One Ecosystem 5: e50452
doi: 10.3897/oneeco.5.e50452

OPEN ACCESS

Review Article

A 'Weight of Evidence' approach to evaluating structural equation models

James B Grace ‡

‡ U.S. Geological Survey, Lafayette, Louisiana, United States of America

Corresponding author: James B Grace (gracej@usgs.gov)

Academic editor: Stoyan Nedkov

Received: 24 Jan 2020 | Accepted: 10 Mar 2020 | Published: 13 Mar 2020

Citation: Grace JB (2020) A 'Weight of Evidence' approach to evaluating structural equation models. One Ecosystem 5: e50452. https://doi.org/10.3897/oneeco.5.e50452

Abstract

It is possible that model selection has been the most researched and most discussed topic in the history of both statistics and structural equation modeling (SEM). The reason for this is because selecting one model for interpretive use from amongst many possible models is both essential and difficult. The published protocols and advice for model evaluation and selection in SEM studies are complex and difficult to integrate with current approaches used in biology. Opposition to the use of *p*-values and decision thresholds has been voiced by the statistics community, yet certain phases of model evaluation have been historically tied to reliance on *p*-values. In this paper, I outline an approach to model evaluation, comparison and selection based on a weight-of-evidence paradigm. The details and proposed sequence of steps are illustrated using a real-world example. At the end of the paper, I briefly discuss the current state of knowledge and a possible direction for future studies.

Keywords

structural equation modelling, path models, model evaluation, model selection, weight-of-evidence

Introduction

Model selection is one of the more challenging aspects of structural equation modeling. The selection decision typically follows a multi-step process of model evaluation that considers numerous possible models and various types of evidence. Traditionally, this process has depended strongly on the use of *p*-values and the concept of dichotomous hypothesis tests. There have been numerous calls from the statistics community for a cessation of *p*-value-based hypothesis testing, especially as of late (see commentary by Amrhein et al. (2019) in *Nature* Magazine). The present state of the methodology literature may leave investigators with some uncertainty about how to use the information reported by software packages so as to conduct model evaluation and selection. In this paper, I suggest an approach to model evaluation in SEM that weighs multiple sources of information to guide model selection, based on a 'weight of evidence' (WOE) paradigm. A primary objective is to show how *p*-values may be used within a context in which they are not the final arbiters of model selection decisions. The presentation focuses on SEM conducted using traditional global estimation methods where dealing with *p*-values is unavoidable because of their association with the chi-square global fit test. A basic familiarity with the methodology is assumed in this presentation. For those interested in these methods who need background information, accessible treatments include Grace (2006), Kline (2016) and Shipley (2016).

When evaluating models estimated using traditional methods, there are two instances where SEM investigations encounter *p*-values:

1. associated with measures of global model fit and
2. associated with test statistics for individual parameters.

The likelihood-ratio X^2 test statistic was originally proposed as the first and final judge of global model fit (see Tomer (2003) for a discussion of the history). There are dual requirements within this tradition. The first was that the *p*-value associated with the model X^2 statistic should be > 0.05 to signify that there is no major discrepancy between observed and model-implied covariances. The second requirement was that all links retained in a model should meet the expectation that removal of any one of them would result in an increase in the X^2 statistic $\geq$ 3.84, the single-degree of freedom criterion value associated with $p = 0.05$.

Throughout the modern history of SEM, which can be thought of as the time since the LISREL synthesis in the early 1970s (Jöreskog 1970), there have been concerns about the use of *p*-values. A central concern that arose early in the history of SEM comes from the fact that, because $X^2 = (n-1)*F_{ml}$, where F_{ml} is the magnitude of the maximum likelihood discrepancy function, a finding of global mis-fit is very sensitive to the size of the data sample. This sensitivity means that dichotomous determinations of model adequacy and parameter significance are not properties of the system, but instead, properties of the size of the sample in hand. For this reason, SEMers have for decades sought alternative approaches to model comparison and evaluation (discussed below).

Within the field of ecology, Burnham and Anderson (2002) have championed the replacement of *p*-value-based dichotomous hypothesis tests with model comparisons, based on information measures such as the Akaike Information Criterion (*AIC*). There has been some resistance to the idea of shifting away from *p*-value-based dichotomous significant testing (Murtaugh 2014; Burnham and Anderson 2014) and, at present, we see plenty of uses of both approaches. Nonetheless, the use of model comparisons and multimodel inference is now perhaps the dominant paradigm in ecology. This is certainly not the case in all scientific disciplines. In the health sciences, for example, there is a major emphasis on exposure-response studies and for these, classic group-difference testing commonly relies on *p*-value-based significance tests.

Shifting away from a strict reliance on dichotomous hypothesis testing towards model comparisons has implications for model selection in SEM studies. Results reported by SEM software includes *p*-value-based indices. Experience tells us that *p*-values are useful indicators of model-data relations, but today's general advice is to not use them for dichotomous significance testing. In this paper, I illustrate an approach to model evaluation and selection that utilises *p*-values as information, while basing model selection on a comparison of the total weight of evidence. I first discuss the types of evidence that might be considered, including the role for expert knowledge. Second, I propose a sequence of steps to follow. Third, I illustrate the proposed process using an ecological example. Finally, I briefly consider the challenges facing the quest for a single “best” index and possibilities for resolution of this issue.

Types of Evidence that can be Considered in Model Evaluation

There are numerous types of evidence to be weighed when evaluating and comparing models. First and foremost is the scientific knowledge of the investigative team.

The Role of the Investigator in Model Evaluation and Selection

As an explanatory method, SEM requires the scientist to play an active role in the model evaluation process. A priori scientific knowledge is essential for the construction of the initial models. However, there may be a tendency for those beginning to use SEM to imagine that model evaluation, based on the data, is a tightly-scripted process defined by the rules of statistics. Earlier treatments of SEM tend to reinforce this impression. This is perhaps true of my own writing (e.g. comparing the presentations in Grace 2006 to the current paper), but also the writing in more general treatments of SEM (Kline 1998, 1st Edition compared to Kline 2016, 5th Edition). Some of the shift in recommendations reflects a broader shift in the view of the role of *p*-values and strict hypothesis tests. Experience with SEM applications to real-world problems also teaches us that the exercise of scientific judgement is an essential part of model evaluation.

The goal of model selection is not simply to describe the relationships in data. In my view, the goal is to balance twin objectives. First is the narrow task of evaluating the model-data inter-relationships. We must address the specific question, “What do these data say about

the hypothesis?" SEM philosophy, however, imagines a sequence of studies and a process of sequential learning (Grace and Irvine 2020). Our initial model for each analysis represents an accumulation of knowledge from prior investigations. The evidence obtained from the analysis then updates our understanding with new information. We should assume, I feel, that there will be additional studies to follow that will test and strengthen our understanding. This looking forward motivates us to a second objective, which is to select a final model in this study that can serve to help us construct the initial model for the next study. These twin objectives motivate us to balance the reliance on empirical evidence in the current data sample against theoretical knowledge about the underlying mechanisms. Here, I describe more specifically the role of the scientist, as this topic is rarely covered in explicit fashion (e.g. Larson and Grace 2004; Grace and Irvine 2020.

Evidence Type #1: A Priori Scientific Knowledge

The initial model construction process in SEM relies heavily on investigator knowledge. The reasoning process adopted during model construction (Grace and Irvine 2020) needs to be maintained during model evaluation as well. It will not always be necessary to explicitly consider alternative models. There is an aspiration in SEM that, through sequential studies, one will reach a confirmatory stage where new data may refine estimates, but not change our minds about the causal structure of the hypothesis. This paper addresses the more typical case where model refinements are proposed and competing models considered.

A Need to Support and Defend Critical Assumptions – There is a certain minimum amount of a priori knowledge that is required in order to use SEM. SE models include so-called "untestable" assumptions, along with assumptions that can be falsified by an appropriate dataset. The most fundamental untestable assumptions are the directions of the arrows. Often, in ecological studies, the investigator is able to justify arrow directionality. More challenging is for investigators to justify things that are omitted from their models. Of course, models must limit their scope to the essential components. There are rules relating to what can and cannot be omitted. First, only include variables that are essential to the modeling objectives. Often scientists attempt to include all measured variables and produce models whose complexity exceed the capacity of the data sample size (too many parameters per sample). Second, omitting variables that have strong effects on two or more of the variables of primary interest can lead to confounding. When very strong confounding relationships are omitted, they have the potential to bias conclusions.

For links not included in the initial model, post-estimation evaluations should reveal whether these assumptions need to be reconsidered. For direct links, these are straightforward to detect and include. However, we should not ignore the possibility that errors will be non-independent. Correlated errors are definite indications of omitted confounders. Finding error correlations may spark a need to consider model modifications. It is wise to consider how one might interpret such findings based on a priori knowledge, as these represent factors being omitted from explicit inclusion in the initial model.

