Chapter – 6
Alternatives to Traditional Model
Comparison Strategies for Covariance Structure Models
Chapter – 6
Alternatives to Traditional Model Comparison Strategies for Covariance Structure Models

6.1 Introduction

In this chapter I discuss two related issues relevant to traditional methods of comparing alternative covariance structure models (CSM) in the context of ecological research. Use of the traditional test of parametrically nested models in applications of CSM (the $\chi^2$ difference or likelihood ratio [LR] test) suffers from several limitations, as discussed by numerous methodologists (MacCallum, Browne, & Cai, 2005). Our primary objection is that the traditional approach to comparing models is predicated on the assumption that it is possible for two models to have identical fit in the population. We argue instead that any method of model comparison which assumes that a point hypothesis of equal fit can hold exactly in the population (e.g., the LR test) is fundamentally flawed. We discuss two alternative approaches to the LR test which avoid the necessity of hypothesizing that two models share identical fit in the population. One approach concerns framing the hypothesis of interest differently, which naturally leads to questions of how to assess statistical power and appropriate sample size. The other approach concerns a radical realignment of how researchers approach model evaluation, avoiding traditional null hypothesis testing altogether in favor of identifying the model that maximizes generalizability.

Power presents a recurrent problem to those familiar with null hypothesis significance testing (NHST). How large should a sample be in order to have
adequate probability of rejecting a false null hypothesis? What is the probability of rejecting a false null if our sample is of size $N$? These questions present special challenges in the context of CSM because the relative status of null and alternative hypotheses are interchanged from their familiar positions — the null hypothesis in CSM represents the theory under scrutiny, and power is framed in terms of the sample size necessary to reject a false model. Traditional goodness of-fit tests deal with the null hypothesis under which the model fits exactly in the population (exact fit test). Point hypotheses tested by the exact fit test are likely never true in practice, so how should power be conceptualized? We present an alternative strategy extending earlier work on power for tests of close fit (rather than exact fit) of single models to tests of small difference (rather than no difference) in comparisons of nested models. The null hypothesis in a test of small difference states that the model fits nearly as well, but not the same, as a less constrained model.

Another alternative to traditional methods of model assessment is to avoid the hypothesis-testing framework altogether, instead adopting a model selection approach that uses comparative reliability as the criterion for selecting a model as superior to its rivals (Weakliem, 2004). Specifically, we argue that the evaluation of models against arbitrary benchmarks of fit gets the researcher nowhere — only in the context of model comparison can science advance meaningfully (Burnham & Anderson, 2004). Maximizing generalizability involves ranking competing models against one another in terms of their ability to fit present and future data. Adopting this model selection strategy, however, necessitates proper quantification
of model complexity — the average ability of a model to fit any given data. Most model fit indices include an adjustment for complexity that is a simple function of the number of free model parameters. We argue that this adjustment is insufficient; the average ability of a model to fit data is not completely governed by the number of parameters. Consequently, we present and illustrate the use of a new information-theoretic selection criterion that quantifies complexity in a more appropriate manner. This, in turn, permits the adoption of an appropriate model selection strategy that avoids pitfalls associated with LR tests.

I begin by providing a review of the traditional representation of the covariance structure model (with mean structure), with an emphasis on its application to multiple groups. We then describe advantages granted by adopting a model comparison perspective in CSM. One way around the problems with traditional approaches is to change the hypothesis under scrutiny to a more realistic one. In describing this alternative approach, we describe an approach to power analysis in CSM involving an extension of recently introduced methods to nested model scenarios. Following our discussion of power, we further explore the potential value of adopting a model selection approach that avoids hypothesis testing — and thus most problems associated with LR tests—altogether. In the process, we introduce the topic of model complexity, suggesting and illustrating the use of a new selection criterion that permits appropriate model comparison even for no nested models.
6.2 COVARIANCE STRUCTURE MODELING

Covariance structure modeling (CSM) is an application of the general linear model combining aspects of factor analysis and path analysis. In CSM, the model expresses a pattern of relationships among a collection of observed (manifest) and unobserved (latent) variables. These relationships are expressed as free parameters representing path coefficients, variances, and covariance's, as well as other parameters constrained to specific, theory-implied values or to functions of other parameters. For simplicity, we restrict attention to the ally model (LISREL Submodel 3B; Joreskog & Sorbom, 1996), which involves only four parameter matrices, although the points we discuss later apply more broadly.

