees should be requiring authors to show awareness of such testing.

While Smart-PLS is commonly used, there are other SEM programs, both CB- and PLS-based. Of the latter, Ned Kock's Warp-PLS has been used by the author and offers a good alternative to Smart-PLS. It too has a graphical interface, has a stable platform, is based upon DiDijkstra's consistent PLS (PLSc) technique and can handle non-linear data. Kock (2018) suggest that this PLSc technique provides a reliability measure that "appears to be, in many contexts, a better approximation of the true reliability than the reliability measures usually reported in PLS-based SEM contexts" (p. 8). It includes a number of features, including the testing for Simpson's paradox should that be present in the data. Similar to Smart-PLS, it provides for inter-group testing and assessments, controls for endogeneity assessment, all the conventional goodness of fit measures plus additional measures of reliability and even includes within the program a routine to assess whether the sample is sufficient. Sub-routines exist for handling non-linear relationships. Other programs used by the author include the CB-based EQS program, which has its own advantages, i.e. creating (at the time when used) more statistical tests than AMOS. It too now has its own graphics interface.

Summary

The above text has offered a personal opinion about the usage of SEM programs in tourism research with the objective of indicating in which ways referees might help authors improve the quality of their papers, or certainly improve the quality of those that are published.

A check list for referees offers as a summary a potential check list that reviewers might wish to use when refereeing an article for a journal.

Criterion

Q1. Is a copy of the questionnaire available?

Q2. Is there a discussion of why a given form of SEM used?

Q3. Does such a discussion provide at least a brief comparison of SEM techniques to justify the choice made?

Q4. Is there a discussion about why sample was pertinent?

Q5. Is there any testing of the adequacy of the sample size?

Q6. Are basic data about reliabilities, such as AVEs and CR, available?

Q7. Did the tables provide meaningful descriptions of labels and not just a list of non-standard acronyms?

Q8. Are the items of observable measures fully described?

Q9. Was there any qualitative stage (of any kind) involved prior to setting questions?

Q10. Or were questions based on previous scales? Is there a reasonable description of those scales?

Q11. Is there a justification offered for the selection of the observable items used for measurement?

Q12. Are the descriptive statistics provided? Are they discussed prior to building the model?

Q13. Latent vs formative relationship – is this discussed?

Q14. Are the hypotheses of interest (or self-evident)?

Q15. Is there a discussion about possible non-response on items? Is there clarity about any imputation method to ensure that as much of the available data is actually used in the calculations?

Q16. Are the data linear or non-linear? If the latter, how is this managed?

Q17. Are there tests of multi-collinearity, independence of the variables and other pertinent tests of data integrity?

Q18. Did the paper specify the software selected and commented, even if briefly, as to why it was chosen?

Q19. Are the models of interest? Can the author clearly state the importance of the research?

Note that the final item is a core question for any research exercise. As a referee and editor, it seemed to the author that there were simply too many articles on the linkage that satisfied tourists would recommend a given destination – a rather tired observations that had significant amounts of prior studies that had not used SEM but had shown the relationship to exist. It is true that several articles exist in marketing that have used SEM techniques (Wong, 2013). Similarly, marketing issues relating to tourist destination promotion, tourist behaviours and perception have long proven to be a core subject in the tourism literature – but several new issues have emerged. These include the role of the internet with peer to peer services, over-tourism, climate change, the role of and use of big data and 5G. Note that these topics are of significance for tourism, and all equally rich topics that can be analysed by techniques that include SEM. There seems, therefore, many opportunities to widen our literature beyond tourist satisfaction; however, tourist satisfaction in a future of constrained choice might yet still offer many research opportunities.

It is also suggested that referees and authors in tourism occasionally refer to the more specialist statistical journals and the associated specialist blog and on-line support materials that exist, and often associated with specific software discussion groups. Currently, there are a number of sub-routines being suggested using “add-ons” written in python, and these will be of help to referees.

References

Astrachan, C.B., Patel, V.K. and Wanzenried, G. (2014), "A comparative study of CB-SEM and PLS-SEM for theory development in family firm research", *Journal of Family Business Strategy*, Vol. 5 No. 1, pp. 116-128.

Becker, J.M., Klein, K. and Wetzels, M. (2012), "Hierarchical latent variable models in PLS-SEM: Guidelines for using reflective-formative type models", *Long Range Planning*, Vol. 45 No. 5-6, pp. 359-394.

Butler, R.W. (1980), "The concept of a tourist area cycle of evolution: implications for management of resources", *The Canadian Geographer/Le Géographe Canadien*, Vol. 24 No. 1, pp. 5-12.

Chin, W.W. (1998), "The partial least squares approach to structural equations modeling", in Marcoulides, G.A. (Ed.), *Modern Methods for Business Research*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 295-335.