Investigator's Opinion of the Strength of Theory versus Strength of Data – A point that is rarely discussed outside of the sphere of Bayesian statistics is how to weight a priori scientific knowledge against the information content of a data set. This omission exists despite the fact that, in many ecological studies, analyses are based on very limited samples. In cases where there is a large and representative dataset for use in an analysis, one should be prepared to consider the data-derived estimates as the final arbiter for drawing inferences. In some cases, our a priori knowledge may be stronger than the dataset available for modeling.

A broad view of the quantitative sciences must recognise that we aspire for our models to transition over time from assumption-testing to assumption-based. Numerous sub-disciplines within ecology rely on assumption-based models. So-called 'mechanistic models' incorporate processes that operate on biological systems with enough regularity that the form of the model is accepted as given and data are used purely to estimate the parameters. Population models often fall into this group. For this model type, some of the processes may be of known functional form, while others may be of unknown form. When studying multi-species ecological communities, we often encounter mechanisms that are contingent on so many factors that relationship forms cannot be taken as given (e.g. effects of species additions or removals; Smith and Knapp 2003).

The preceding material is presented to make two important points relevant to model evaluation and selection within SEM. First, when alternative models are suggested, based on empirical results, we should avoid constructing alternative models for consideration that we know are false representations of the system. Perhaps a birth rate estimate is low and its 95% confidence interval includes a value of zero. Do we prune the model to adhere to the principle of parsimony? That would mean we might end up presenting a final model that, by omission, suggests that births are not a contributing factor for population size. Scientific logic would suggest that we should not prune in this case, but what are the consequences? I will address this question in the context of our illustrative example in the section below. In summary, the rule I would suggest is for the investigator to not include models in your comparison set that you, as a scientist, are not willing to defend.

Evidence Type #2: *P*-values

Historically, the use of *p*-values came into widespread use as a part of the machinery for null-hypothesis testing. *P*-values have traditionally been used for dichotomous decision-making and have been central to the practical application of statistical methods. Within a WOE paradigm, *p*-values can be used as continuous quantitative indicators without their being used as strict cutoffs. The case for this type of use of *p*-values has been recently articulated by McShane et al. (2019) who argue, "We propose that the *p*-value be demoted from its threshold screening role and instead, treated continuously, be considered ... as just one among many pieces of evidence."

As mentioned in the Introduction, for SEM applications that utilise global estimation methods (e.g. LISREL, Mplus, AMOS or lavaan), model estimation returns an initial set of measures that quantify the overall correspondence between observed and model-implied

covariances. Immediate focus is directed to the X^2 statistic, summarising model-data fit and its *p*-value. The *p*-value, in this case, has a counter-intuitive meaning in that it represents the probability that the observed data deviate from model expectations, which would imply the model structure is inappropriate for obtaining estimates from the data. When conducting SEM, there is an immediate decision that has to be made after the initial model is estimated, which is to decide whether there are one or more important links omitted. If there are, the reported parameter estimates are not to be trusted. This need to make sure the network is not missing important links has led to a history of reliance on dichotomous decision-making and this is perhaps unavoidable. The next section will discuss alternative indices for global fit assessment, but regardless of the fit index used, the question still arises, “Should we keep the original model or create a more complete one before we proceed?”

Evidence Type #3: Global Fit Judged by Approximate Fit Indices

There are a number of factors that limit our ability to provide an omnibus model evaluation using the X^2 statistic. For example, the continuous increase in statistical power with increasing *N* (sample size) of course means appreciable differences between data and model could be missed when *N* is small, but trivial differences could be flagged when *N* is large. Additionally, it is well documented that the X^2 statistic can hide various types of mis-specification simply because it is a summary statistic for the entire model (Steiger 2007). As a result of these problems, a great many alternative measures of global model fit have been proposed and evaluated.

Kline (2016) provides an overview of the approximate fit indices developed for SEM. Most of these are not proper for use in significance tests, but continuous measures of model-data correspondence. That said, the urge to perform model sufficiency testing has led to various attempts to create thresholds for approximate fit indices (e.g. Hu and Bentler 1999). Kline (2016) places approximate fit indices into three primary categories, (1) absolute fit, (2) comparative fit and (3) parsimony-adjusted. Information metrics like *AIC*, which are technically model comparison measures, are discussed as a separate topic in a later section.

Regarding approximate fit indices, it is probably true that some are, on average, better than others. However, simulation studies indicate that the capacity to detect mis-specifications based on recommended thresholds depends on the particular mis-specification (Marsh et al. 2004). Some authors have even suggested such indices and associated thresholds not be used at all (Barrett 2007), while other suggest they do have a role to play (Mulaic 2009). The current practice amongst SEM users, particularly those in the social sciences using latent variable models, has come to be pluralistic and individualistic, with each investigator potentially relying on a unique combination of pieces of evidence. The fundamental problem with the approaches considered is that there is no clear way to combine the different types of evidence.

Under the WOE approach, I will demonstrate in this paper approximate fit indices can be useful measures to report. Kline (2016) suggests the following should be reported: (a) the

Root Mean Square Error of Approximation (RMSEA) and its 90% confidence interval, (b) the Comparative Fit Index (CFI) and (c) the Standardized Root Mean Square Residual (SRMR). Simulation studies have shown that none of these indices is sufficient for detecting all types of mis-specifications. Each provides some quantification of evidence nonetheless and a description of how they are judged is presented below.

Evidence Type #4: Modification Indices and Residual Relationships

The second phase in evaluating models after first examining global fit measures is often to search for indications of what changes could be made to improve model-data concordance. Specially designed for this purpose are so-called modification indices (MI). All global-estimation-based software packages, with which I am familiar, report this information upon request. The critical role of the investigator's scientific judgement comes into sharp focus once one tries to make sense of the MI table provided for a model that is mis-specified.

MI values are expressed in terms of the drop in the X^2 statistic that would be observed if a link were added to the model. The categories of possible additions include, (a) regressions, (b) latent variable loadings, (c) error correlations and (d) variance constraints. The modification indices are not arrived at by actually fitting alternative models. Rather, they are approximations and therefore do not always correspond to the changes that will be observed.

Perhaps the best way to gain some intuition about the challenge MI values attempt to overcome is to look at the raw materials for computing evidence of mis-specification, which are the residuals. In this case, the residuals are not those between predicted and observed individual data values, but instead, between the observed and model-implied variance-covariance matrices. Requesting to see residuals in a standardised metric will illustrate where model-implied and observed matrices are most discrepant. Because As the parts of a model are intercorrelated, there are many different model changes that might be implied. From a very practical standpoint, the investigator must realise that any change in the model can potentially resolve many of the listed modification possibilities. Therefore, one should decide on a single addition to the model before re-estimation. It is essential that the chosen modifications make substantive sense, because they must explain to the reviewers the scientific basis for the modifications of their initial model (see Grace 2006 for an in-depth discussion of this issue).

When working with models having latent variables with multiple indicators, MI tables sometimes return no usable advice, even though global model fit is poor. In this case, it becomes essential to consult the residual matrix unless theoretically predefined alternative specifications are available. While matrices of residuals provide a more undistilled source of information, they are commonly a fundamental source of evidence for selecting alternative models for consideration.

Evidence Type #5: Information Measures - *AIC* and *BIC*

As with all other types of fit measures, information-based measures have a long history of use in SEM. This is a complex topic that I will treat lightly because the jury is still out on whether universally-applicable recommendations are even possible. Also Complicating things is also the sheer variety of information metrics that have been proposed for use. Fortunately, a recent review of past studies and set of simulation studies by Lin et al. (2017) provide a basis for summarising the topic for SEM users.

Two types of information measures have captured most of the recommendations, *AIC*-type indices and *BIC*-type indices. The Akaike Information Criterion (*AIC*) was proposed for use in 1974 (Akaike 1974), while Schwarz (1978) proposed a Bayesian alternative (*BIC*). In Burnham and Anderson (2002), the foundational arguments were laid out for using *AIC* within a multimodel inference system as a replacement for *p*-value-based null hypothesis testing. Much discussion of this index and comparisons with other approaches, including *BIC*, have taken place and a concise summary of the discussion can be found in Brewer et al. 2016) and Aho et al. 2014). The result of all this attention has led to widespread adoption of information-based multimodel comparisons in the natural sciences. The separate history of discussion of the same issues amongst SEM practitioners has led to a distinct body of literature where practices and recommendations vary widely. One lack of overlap relates to the sample-size corrected version of *AIC*, known as *AICc*. Since it is not theoretically justified for multivariate models, it has not been included in SEM studies. For univariate models, it has been shown to outperform *AIC* under small sample sizes and is the default index for many studies in the natural sciences.

In their book, Burnham and Anderson (2004) suggest models separated by more than 2 AIC units could be seen as distinct, while later (Burnham et al. 2011), they suggest 4-7 units might be a better criterion. As I will show below, multimodel inference should use the full set of models evaluated for summarising the evidence for any one of the set.