6.3 The Importance of CSM to Ecological Research

There are several advantages associated with CSM that make it especially appropriate for addressing hypotheses in the context of ecological models. First, CSM permits the specification and testing of complex causal and co-relational hypotheses. Sets of hypotheses can be tested simultaneously by constraining model parameters to particular values, or equal to one another within or across multiple groups or occasions of measurement, in ways consistent with theoretical predictions. Second, by permitting several measured variables to serve as indicators of unobserved latent variables, CSM separates meaningful variance from variance specific to items, allowing researchers to test structural hypotheses relating constructs that are not directly observed. Third, CSM is appropriate for testing co-relational or causal hypotheses using either (or both) experimental or
observational data. One of the central ideas behind ecological modeling is that there is much knowledge to be gained by collecting data observed in context that would be difficult or impossible to learn under artificial conditions. Finally, CSM is a flexible modeling approach that can easily accommodate many novel modeling problems.

6.4 The Importance of Adopting a Model Comparison Perspective

In practice, CSMs are typically evaluated against benchmark criteria of good fit. Based on how well a model fits data relative to these criteria, the model is usually said to fit well or poorly in an absolute sense. The reasoning underlying this strategy of gauging a model’s potential usefulness is predicated on an approach to science termed falsifications, which holds that evidence accumulates for theories when their predictions are subjected to, and pass, realistic “risky” tests. If a model passes such a test under conditions where it would be expected to fail if false (i.e., if it shows good fit), evidence accumulates in favor of the theory whose predictions the model represents. If it fails, the model is either rejected or modified, with implications for the revision or abandonment of the theory. Ideally, a model is subjected to repeated risky tests to give a better idea of its long-term performance, but replication is unfortunately rare in the social sciences.

An alternative philosophical perspective maintains that the evaluation of models in isolation tells us very little, and that the fit of a model to a particular data set is nearly uninformative. Rather, science progresses more rapidly if
competing theories are compared to one another in terms of their abilities to fit existing data and, as we will discuss, their abilities to fit future data arising from the same latent process (Lakatos, 1970; MacCallum, 2003). This approach is sometimes termed strong inference (Platt, 1964), and involves model comparison as a signature feature. We know from the outset that no model can be literally true in all of its particulars, unless one is extraordinarily lucky or possesses divinely inspired theory-designing skills. But it stands to reason that, given a set of alternative models, one of those models probably represents the objectively true data-generating process better than other models do. It is the researcher’s task to identify this model and use it as the best working hypothesis until an even more appropriate model is identified (which, by design, inevitably happens). Every time a model is selected as the optimal one from a pool of rivals, evidence accumulates in its favor. This process of rejecting alternative explanations and modifying and re-testing models against new data continues ad infinitum, permitting scientists to constantly update their best working hypotheses about the unobserved processes underlying human behavior.

Because no model is literally true, there is an obvious logical problem in testing the null hypothesis that a model fits data perfectly in the population. Yet, this is precisely the hypothesis tested by the popular LR test of model fit. Moreover, most fit indices require the researcher to choose arbitrary values to represent benchmarks of good fit. A model comparison approach goes far in avoiding these problems, although it cannot avoid them altogether. Most damning, it is possible to assert apriori that the hypothesis tested with the $\chi^2$ statistic — that
a model fits exactly in the population or that two models share exactly the same fit — is false in virtually every setting (Bentler & Bonett, 1980; Tucker & Lewis, 1973). A model selection approach avoids the pitfalls inherent in hypothesis testing by avoiding such tests altogether.

In addition to adhering more closely to scientific ideals and circumventing logical problems inherent in testing isolated models, the practice of model comparison avoids some problems associated with confirmation bias. Confirmation bias reflects the tendency for scientists unconsciously to increase the odds of supporting a preferred hypothesis (Greenwald, Pratkanis, Leippe, & Baumgardner, 1986). Regardless of why or how much the deck is stacked in favor of the researcher’s preferred model in terms of absolute fit, one model is virtually guaranteed to outperform its rivals. Model comparison does not entirely eliminate confirmation bias, but it certainly has the potential to improve the researcher’s objectivity.

In the foregoing we have explained that the popular LR test is fundamentally flawed in that the hypothesis it tests is rarely or never true in practice; thus, persistent and frequent use of the LR test is of questionable utility. We have also explained that adopting a model selection approach, in which at least two theory-inspired models are compared, has potentially greater scientific potential. In the following two broad sections, we outline some practical solutions to logical problems imposed by use of the traditional LR tests of model fit in ecological research. The first suggested approach emphasizes the utility of avoiding the hypothesis that two models have identical fit in favor of a hypothesis
that the difference is within tolerable limits. This approach recognizes that no model can realistically fit perfectly in the population, and points out that shifting the focus to a less stringent hypothesis is more logical, yet has consequences for statistical power and identifying the necessary sample size. We describe and discuss methods that can be used to address these problems. The second section focuses more closely on the model selection perspective just outlined, emphasizing that model fit is overrated as a criterion for the success or usefulness of a theory. Rather, more attention should be paid to a model’s ability to cross-validate, or generalize, relative to competing models. Special attention is devoted to a new model selection criterion that considers aspects of model complexity beyond simply the number of free parameters.