Dijkstra, T.K. and Henseler, J. (2015), "Consistent partial least squares path modeling", *MIS Quarterly*, Vol. 39 No. 2, pp. 297-316.

Dolnicar, S. (2015), "In future, I would love to see [. . .] a reflection on the state of quantitative tourism research", *Tourism Review*, Vol. 70 No. 4, pp. 259-263.

Friedman, N. Goldszmidt, M. and Wyner, A. (2013), "Data analysis with bayesian networks: a bootstrap approach", arXiv preprint arXiv:1301.6695.

Hair, J.F., Jr, Matthews, L.M., Matthews, R.L. and Sarstedt, M. (2017b), "PLS-SEM or CB-SEM: updated guidelines on which method to use", *International Journal of Multivariate Data Analysis*, Vol. 1 No. 2, pp. 107-123.

Hair, J.F., Jr, Hult, G.T.M., Ringle, C.M. and Sarstedt, M. (2017a), *A Primer on Partial Least Squares Structural Equation Modelling (PLS-SEM)*, 2nd ed., Sage Publications, Thousand Oaks, Calif.

Henseler, J., Dijkstra, T.K., Sarstedt, M., Diamantopoulos, A., Straub, D.W., Ketchen, D.J. and Calantone, R.J. (2014), "Common Beliefs and Reality about Partial Least Squares: Comments. Rönkkö & Evermann (2013)", *Organizational Research Methods*, Vol. 17 No. 2, pp. 182-209.

Jannoo, Z., Yap, B.W., Auchoybur, N. and Lazim, M.A. (2014), "The effect of nonnormality on CB-SEM and PLS-SEM path estimates", *International Journal of Mathematical, Computational, Physical and Quantum Engineering*, Vol. 8 No. 2, pp. 285-291.

Kock, N. (2018), *Warp-PLS 6.0 Manual*, Script-Warp Systems, Laredo, Tx.

Lee, N. and Cadogan, J.W. (2013), "Problems with formative and higher-order reflective variables", *Journal of Business Research*, Vol. 66 No. 2, pp. 242-247.

Li, P., Ryan, C. and Zhou, B. (2020), "The motivations of Chinese hikers: data from ningbo", *Current Issues in Tourism (SSCI)*, doi: 10.1080/13683500.2019.1646224.

Mendes, M. and Pala, A. (2003), "Type I error rate and power of three normality tests", *Information Technology Journal*, Vol. 2 No. 2, pp. 135-139.

Pearl, J. and Mackenzie, D. (2018), *The Book of Why: The New Science of Cause and Effect*, Basic Books, New York, NY.

Razali, N.M. and Wah, Y.B. (2011), "Power comparisons of Shapiro-Wilk, Lolmogorov-Smirnov, lilliefors and Anderson-Darling tests", *Journal of Statistical Modeling and Analytics*, Vol. 2 No. 1, pp. 21-33.

Ryan, C. (2018), "Keynote: Publishing research: the new challenges", *Apac-Chris Conference*, Sun Yat-sen University, Guangzhou, China.

Ryan, C. (2020), *Advanced Tourism Destination Management*, Edward Elgar Publishing, Cheltenham.

Sarstedt, M., Hair, J.F., Ringle, C. and Hair, J.F. (2018), "Partial least squares structural equation modeling", in Homburg, C., Klarman, M. and Vomberg, A. (Eds), *Handbook of Market Research*, Springer, Germany, pp. 1-40.

Sarstedt, M., Ringle, C.M., Smith, D., Reams, R. and Hair, J.F. Jr (2014), "Partial least squares structural equation modeling (PLS-SEM): a useful tool for family business researchers", *Journal of Family Business Strategy*, Vol. 5 No. 1, pp. 105-115.

Westland, J.C. (2010), "Lower bounds on sample size in structural equation modeling", *Electronic Commerce Research and Applications*, Vol. 9 No. 6, pp. 476-487.

Westland, J.C. (2012), "Erratum to ''lower bounds on sample size in structural equation modeling", *Electronic Commerce Research and Applications*, Vol. 9 No. 6, pp. 476-487.

Wong, K.K.-K. (2013), "Partial least squares structural equation modelling (PLS-SEM) techniques using Smart-PLS", *Marketing Bulletin*, Vol. 24 No. 1, pp. 1-32.

Woodside, A.G. (2017), "Releasing the death-grip of null hypothesis statistical testing ($p < 0.05$): applying complexity theory and somewhat precise outcome testing (SPOT ", *Journal of Global Scholars of Marketing Science*, Vol. 27 No. 1, pp. 1-15.

Corresponding author
Chris Ryan can be contacted at: caryan@waikato.ac.nz