In this paper, I do not wish to attempt to propose a definitive answer to the question of which information index is best nor consider the detailed studies and arguments association with that question. My intent is to show how the various types of evidence, associated with SEM, can be used to build up a set of candidate models for comparison and how information measures can be used to assist in the comparisons. For that purpose, I will rely on the following synoptic view, taken from various sources.

The relative performance of different information measures has been shown to vary with a number of factors, including (a) sample size, (b) the composition of candidate model sets, (c) the magnitude of effects to be detected and (d) heterogeneity in the data. I will try to summarise our current understanding (and my own experience) in a few summary statements:

1. The behaviour of *AIC* versus *BIC* indices can be anticipated by the fact that the latter group imposes stronger penalties for model complexity. Variant types of *AIC* (e.g. *Consistent AIC - CAIC*) and of *BIC* (e.g. Haughton's BIC - *HBIC* and sample-

size adjusted BIC -*ABIC*) all have different types of behaviour. The performance of different indices depends strongly on the above factors. There is no single indicator that is superior across all assumptions and conditions.

2. There are two key questions for the investigator to consider that influence the guidance to take home from simulation studies. First, is the true data-generating process complex and with tapering effect sizes? Second, is it likely that not all the important variables in the true model are in your candidate models?
3. If you answer yes to both questions in number 2, *AIC* and *ABIC* are perhaps the best choices up to $N = 400$. Above that, *HBIC* is a good choice.
4. If your answer to only the first question is yes, *AIC* remains a consistent performer up to $N = 300$, but *ABIC* is not as consistent.

Evidence Type #6: d-separation Tests

In this paper, my focus is on globally-estimated models where the investigator must contend with multiple forms of evidence encountered within a sequential evaluation of overall model fit and individual links included in the model. However, many investigators use local estimation methods (e.g. Lefcheck 2016). For this reason, I briefly describe methods for d-separation (d-sep) testing here. Further, Kline (2016) (Chapter 11) discusses the potential for including evidence from local fit indicators, such as d-sep tests, in the evaluation of globally-fit models.

Pearl's redescription of SEM in foundational terms, referred to as the Structural Causal Model (Pearl 2000), includes the proposal that the testable implications of models can be expressed in terms of d-separation tests. Shipley (2000) subsequently developed formal d-sep tests that could be used for empirical evaluation of conditional independence claims in recursive path models. Initially, Shipley's test statistic was based on *p*-values, which were used in conjunction with the testing of individual independence claims. They were also used in developing an overall test statistic, the *C* score, which is a function of the sum of *p*-values across the entire set of independence claims evaluated.

In 2013, Shipley subsequently developed a version of his method based on *AIC* (*Shipley 2013*). He demonstrated that, for a set of equations whose parameters are estimated using maximum likelihood, the *C* statistic equates to a maximum likelihood quantity as long as null probabilities, from which a *C* statistic is computed, are maximum-likelihood probabilities. Fulfilling this requirement makes it possible to compute *AIC* from a *C* score. He went on to show how to apply the *AIC* statistic to model comparisons based on d-sep criteria.

Most recently, Shipley and Douma (2019) have revisited the use of *AIC* in conjunction with locally-estimated model comparisons. They pointed out that the "*d-sep AIC*", which is the name they suggested for their original computation, only captures evidence contained within the conditional independence tests and not within the entire model. To remedy this limitation, they developed a "*full-model AIC*" that takes into account both the causal topology of the model and the parameter estimates.

Proposed Sequence for a Weight-of-Evidence Approach to Model Evaluation and Selection

The use of the above-described types of evidence is illustrated next. To assist in the process, I provide a sequence of steps for a weight-of-evidence approach.

1. Consideration of Sample Size - The first piece of evidence to consider is the number of samples, *N*. Sample size has a huge influence on model evaluation. One can anticipate a great deal about the value of various model fit indices based on sample size. A first-cut distinction is often made between studies depending on whether *N* is less than or greater than 200. Other sample size distinctions may be informative to the investigator based on their past experience, as well as other factors such as model complexity and model specification details.

2. Examination of Model X^2 Statistic, Model Degrees of Freedom and associated *P*-value – Below a sample size of 200, the X^2 statistic and associated *p*-value are informative. That said, the criterion of $p \geq 0.05$ is known to be imprecise. When $N < 200$, a *p*-value falling below the 0.05 threshold is a strong indication that one should consider alternative models that include additional links. Above the 0.05 criterion, increasing values of *p* provide increasing support for the presumption that the model under consideration is not leaving out links representing important processes. It is common to find support for adding links when $p > 0.05$. Based on personal experience, we might think of some much larger value, such as 0.50, as a point at which any omitted links are likely weak, but additional pieces of information are nearly always worth examining unless *p* is quite high. Above a sample size of 200, one can expect to need to use Approximate Fit Indices to convince reviewers that the model is not leaving out important links and thus suitable for interpretation.

3. Examination of Select Approximate Fit Indices – As mentioned above, Kline (2016) suggests reporting certain Approximate Fit Indices.

a. The Root Mean Square Error of Approximation (*RMSEA*) is appealing because it is accompanied by upper and lower values for a 90% confidence interval. The *RMSEA* differs from the X^2 statistic as it measures departures from approximate/close fit instead of perfect fit. "Approximate" or "close" fit is described as the situation where the $X^2 \leq$ df, while perfect fit is where $X^2 = 0$. The *RMSEA* is scaled as a "badness of fit" measure, where 0 means perfect fit. The index is known to be a poor decision criterion by itself. That said, when the lower confidence interval is 0, it is supposed to mean that approximate fit is within the range of support offered by the data. When the associated *p*-value is less than 0.05, it suggests some type of misfit (due to omitted linkage).

b. The Comparative Fit Index (*CFI*) compares the departure from close fit (just described) to what we would find for a null model (all parameters = 0.0). Its values are scaled to range from 1.0 to 0. It is widely reported by investigators, especially when $N > 200$, because it is completely unaffected by sample size. It is not a reliable index for model selection by itself (though the unscaled version, *RNI* – the Relative Noncentrality Index) has been shown to

perform as well as *AIC* by Bollen et al. (2014). Hu and Bentler (1999) have suggested a criterion of $CFI \geq 0.95$, though this is probably overly restrictive when $N < 200$ because of the limited power to detect effects.

c. The Standardized Root Mean Square Residual (*SRMR*) is computed as the absolute value of the mean of the standardised residual covariances. A value of 0 means perfect fit and a value > 0.10 is of concern. To interpret this measure appropriately, it is best to examine the matrix of standardised residual covariances since adequate mean fit may hide specific deviations representing important mis-specifications.

4. Examination of Modification Indices and Covariance Residuals – The matrix of covariance residuals provides the raw materials used to compute overall model fit, as well as modification indices. Their examination is sometimes an important supplement to the MI table. The MI values themselves are expressed in terms of the drop in the X^2 statistic that would be expected if a link were added to a model. The categories of possible additions include, (a) regressions, (b) latent variable loadings, (c) error correlations and (d) the relaxation of constraints (e.g. if any parameters have been given fixed values). As any given covariance residual can be resolved in several different ways, it is important to consider the scientific interpretability before examining the MI table. For example, in this paper we are not considering latent variable models. For this reason, we may want to request only a subset of the MI types, such as direct effects (~) and possibly error correlations (~~). The single-degree-of-freedom X^2 criterion value of 3.84 is often used to provide information on the interpretation of MI values. The use of this information is considered in the next section.

5. Considering Alternative Models Containing Additional Linkages/Parameters – Unless overall model fit is very close, it is desirable to consider alternative models. Failing to find a credible alternative model is itself a form of support for the model under consideration. It is critical that all alternative models estimated be scientifically plausible. The entire foundation for multimodel comparisons is based on the premise that all models compared are scientifically defensible. One simple tip is to consider in advance the signs (positive or negative) of links added to a model. At the point of examining modification indices, it is possible to not only examine the magnitude of the MI, but also the sign of the expected parameter. Interpretation of added coefficients can be very dependent on whether the sign of the coefficient corresponds to the type of mechanisms suggested. Sometimes, the expected parameter change (“epc”) is of opposite sign to the mechanistic interpretation for adding a link and this may influence how the investigator proceeds. When considering MI values in terms of their magnitudes, the investigator should recognise that it is only profitable to make one addition to a model at a time. Making a single change to a model can lead to a completely different set of MI values for the remaining omitted links under consideration. It is common to focus on the suggested modifications with the largest MI values, though the interpretability of suggested changes is of paramount importance. The investigator will usually notice several possible modifications that are projected to have identical effects on model fit. I will elaborate the logic to employ in such a situation in the context of our empirical example.

6. Repeat of Steps 1 – 5 as needed – When a link or additional parameter is added to form a new model, the steps just described will need to be repeated. At this point in the process, we are ignoring the second-level question of whether the included links (freely estimated parameters) in a model are supported. The inclusion of weak or non-supported effects in models has only a minor effect on overall model fit because they always reduce the raw discrepancy and only impact the model X^2 statistic by increasing the number of parameters being estimated (K). As omitted links can have large effects on model fit and on estimated values for the other parameters, I always recommend addressing the question of whether important processes are omitted from the model before worrying about simplification.