6.5 Concluding Remarks

There are two broad issues that we wish to emphasize to close this section on power analysis and specification of the null hypothesis when performing comparisons of nested models. The first issue is the choice of pairs of RMSEA values. Essentially the results of any application of any of the methods we described are contingent on the particular RMSEA values that the user selects. Here we can offer only some general principles. For a more thorough discussion of this issue we refer the reader to MacCallum et al. (2006). For specifying RMSEA values for testing a null hypothesis of a small difference in fit, the user should regard the Good-Enough Principle (Serlin & Lapsley, 1985) as the objective, and pick RMSEA values for Models A and B that represent a difference so small that
the user is willing to ignore it. In the context of power analysis, the relevant
general principle would be to choose values that represent a difference that the
investigator would wish to have a high probability of detecting. In practice, users
will need to rely on guidelines for the use of RMSEA as mentioned earlier
(Browne & Cudeck, 1993; Steiger, 1994), as well as the characteristics of the
models under comparison.

The second issue has to do with the assumptions involved in our
developments. All of the methodological developments presented thus far rely on
well known distribution theory and its assumptions. Specifically, we make
extensive use of the assumptions that ensure the chi-squared ness of the LR test
statistic $T$, for both the central and non central cases. These include multivariate
normality, the standard set of regularity conditions on the likelihood to carryout
asymptotic expansions, and the population drift assumption (Steiger et al., 1985).
As always, however, such assumptions never hold exactly in the real world, so the
user should always be cautious in the application of these methods in data analysis
and should watch for potential pitfalls due to assumption violations. MacCallum et
al. (2006) discuss the consequences of such violations.

6.6 MODEL SELECTION AND MODEL COMPLEXITY

Model Selection and Generalizability

In the preceding section we provide and illustrate methods for comparing
rival models in terms of a noncentrality-based fit index, RMSEA. We suggest that
this strategy is appropriate for statistically comparing the fit of rival,
parametrically nested models, but the procedure depends in part on the researcher’s judgment of appropriate choices for $\epsilon^A$ and $\epsilon^B$, or what, in the researcher’s judgment, constitutes the smallest difference in fit that it would be interesting to detect. In practice, a model can demonstrate good fit for any number of reasons, including a theory’s proximity to the objective truth (or verisimilitude; Meehl, 1990), random chance, simply having many free parameters, or by possessing a structure allowing parameters to assume values which lead to good model fit for many different data patterns—even those generated by other processes not considered by the researcher. In other words, models can demonstrate close fit to data for reasons other than being “correct,” even if one grants that true models are possible to specify (we do not), so good fit should represent only one criterion by which we judge a model’s usefulness or quality.

Another criterion of model success that has found much support in mathematical psychology and the cognitive modeling literature is generalizability (or reliability). The idea here is that it is not sufficient for a model to show good fit to the data in hand. If a model is to be useful, it should predict other data generated by the same latent process, or capture the regularities underlying data consisting of signal and noise. If a model is highly complex, refitting the model to new data from scratch will not advance our knowledge by much; if a model’s structure is complex enough to show good fit to one data set, it may be complex enough to show good fit to many other data sets simply by adjusting its parameters. In other words, pure goodness of fit represents fit to signal plus
fit to noise. However, if model parameters are fixed to values estimated in one setting, and the model still demonstrates good fit in a second sample (i.e., if the model *cross-validates* well), the model has gained considerable support. A model’s potential to cross-validate well is its generalizability, and it is possible to quantify generalizability based only on knowledge of the model’s formant of its fit to a given data set. By quantifying a model’s potential to cross validate, generalizability avoids problems associated with good fit arising from fitting error or from a model’s flexibility. It also does not rely on unsupportable assumptions regarding a model’s absolute truth or falsity. Therefore, generalizability is arguably a better criterion for model retention than is goodness of fit per se (Pitt & Myung, 2002).

Earlier we stated that adopting a model selection perspective requires a fundamental shift in how researchers approach model evaluation. Traditional hypothesis testing based on LR tests results in a dichotomous accept–reject decision without quantifying how much confidence one should place in a model, or how much relative confidence one should place in each member of a set of rival models. In model comparison, on the other hand, no null hypothesis is tested (Burnham & Anderson, 2004). The appropriate sample size is not selected based on power to reject hypotheses of exact or close fit (obviously, since no such hypotheses are tested), but rather to attain acceptable levels of precision of parameter estimates. Rather than retaining or discarding models on a strict accept–reject basis, models are ranked in terms of their generalizability, a notion that
combines fit with parsimony, both of which are hallmark characteristics of a good model.