7. Model Simplification – The model simplification process addresses the question of whether there is empirical support for all of the included linkages. The reported statistics for individual parameters now come into play. Historically, the *p*-values, associated with individual parameters, have been used as guideposts, but not for deciding whether a parameter should be set to zero. It has long been the practice within SEM under global estimation to use *p*-values as a continuous quantitative measure of the variability associated with a parameter estimate. Finding a parameter estimate with a *p*-value at or above 0.05 raises the question as to what magnitude of change would be seen in global model fit if that parameter were set to zero (i.e. removing a link). The standard approach has been to use a single-degree-of-freedom X^2 test to judge whether a link should be removed. Under a WOE paradigm, *p*-values at or above 0.05 suggest the construction of alternative models to evaluate, using multi-model comparison.

8. Selection of Candidate Models for Comparison – The approach, described thus far, encourages a liberal consideration of alternative models, under the provision that they represent valid competing scientific explanations. A single model for scientific inference purposes is to be selected from the set. The suggestion of model averaging, proposed by Burnham and Anderson (2002), has been shown to be inappropriate for interpretive science based on both statistical (Cade 2015) and scientific (Grace and Irvine 2020) grounds.

9. Model Comparison, Weighing of Evidence and Model Selection – Model selection, based on multimodel comparisons using information criteria, as championed by Burnham and Anderson (2002), provides an appealing framework for use in a WOE process. A detailed description of multimodel comparison is not provided here, but its use will be illustrated in the next section. In addition to the quantities mentioned in this section are the types of scientific expert judgement described previously. Since *AIC* difference categories have been described in terms of the concept of model equivalency, the framework readily integrates different forms of evidence and allows the investigator to judge the consequences of selecting any model other than the one with the lowest metric value.

An Ecological Example

To illustrate the ideas presented in this paper, I will rely on an example related to the biological control of invasive plants. The invasive plant *Euphoria esula* (leafy spurge) is considered a threat to the ecological and economic integrity of grasslands in the north-central United States. The example, presented here, is derived from a study conducted at the Theodore Roosevelt National Park in North Dakota (Larson and Grace 2004). Following establishment of spurge in the Park in the 1970s, a biocontrol programme was initiated in the 1980s. Prior to the study described here, there were more than 1,800 releases of two flea beetle species, both of which are obligate feeders on spurge, throughout the Park. In 1999, permanent plots were established for the monitoring of plant and beetle density dynamics.

In this paper, I have used the results published by Larson and Grace (2004) to demonstrate a WOE approach. In order to base illustrations on a “known” situation, the published SEM model was adopted as the “true” model and published parameter estimates were used to simulate a large sample (N = 10,000), which was converted to a variance-covariance matrix. The development of a large-sample covariance matrix allows me to use a single representative dataset and then specify a real-world sample size for illustrative purposes. A WOE approach to model evaluation is illustrated below by starting with a naïve, initial model similar to the one used by the investigators in the original study.

Fig. 1 represents the initial hypothesis used in this illustration. Table 1 provides summary information related to the mechanisms encoded in the figure. Knowledge of plant biology suggested that the change in stems between year0 and year1 would depend on the initial stem density in year0 (Fig. 1, link 1). Since areas where the invasive plant had already established were the object of study, the possibility of self-thinning (negative density dependence) was considered most likely. A bit of information about the life-cycle of the flea beetles helps to explain other parts of the proposed model. The flea beetles feed exclusively on spurge and live on the plant except when they disperse as adults. In the autumn, females lay eggs at, or just below, the soil surface, near the bases of stems. Newly-hatched larvae burrow into the soil and begin feeding on very small plant roots. Larvae feed on progressively larger roots and root buds as they develop. After the larvae overwinter, they resume feeding on plant roots until they pupate in late spring or early summer. Once the larvae pupate, the adult flea beetles emerge from the soil and feed on the plant’s foliage and flowers throughout the growing season, dispersing then as adults. Based on the life cycle, the investigators hypothesised that the numbers of adult beetles should be greatest in plots with the highest numbers of plant stems in the current year due to the resources provided to the adults (Fig. 1, links 2 and 7). It was also hypothesised by the investigators? that beetle population densities could be related to the food supply provided to larvae in the preceding winter, representing a lag effect (Fig. 1, links 5 and 10). Some degree of year-to-year spatial fidelity was expected (Fig. 1, links 3 and 8), while a potential for competitive effects between beetle species was considered (Fig. 1, links 6 and 11). Finally, if biocontrol agents were effective at controlling plant populations, it was

hypothesised that plant densities would decline over time faster where beetle densities were greatest (Fig. 1, links 4 and 9).

Table 1.

Description of ecological linkages numbered in Fig. 1

Link #	Description of potential mechanisms and expected sign of effect
1	Change in stem density is expected to depend on initial density of stems. A positive parameter estimate would indicate positive density dependence, while a negative parameter estimate would indicate negative density dependence (negative effect).
2	Dependence of flea beetle density on plant stem density for *BioA*. (positive effect)
3	Site fidelity for *BioA*. (positive effect)
4	Effect of *BioA* on *StemChg* (expecting a negative effect, if control agent is effective)
5	Lag food effect on *BioA*. (positive effect)
6	Competitive effect of *BioA* on *BioB*. (negative effect)
7	Dependence of flea beetle density on plant stem density for *BioB*. (positive effect)
8	Site fidelity for *BioB*. (positive effect)
9	Effect of *BioB* on *StemChg* (expecting a negative effect, if control agent is effective)
10	Lag food effect on *BioB*. (positive effect)
11	Competitive effect of *BioB* on *BioA*. (negative effect)

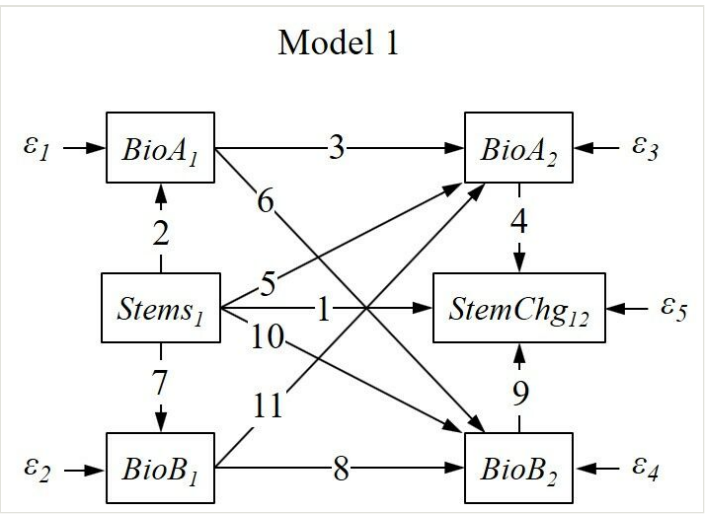

Figure 1.

Structural equation model representing an initial hypothesis regarding the potential effects of two biocontrol flea beetles, *BioA* (*Aphthona nigriscutus*) and *BioB* (*Aphthona lacertosa*), on an invasive plant (*Euphorbia esula*). Stem and beetle densities were measured over two years and the change in stem density between years 1 and 2 computed (*StemChg12*).

For the illustration below, I simulated 10,000 replicates using the lavaan package simulateData function and then captured the covariance matrix as input for illustrations. The illustrations in this paper assume a sample size of 150 (original investigation was based on 165 samples). Code used for data simulation is in the supplementary materials (Suppl. material 1).

Illustration of a WOE Approach using the Biocontrol Example

In this section, I follow the Proposed Sequence described above to illustrate its application. In this example, I start with a single proposed a priori model and work from there. In other cases, we might have multiple candidate models from the beginning to evaluate, which would modify the sequence slightly.

Model 1

Model 1 (Fig. 1) provides our starting point for the selection of a explanation for the data. Covariance matrix input and Model 1 code are in Table 2.

Table 2.

Code for estimating Model 1 (Fig. 1) using lavaan.

### load libraries library(lavaan)					
########## Simulation Study #1 ########## # Covariance matrix for input					
sim.cov <- '					
1.2472					
-0.1492	1.019				
0.8442	-0.178	1.6417			
0.0358	0.704	0.0357	1.5512		
-0.2922	-0.233	-0.1217	-0.5276	1.488	
0.5235	0.149	0.5335	0.3277	-0.655	1.029'

# Convert matrix and name variables 	
sim.cov.dat <- getCov(sim.cov, names = c("BioA1", "BioB1", "BioA2",
"BioB2", "StemChg12", "Stems1"))					

##### Scenario 1 - Set N=150 ##### ### mod1 – initial model					
mod1 <- ' # regressions BioA1 ~ Stems1 BioB1 ~ Stems1 BioA2 ~ BioA1 +Stems1 +BioB1 BioB2 ~ BioB1 +Stems1 +BioA1 StemChg12 ~ Stems1 +BioA2 +BioB2 '					

```
# Estimate model ‘mod1’ using data matrix ‘sim.cov.dat’, N = 150
Mod1.fit <- sem(mod1, sample.cov = sim.cov.dat, sample.nos = 150)
```

Table 3.

Global fit statistics obtained for Model 1.

> summary(mod1.fit, fit.measures=T)	
lavaan 0.6-5 ended normally after 15 iterations	
Estimator	ML
Optimisation method	NLMINB
Number of free parameters	16
Number of observations	150
Model Test User Model:	
Test statistic	9.125
Degrees of freedom	4
P-value (Chi-square)	0.058
Model Test Baseline Model:	
Test statistic	247.786
Degrees of freedom	15
P-value	0.000
User Model versus Baseline Model:	
Comparative Fit Index (CFI)	0.978
Tucker-Lewis Index (TLI)	0.917
Loglikelihood and Information Criteria:	
Loglikelihood user model (H0)	-1274.742
Loglikelihood unrestricted model (H1)	-1270.179
Akaike (AIC)	2581.483
Bayesian (BIC)	2629.654
Sample-size adjusted Bayesian (BIC)	2579.017
Root Mean Square Error of Approximation:	
RMSEA	0.092
90 Percent confidence interval - lower	0.000
90 Percent confidence interval - upper	0.173
P-value RMSEA <= 0.05	0.153
Standardized Root Mean Square Residual:	
SRMR	0.055

Estimation of Model 1, using lavaan, returns the measures of overall model fit shown in Table 3. We can see that lavaan converged rapidly and without warnings. My discussion of results will follow the proposed sequence of examinations described above.

1. Consideration of Sample Size – Since N = 150, all of the fit measures presented can be interpreted in a fairly straightforward fashion.

2. Examination of Model X^2 Statistic, Model Degrees of Freedom and associated *P*-value –While the *p*-value returned is above the criterion of 0.05, it can be anticipated at a sample size of 150 that we have sufficient statistical power to detect modest effect sizes. For this example, it is appropriate to assume that the true model contains a wide range of effect strengths, including some of scientific interest, but of modest size. With a *p*-value of 0.058, it would be naïve to simply accept this model without looking any further.

3. Examination of Select Approximate Fit Indices –

a. *RMSEA* is below 0.10, the lower CI boundary is 0.0 and associated *p*-value is 0.153. All these values suggest approximate fit.

b. *CFI* is estimated to be 0.978, which implies that we are probably not missing any large effects (assuming no interactions, as I do for this illustration).

c. *SRMR* is 0.055, well below the warning level of 0.10.

d. Summary: This evidence suggests that when model-data discrepancies are average across the whole model, there is reasonable correspondence. However, this level of fit could be the result of averaging many equal-sized minor mis-specifications OR averaging many areas of the model with very close fit and a few important model-data mismatches. The limitation of global fit measures is their inability to distinguish these two possibilities.

Examination of Modification Indices and Covariance Residuals – For now, I forego presenting a matrix of standardised residual covariances as their interpretation requires some experience. My typical process is to first examine modification indices and work with those, resorting to an examination of residuals, if it seems necessary. Seven modification suggestions are returned in this case (Table 4). Immediately one recognises that the MI value is identical for all of the suggested changes. All the suggested changes involve the densities of species A, species B or both at time 1. In fact, changes 1, 2 and 5 suggest adding a link between *BioA1* and *BioB1*. Changes 3, 4, 6 and 7 are suggestions to add effects pointing from time 2 to time 1, which are not scientifically plausible. So, the modifications suggest we consider what kind of process could jointly influence the abundances of species A and B at time 1 (other than dependence on stem density). The expected parameter change values for the set of changes under consideration are all negative. This means we should be thinking of processes that could cause a negative association between the two species at the beginning of the study. The investigators imagined several, including (a) different establishment histories for the two biocontrol agents within the landscape sampled, (b) different habitat preferences or (c) some previous interaction between the species, such as competition. All of these mechanisms would best

be represented by an error correlation between *BioA1* and *BioB1*. This alternative model (Model 2) is shown in Fig. 2 and was fitted to the data for further evaluation (step 5).

Table 4.

Modification indices for Model 1.

> # Modification Indices
> subset(modindices(mod1.fit), mi > 2, c("lhs", "op", "rhs", "mi", "epc"))

lhs	op	rhs	mi	epc
1 BioA1	~~	BioB1	7.762	-0.224
2 BioA1	~	BioB1	7.762	-0.226
3 BioA1	~	BioA2	7.762	1.723
4 BioA1	~	BioB2	7.762	-0.341
5 BioB1	~	BioA1	7.762	-0.229
6 BioB1	~	BioA2	7.762	-0.414
7 BioB1	~	BioB2	7.762	-12.411

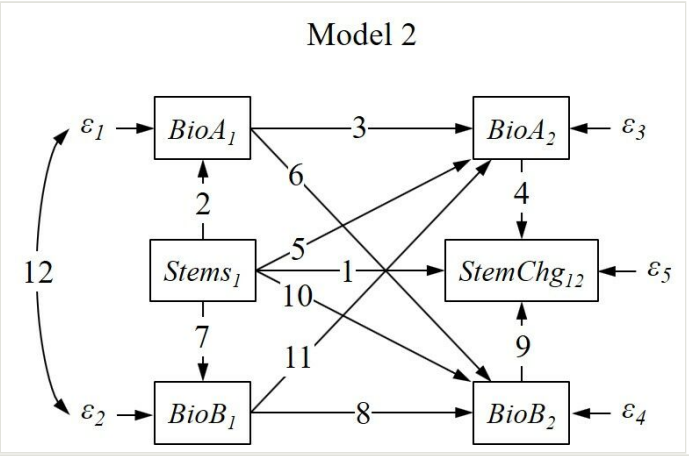

Figure 2.

Model 2, which includes an error correlation between species A and B at time 1 (parameter 12).

Model 2

Fig. 2 represents our new model for evaluation. The modified lavaan codes is shown in Table 5. To consider the evidence in support of this model, we must estimate the new model and repeat the proposed sequence of steps. Less explanation is needed for the second model evaluation, so I condense steps.

Table 5.

Code for estimating Model 2. Parameters are now labelled b1-12.

```
### Model 2 (parameter numbers correspond to those in Fig. 5)
mod2 <- '
# regressions
BioA1 ~ b2*Stems1
BioB1 ~ b7*Stems1
BioA2 ~ b3*BioA1 +b5*Stems1 +b11*BioB1
BioB2 ~ b8*BioB1 +b10*Stems1 +b6*BioA1
StemChg12 ~ b1*Stems1 +b4*BioA2 +b9*BioB2
# error covariance
BioA1 ~~ b12*BioB1'
```

Steps 2-5: Examination of Global Fit Measures and Consideration of Additions to Model 2 – The X^2 dropped from 9.125 to 1.155 (a decline of 8.03) and its *p*-value rose from 0.058 to 0.764 (Table 6). The *RMSEA* estimate is now 0.0 and *CFI* is estimated at 1.0, while *SRMR* shrunk from 0.055 to 0.013. Since the new X^2 is well below 3.84, there is not enough remaining model-data discrepancy to justify any further additions at this point. Therefore, there is no reason to examine modification indices for this revised model. It is worth noting that *AIC* dropped from 2629.654 to 2575.513, a decline of 54.141, a decisive magnitude of improvement.

Table 6.

Global fit statistics obtained for Model 2.

> summary(mod2.fit, fit.measures=T)	
lavaan 0.6-3 ended normally after 16 iterations	
Optimisation method	NLMINB
Number of free parameters	17
Number of observations	150
Estimator	ML
Model Fit Test Statistic	1.155
Degrees of freedom	3
P-value (Chi-square)	0.764
Model test baseline model:	
Minimum Function Test Statistic	247.786
Degrees of freedom	15
P-value	0.000
User model versus baseline model:	
Comparative Fit Index (CFI)	1.000
Tucker-Lewis Index (TLI)	1.040
Loglikelihood and Information Criteria:	

Loglikelihood user model (H0)	-1270.757
Loglikelihood unrestricted model (H1)	-1270.179
Number of free parameters	17
Akaike (AIC)	2575.513
Bayesian (BIC)	2626.694
Sample-size adjusted Bayesian (BIC)	2572.892
Root Mean Square Error of Approximation:	
RMSEA	0.000
90 Percent Confidence Interval	0.000 0.093
P-value RMSEA <= 0.05	0.848
Standardized Root Mean Square Residual:	
SRMR	0.013

Step 7: Considering Simplification for Model 2 – We now turn to judging support for the links included. The information reported by lavaan that is most helpful at this point are the statistics associated with individual parameters. These are presented in Table 7. There are 18 parameters listed, one of which is treated as fixed (the variance of Stems1). Stems1 is our only exogenous variable. In lavaan, exogenous variances are assumed to be those in the observed variance-covariance matrix by default, though that default can be relaxed. As for the rest of the parameters, we should, at this point, consider whether the signs of the parameters correspond to hypothesised mechanisms associated with those parameters. Based on Table 1, we expect negative effects for links 4, 6, 9 and 11 and positive effects for links 2, 3, 5, 7, 8 and 10. Link 1 could be either positive or negative, depending on the stage of establishment of the plant population. In this instance, we expect link 1 to capture negative density dependence, but the data will be the final determinant since there is no theoretical guarantee. The coefficients in Table 7 do not all conform to expectations. The most notable exception is for link 4 in Model 2. Parameter *b4* is positive and its *p*-value suggests (based on experience) some degree of support. Importantly, link 4 is one of the parameters in our model of greatest interest relative to the question of whether biocontrol is being successful. The same parameter for the other biocontrol species (*b9*) is seen to be negative with a *p*-value of 0.002, conforming to expectations. At this point, we must decide how to proceed and the investigators, involved in the original study, chose to create a new model for further examinations.

Table 7.

Parameter-specific statistics for Model 2.

> # For examination of individual parameter support
> parameterEstimates(mod2.fit)

	lhs	op	rhs	label	est	se	z	pvalue	ci.low	ci.up
1	BioA1	~	Stems1	b2	0.509	0.080	6.382	0.000	0.353	0.665
2	BioB1	~	Stems1	b7	0.145	0.080	1.801	0.072	-0.013	0.302

3	BioA2	~	BioA1	b3	0.554	0.085	6.496	0.000	0.387	0.721
4	BioA2	~	Stems1	b5	0.256	0.094	2.718	0.007	0.071	0.440
5	BioA2	~	BioB1	b11	-0.131	0.085	-1.549	0.121	-0.297	0.035
6	BioB2	~	BioB1	b8	0.662	0.085	7.834	0.000	0.497	0.828
7	BioB2	~	Stems1	b10	0.213	0.094	2.266	0.023	0.029	0.397
8	BioB2	~	BioA1	b6	0.018	0.085	0.217	0.828	-0.149	0.186
9	StemChg12	~	Stems1	b1	-0.643	0.091	-7.077	0.000	-0.821	-0.465
10	StemChg12	~	BioA2	b4	0.139	0.069	2.005	0.045	0.003	0.275
11	StemChg12	~	BioB2	b9	-0.208	0.067	-3.078	0.002	-0.340	-0.075
12	BioA1	~~	BioB1	b12	-0.224	0.082	-2.717	0.007	-0.385	-0.062
13	BioA1	~~	BioA1		0.974	0.113	8.660	0.000	0.754	1.195
14	BioB1	~~	BioB1		0.991	0.114	8.660	0.000	0.767	1.215
15	BioA2	~~	BioA2		1.008	0.116	8.660	0.000	0.780	1.236
16	BioB2	~~	BioB2		1.008	0.116	8.660	0.000	0.780	1.236
17	StemChg12	~~	StemChg12		0.968	0.112	8.660	0.000	0.749	1.187
18	Stems1	~~	Stems1		1.022	0.000	NA	NA	1.022	1.022

Model 3

A new model for evaluation is presented in Fig. 3. The lavaan code is given in Table 8.

Table 8.

Code for estimating Model 3.

```
### Model 3 (parameter b4 now estimates an error correlation)
Mod3 <- '
# regressions
BioA1 ~ b2*Stems1
BioB1 ~ b7*Stems1
BioA2 ~ b3*BioA1 +b5*Stems1 +b11*BioB1
BioB2 ~ b8*BioB1 +b10*Stems1 +b6*BioA1
StemChg12 ~ b1*Stems1 +b9*BioB2
# error covariance
BioA1 ~~ b12*BioB1
StemChg12 ~~ b4*BioA2'
```

Steps 2-5: Examination of Global Fit Measures and Consideration of Additions to Model 3 – The fit statistics for Model 3 are similar to those for Model 2, but with even closer fit (full results not presented to economise on space). The X^2 is now 0.133 and *p*-value = 0.988. All other fit statistics suggest there are no other important links missing from this model.

Step 7: Considering Simplification for Model 3 – We now turn to an examination of the parameter-specific statistics for Model 3 (Table 9). For model simplification, we now look at

the parameters with the least support (highest *p*-values). As always, we do not make model modifications we do not wish to defend later. Parameter labels now correspond to the numbered links in Fig. 3. Parameter *b6* has the highest *p*-value, 0.828, which suggests very weak support for that process. This parameter represents an effect of *BioA1* on *BioB2*, which was considered to be an open question at the beginning of the analysis. The a priori hypotheses being considered is that species A has a competitive effect on species B, which would anticipate a negative effect. The returned estimate is near zero and positive. All results suggest we should remove this link from our model, yielding Model 3B (Fig. 4).

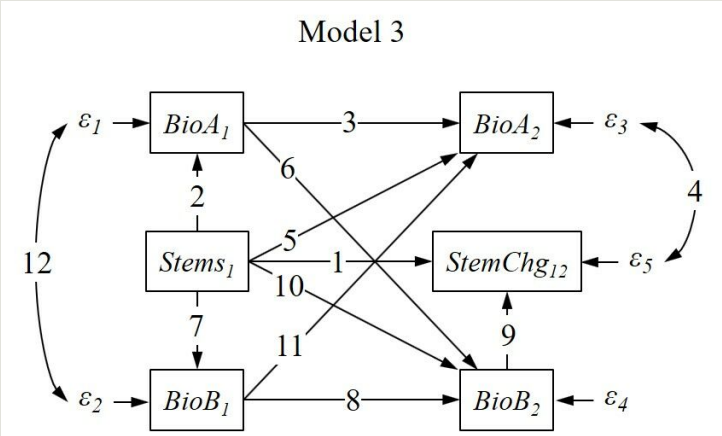

Figure 3.

Model 3, with link from *BioA1* to *BioB2* removed.

Table 9.

Parameter-specific statistics for Model 3.

> # For examination of individual parameter support
> parameterEstimates(mod3.fit)

	lhs	op	rhs	label	est	se	z	pvalue	ci.low	ci.up
1	BioA1	~	Stems1	b2	0.509	0.080	6.382	0.000	0.353	0.665
2	BioB1	~	Stems1	b7	0.145	0.080	1.801	0.072	-0.013	0.302
3	BioA2	~	BioA1	b3	0.551	0.084	6.573	0.000	0.387	0.716
4	BioA2	~	Stems1	b5	0.257	0.094	2.747	0.006	0.074	0.441
5	BioA2	~	BioB1	b11	-0.133	0.084	-1.594	0.111	-0.297	0.031
6	BioB2	~	BioB1	b8	0.662	0.085	7.834	0.000	0.497	0.828
7	BioB2	~	Stems1	b10	0.213	0.094	2.266	0.023	0.029	0.397
8	BioB2	~	BioA1	b6	0.018	0.085	0.217	0.828	-0.149	0.186
9	StemChg12	~	Stems1	b1	-0.565	0.083	-6.786	0.000	-0.729	-0.402
10	Stemchg12	~	BioB2	b9	-0.224	0.067	-3.332	0.001	-0.355	-0.092

11	BioA1	~~	BioB1	b12	-0.224	0.082	-2.717	0.007	-0.385	-0.062
12	BioA2	~~	StemChg12	b4	0.181	0.083	2.184	0.029	0.019	0.344

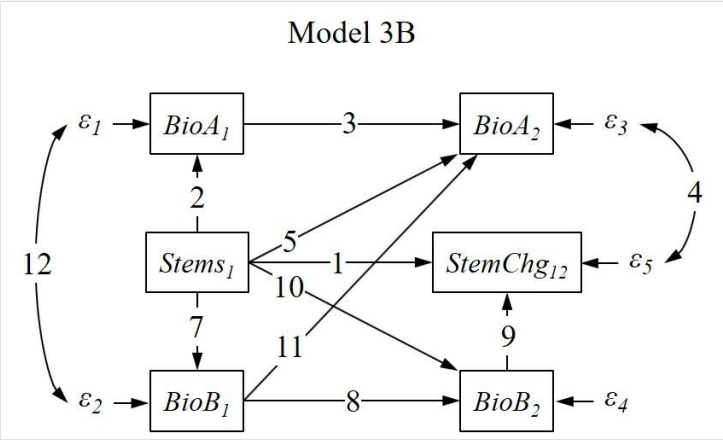

Figure 4.

Model 3B.

Model 3B, Step 7: Considering Simplification for Model 3B – Model 3B was created by removing link 6 from Model 3. This is done by setting parameter *b6* to zero (Table 10). We now re-examine *p*-values for those showing weak support (Table 11). We now see that parameter *b11* has a *p*-value of 0.111. Parameter *b11* corresponds to the link in Fig. 4 from *BioB1* to *BioA2*. As with parameter *b6*, which we already fixed to zero, with link 6 we are looking for evidence of a cross-year competitive effect of species B on species A. It is entirely possible for competition to be asymmetric, so failing to find the effect of species A on B (*b6* set to zero) does not automatically preclude an effect of B on A (*b11*). Still, in this case, setting *b11* to zero seems like the next step, which leads to a model that is simplified further, Model 3C (not shown, but presented in Suppl. material 1).

Table 10.

Code for estimating Model 3B.

```
### Model 3B (parameter b6 now set to zero)

Mod3B <- '
# regressions
BioA1 ~ b2*Stems1
BioB1 ~ b7*Stems1
BioA2 ~ b3*BioA1 +b5*Stems1 +b11*BioB1
BioB2 ~ b8*BioB1 +b10*Stems1 +b6*BioA1
StemChg12 ~ b1*Stems1 +b9*BioB2
# error covariance
BioA1 ~~ b12*BioB1
StemChg12 ~~ b4*BioA2
# set parameters to zero
b6 == 0'
```

Table 11.

Parameter-specific statistics for Model 3B.

> # For examination of individual parameter support
> parameterEstimates(mod3B.fit)

	lhs	op	rhs	label	est	se	z	pvalue	ci.low	ci.up
1	BioA1	~	Stems1	b2	0.509	0.080	6.382	0.000	0.353	0.665
2	BioB1	~	Stems1	b7	0.145	0.080	1.801	0.072	-0.013	0.302
3	BioA2	~	BioA1	b3	0.551	0.084	6.573	0.000	0.387	0.716
4	BioA2	~	Stems1	b5	0.257	0.094	2.747	0.006	0.074	0.441
5	BioA2	~	BioB1	b11	-0.133	0.084	-1.594	0.111	-0.297	0.031
6	BioB2	~	BioB1	b8	0.658	0.082	7.993	0.000	0.497	0.820
7	BioB2	~	Stems1	b10	0.223	0.082	2.723	0.006	0.063	0.384
8	BioB2	~	BioA1	b6	0.000	0.000	NA	NA	0.000	0.000
9	StemChg12	~	Stems1	b1	-0.565	0.083	-6.786	0.000	-0.729	-0.402
10	StemChg12	~	BioB2	b9	-0.224	0.067	-3.332	0.001	-0.355	-0.092
11	BioA1	~~	BioB1	b12	-0.224	0.082	-2.717	0.007	-0.385	-0.062
12	BioA2	~~	StemChg12	b4	0.181	0.083	2.184	0.029	0.019	0.344

Model 3C, Step 7: Considering Simplification for Model 3C – We now direct our attention to Table 11 and parameter *b7*, the effect of *Stems1* on *BioB1*, which has a *p*-value of 0.072. Here, we encounter a parameter with marginal empirical support but with very strong theoretical support. The investigators (Larson and Grace 2004) made the decision to leave this link in all models because knowledge of the biology and evidence from the field, guarantee that the biocontrol beetles depend on the abundance of the invasive plant. Both data and field observations (as well as subsequent studies with longer time courses, Larson et al. (2008)) confirm this process, but contribute to an understanding of how the dynamics of beetles induces substantial variability in the measured association. Later, I will show the consequences of retaining *b7* as a free parameter (or setting it to zero) for other model parameter estimates. This will give the reader a feel for what retaining a weakly-supported link implies. For now, we leave link 7 to be freely estimated and look further at the output. The only remaining parameter with marginal support is *b4*, the error correlation between *BioA2* and *StemChg12* (p = 0.031). This level of support is usually indicative of support for retaining a link. To evaluate this expectation, I estimated one additional model, Model 3D, in which *b4* == 0. Results for Model 4 show no signs of mis-specification and no obvious opportunities for justifiable modifications. This completes our process of creating alternative models.

Step 8: Selection of Candidate Models for Comparison

In this situation, we must decide which of the models that have been estimated are scientifically defensible. Model 1 was found to be obviously mis-specified due to lack of

empirical support and is not a contender for model selection. Model 2, while exhibiting a close model-data fit, was found to lack theoretical support for the originally hypothesised effect of *BioA2* on *StemChg12*. Model 3 was created to solve that problem. Subsequent models examined (Models 3B-D) were all attempts to evaluate simpler versions of Model 3 and are all scientifically defensible. The appropriate model comparison set in this case includes all versions of Model 3.

Step 9: Model Comparison, Weighing of Evidence and Model Selection

A model comparison table is presented in Table 12. There are two models that are virtually identical, Models 3B and 3C. This means the choice is up to the scientists to justify. This observed result could be interpreted as support for an effect of *BioB1* on *BioA2* (Model 3B), though certainly any such effect is weak and variable. Table 13 provides details. Parameter *b11* is negative, consistent with theoretical expectations and has a two-tailed *p*-value of 0.111. Biologically, the debated process is potentially important because it may be that one of the species is a superior biocontrol agent and introducing a less effective biocontrol agent that competes with the first mentioned should be considered. The authors of the original study included this link (actual sample size was slightly larger and support slightly stronger, though scientific interest was the final determinant). Results for Model 3B are presented in Table 13.

Table 12.

Model comparison table.

```
> ##### Multimodel Comparisons
> library (AICcmodavg)
> aictab(list(mod3.fit, mod3B.fit, mod3C.fit, mod3D.fit),
+ c("MOD3", "MOD3B","MOD3C","MOD3D"))
```

Model selection based on AICc:

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
MOD3B	16	2576.63	0.00	0.39	0.39	-1270.27
MOD3C	15	2576.64	0.01	0.39	0.77	-1271.53
MOD3D	14	2579.03	2.40	0.12	0.89	-1273.96
MOD3	17	2579.13	2.50	0.11	1.00	-1270.25

Table 13.

Select results for model selected for interpretation, Model 3B.

Model Fit						
Estimator					ML	
Model Fit Test Statistic					0.180	
Degrees of freedom					4	
P-value (Chi-square)					0.996	

Comparative Fit Index (CFI)					1.000	
RMSEA					0.000	
90 Percent Confidence Interval				0.000 0.000		
P-value RMSEA <= 0.05					0.998	
SRMR					0.006	
Regressions:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
BioA1 ~						
Stems1	(b2)	0.509	0.080	6.382	0.000	0.462
BioB1 ~						
Stems1	(b7)	0.145	0.080	1.801	0.072	0.146
BioA2 ~						
BioA1	(b3)	0.551	0.084	6.573	0.000	0.481
Stems1	(b5)	0.257	0.094	2.747	0.006	0.204
BioB1	(b11)	-0.133	0.084	-1.594	0.111	-0.105
BioB2 ~						
BioB1	(b8)	0.658	0.082	7.993	0.000	0.534
Stems1	(b10)	0.223	0.082	2.723	0.006	0.182
BioA1	(b6)	0.000				0.000
StemChg12 ~						
Stems1	(b1)	-0.565	0.083	-6.786	0.000	-0.470
BioB2	(b9)	-0.224	0.067	-3.332	0.001	-0.228
Covariances:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
.BioA1 ~~						
.BioB1	(b12)	-0.224	0.082	-2.717	0.007	-0.227
.BioA2 ~~						
.StmChg12	(b4)	0.181	0.083	2.184	0.029	0.181
R-Square:						
		Estimate				
BioA1		0.214				
BioB1		0.021				
BioA2		0.381				
BioB2		0.346				
StemChg12		0.328				

Further Results for Selected Model, Model 3B

Additional summary results for the selected model are shown in Table 13. Included are both the standardised parameter estimates and the *R-squares*. Often standardised parameter estimates are presented and used for conveying model findings. I refer the reader to the original paper (Larson and Grace 2004), if they have a deeper interest in the example.

A question raised earlier in the paper was about the consequences of retaining a weakly-supported link in a model for the other model parameters. It is known that leaving out an important link can have a major impact on estimated parameter values for the included links. This sensitivity is illustrated by the fact that model X^2 can drop abruptly when a single link is added (we observed a drop of over 8 points when we added link 12 to create Model 2). The elimination of link 6 from Model 3, however, only increased model X^2 by a tiny amount (0.05), which is typical when parameters with little support are removed. More to the point in this paper is the question of what would happen if we were to set the dependence of beetle species B on plant stems (parameter *b7*) to zero? The results from such a change are presented in Table 14. Model fit measures increase noticeably, but overall fit remains good. Comparing standardised parameter estimates to those in Table 13 shows the only non-trivial change is for the parameter set to zero (drops from 0.15 to 0.0). All other parameter estimates are very close to the same values as before.

Table 14.

Results if *b6* in Model 3B were to be set to zero (compare to results in Table 13).

Model Fit						
Estimator					ML	
Model Fit Test Statistic					3.390	
Degrees of freedom					5	
P-value (Chi-square)					0.640	
Comparative Fit Index (CFI)					1.000	
RMSEA					0.000	
90 Percent Confidence Interval				0.000 0.092		
P-value RMSEA <= 0.05					0.793	
SRMR					0.050	
Regressions:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
BioA1 ~						
Stems1	(b2)	0.541	0.078	6.974	0.000	0.485
BioB1 ~						
Stems1	(b7)	0.000	NA			0.000

BioA2 ~						
BioA1	(b3)	0.551	0.084	6.573	0.000	0.481
Stems1	(b5)	0.257	0.093	2.769	0.006	0.201
BioB1	(b11)	-0.133	0.083	-1.610	0.107	-0.104
BioB2 ~						
BioB1	(b8)	0.658	0.081	8.079	0.000	0.541
Stems1	(b10)	0.223	0.081	2.752	0.006	0.184
BioA1	(b6)	0.000				0.000
StemChg12 ~						
Stems1	(b1)	-0.565	0.082	-6.903	0.000	-0.474
BioB2	(b9)	-0.224	0.067	-3.343	0.001	-0.227
Covariances:						
		Estimate	Std.Err	z-value	P(>|z|)	Std.all
.BioA1 ~~						
.BioB1	(b12)	-0.228	0.083	-2.743	0.006	-0.230
.BioA2 ~~						
.StmChg12	(b4)	0.181	0.083	2.184	0.029	0.181
R-Square:						
		Estimate				
BioA1		0.235				
BioB1		0.000				
BioA2		0.397				
BioB2		0.327				
StemChg12		0.316				

Summary Thoughts and Future Directions

In this paper, I describe a way to bring the necessary evaluation of *p*-values into what is ultimately a model comparison process. The advice provided from main-stream SEM, which is very heavily influenced by the study of complex latent variable models, is both exhaustive and exhausting for the ecologist. This paper provides updated advice for practitioners who rely on global estimation software packages for their SEM analyses.

At the present time, the greatest challenge for future studies is to provide defensible advice for the use of information measures in model comparisons. Most researchers investigating this topic have sought to identify a single index for all-purpose use. The most visible discussions in the field of ecology have debated the use of *AIC* versus *BIC*. There are also a great many variants of *AIC* and *BIC* that have been developed and discussed. For those in ecology, the recent simulation study by Lin et al. (2017) is perhaps most instructive. It is undeniable at the present time that the ideal information measure for model comparisons

varies, depending on the assumptions you make about the complexity of the underlying true data-generating process, the sample size, the strengths of effects of interest and other properties of the data. I expect a decision tree or matrix of recommendations will be the ultimate solution, though certainly there are advanced approaches being studied (e.g. Brewer et al. 2016).

The presentation here would be incomplete without mentioning that the ultimate solution to selecting the best model involves the data itself and not the methods of analysis. The bigger and better the sample, the more confidence we may have in the conclusions. If our goal is to generalise beyond the current sample, there is no substitute for sound mechanistic knowledge and sequential learning across linked studies (Grace and Irvine 2020).

Acknowledgements

I thank Lori Randall, USGS, Maria Felipe-Lucia, Helmholtz Center for Environmental Research and Frank Pennekamp, University of Zurich, for helpful review comments and suggestions. This work was supported by the USGS Land Change Science and Ecosystems Programs. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Hosting institution

U.S. Geological Survey

Conflicts of interest

No conflicts of interest.

References

- Aho K, Derryberry D, Peterson T (2014) Model selection for ecologists: the worldviews of AIC and BIC. Ecology 95: 631-636. https://doi.org/10.1890/13-1452.1
- Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716-723. https://doi.org/10.1109/TAC.1974.1100705
- Amrhein V, Greenland S, McShane B (2019) Scientists rise up against statistical significance. Nature 567: 305-307. https://doi.org/10.1038/d41586-019-00857-9
- Barrett P (2007) Structural equation modelling: Adjudging model fit. Personality and Individual Differences 42: 815-824. https://doi.org/10.1016/j.paid.2006.09.018
- Bollen K, Harden J, Ray S, Zavisca J (2014) BIC and alternative Bayesian information criteria in the selection of structural equation models. Structural Equation Modeling 21: 1-19. https://doi.org/10.1080/10705511.2014.856691

- Brewer MJ, Butler A, Cooksley SL (2016) The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. Methods in Ecology and Evolution 7: 679-692. https://doi.org/10.1111/2041-210X.12541
- Burnham K, Anderson D, Huyvaert K (2011) AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behavioral Ecology and Sociobiology 65: 23-35. https://doi.org/10.1007/s00265-010-1029-6
- Burnham K, Anderson D (2014) P values are only an index to evidence: 20th-vs. 21st-century statistical science. Ecology 95: 627-630. https://doi.org/10.1890/13-1066.1
- Burnham KP, Anderson DR (2002) Model selection and multimodel inference. Springer Publishers, New York, NY, USA.
- Cade BS (2015) Model averaging and muddled multimodel inferences. Ecology 96: 2370-2382. https://doi.org/10.1890/14-1639.1
- Grace J (2006) Structural Equation Modeling and Natural Systems. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511617799
- Grace JB, Irvine KM (2020) Scientist's guide to developing explanatory statistical models using causal analysis principles. Ecology e02962. https://doi.org/10.1002/ecy.2962
- Hu LT, Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling 6: 1-55. https://doi.org/10.1080/10705519909540118
- Jöreskog KG (1970) A general method for analysis of covariance structures. Biometrika 57: 239-251. https://doi.org/10.1093/biomet/57.2.239
- Kline RB (1998) Principles and practice of structural equation modeling. First Edition. Guilford Press, New York, NY, USA.
- Kline RB (2016) Principles and practice of structural equation modeling. Fourth Edition. Guilford Press, New York, NY, USA.
- Larson DL, Grace JB (2004) Temporal dynamics of leafy spurge (*Euphorbia esula*) and two species of flea beetles (*Aphthona* spp.) used as biological control agents. Biological Control 29: 207-214. https://doi.org/10.1016/S1049-9644(03)00156-7
- Larson DL, Grace JB, Larson JL (2008) Long-term dynamics of leafy spurge (Euphorbia esula) and its biocontrol agent, flea beetles in the genus *Aphthona* . Biological Control 47: 250-256. https://doi.org/10.1016/j.biocontrol.2008.07.016
- Lefcheck JS (2016) piecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. Methods in Ecology and Evolution 7: 573-579. https://doi.org/10.1111/2041-210X.12512
- Lin LC, Huang PH, Weng LJ (2017) Selecting path models in SEM: A comparison of model selection criteria. Structural Equation Modeling 24: 855-869. https://doi.org/10.1080/10705511.2017.1363652
- Marsh HW, Hau KT, Wen Z (2004) In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Structural Equation Modeling 11: 320-341. https://doi.org/10.1207/s15328007sem1103_2
- McShane BB, Gal D, Gelman A, Robert C, Tackett JL (2019) Abandon statistical significance. American Statistician 73: 235-245. https://doi.org/10.1080/00031305.2018.1527253

- Mulaic S (2009) Linear causal modeling with structural equations. CRC Press, New York, NY, USA. https://doi.org/10.1201/9781439800393
- Murtaugh P (2014) In defense of P values. Ecology 95: 611-617. https://doi.org/10.1890/13-0590.1
- Pearl J (2000) Causality. Cambridge University Press, Cambridge, UK.
- Schwarz G (1978) Estimating the dimension of a mode. The Annals of Statistics 6: 461-464. https://doi.org/10.1214/aos/1176344136
- Shipley B (2000) A new inferential test for path models based on directed acyclic graphs. Structural Equation Modeling 7: 206-218. https://doi.org/10.1207/S15328007SEM0702_4
- Shipley B (2013) The AIC model selection method applied to path analytic models compared using a d-separation test. Ecology 94: 560-564. https://doi.org/10.1890/12-0976.1
- Shipley B (2016) Cause and correlation in biology. Second Edition. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9781139979573
- Shipley B, Douma JC (2019) Generalized AIC and chi-squared statistics for path models consistent with directed acyclic graphs. Ecology https://doi.org/10.1002/ecy.2960
- Smith M, Knapp A (2003) Dominant species maintain ecosystem function with non-random species loss. Ecology Letters 6: 509-517. https://doi.org/10.1046/j.1461-0248.2003.00454.x
- Steiger JH (2007) Understanding the limitations of global fit assessment in structural equation modeling. Personality and Individual Differences 42: 893-989. https://doi.org/10.1016/j.paid.2006.09.017
- Tomer A (2003) A short history of structural equation models. In: Pugesek B, Tomer A, von Eye A (Eds) Structural equation modeling: Applications in ecological and evolutionary biology. Cambridge University Press, Cambridge, UK, 85-124 pp. https://doi.org/10.1017/CBO9780511542138.005

Supplementary material

Suppl. material 1: A 'Weight of Evidence' Approach to Evaluating Structural Equation Models- Supplement1 doi

Authors: Grace, JB
Data type: R code
Brief description: This text file contains the R code used to develop the demonstrations included in Grace JB (2020) A 'weight of evidence' approach to evaluating structural equation models. One Ecosystem
Download file (5.19 kb)