The model selection approach does not require that any of the rival models be correct, or even (counter intuitively) that any of the models fit well in an absolute sense. The process is designed in such a way that researchers will gravitate toward successively better models after repeated model comparisons. The more such comparisons a particular model survives, the better its track record becomes, and the more support it accrues. Therefore, it is incumbent upon scientists to devise models that are not only superior to competing models, but also perform well in an absolute sense. Such models will, in the long run, possess higher probabilities of surviving risky tests, facilitate substantive explanation, predict future data, and lead to the formulation of novel hypotheses. But, again, the model selection strategy we advocate does not require that any of the competing models be correct or even close to correct in the absolute sense.

**Information-Theoretic Criteria**

In contrast to model selection methods rooted in Bayesian or frequents traditions, much research points to information theory as a likely source for the optimal model selection criterion. Selection criteria based on information theory seek to locate the one model, out of a pool of rival models, which shows the optimal fidelity, or signal-to-noise ratio; this is the model that demonstrates the best balance between fit and parsimony. This balance was termed *generalizability* earlier. Several popular model selection criteria were either derived from, or are
closely related to, information theory. The most popular such criteria are the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwartz, 1978). Excellent treatments of AIC and BIC can be found elsewhere (e.g., Burnham & Anderson, 2002, 2004; Kuha, 2004).

Many information-based criteria may be construed as attempts to estimate the Kullback–Leibler (K–L) distance. The K-L distance is the (unknown) information lost by representing the true latent process with an approximating model (Burnham & Anderson, 2004). Even though we cannot compute the K–L distance directly because there is one term in the K–L distance definition that is not possible to estimate, we can approximate relative K–L distance in various ways by combining knowledge of the data with knowledge of the models under scrutiny. Of great importance for model comparison, the ability to approximate relative K–L distance permits the ranking of models in terms of their estimated verisimilitude, tempered by our uncertainty about the degree of approximation. In other words, using information-based criteria, models can be ranked in terms of estimated generalizability.

**Minimum Description Length and the Normalized Maximum Likelihood**

Information-based criteria such as AIC and BIC are used with great frequency in model comparisons and with increasing frequency in applications of CSM. However, they suffer from at least two major drawbacks. First, they employ complexity adjustments that are functions only of the number of free model parameters. Second, they implicitly require the strong assumption that a correct
model exists. We focus instead on a newer criterion that remains relatively unknown in the social sciences, yet we feel has great promise for application in model selection. This is the principle of minimum description length (MDL: Grunwald, 2000; Myung, Navarro, & Pitt, 2005; Rissanen, 1996, 2001; Stine, 2004). The MDL principle involves construing data as compressible strings, and conceiving of models as compression codes. If models are viewed as data compression codes, the optimal code would be one that compresses (or simply represents) the data with the greatest fidelity. With relevance to the limitations of criteria such as AIC and BIC, the MDL principle involves no assumption that a true model exists. If one accepts that a model’s proximity to the truth is either undefined (i.e., that the notion of a true model is merely a convenience and bears no direct relation to reality) or is at any rate impossible to determine, then the MDL principle offers a viable alternative to traditional methods of model selection. Excellent discussions of the MDL principle can be found in Grunwald (2000), Grunwald, Myung, and Pitt (2005), Hansen and Yu (2001), and Markon and Krueger (2004). Three quantifications of the MDL principle are normalized maximum likelihood (NML), Fisher information approximation (FIA), and stochastic information complexity (SIC). NML is quantified as:

$$NML = \frac{L(y|\hat{\theta})}{\int_{\mathcal{S}} L(z|\hat{\theta}(z))dz}, \quad (14)$$

or the likelihood of the data given the model divided by the sum of all such likelihoods. FIA is quantified as
an approximation to the negative logarithm of NML that makes use of the number of free parameters \((q)\) and the determinant of the Fisher information matrix, \(I(\theta)\).

SIC, an approximation to FIA that is typically more tractable in practice, is quantified as:

\[
SIC = -\ln L(\mathbf{y}|\hat{\mathbf{\theta}}) + \frac{1}{2} \ln |nI(\theta)|. \tag{16}
\]

The Appendix (see Quant.KU.edu) contains more detailed discussion of these criteria. NML, FIA, and SIC all represent model fit penalized by the model’s average ability to fit any given data.

NML is similar in spirit to selection criteria such as AIC and BIC in several respects, save that preferable models are associated with higher values of NML but with lower values of AIC or BIC. All of these criteria can be framed as functions of the likelihood value adjusted for model complexity, although the complexity correction assumes different forms for different criteria. NML differs from criteria like AIC and BIC mainly in that not every parameter is penalized to the same extent. NML imposes an adjustment commensurate with the degree to which each free parameter increases complexity, as reflected in the model’s general data-fitting capacity. Consequently, NML does not assume (as do AIC and BIC) that each parameter contributes equally to goodness of fit. Therefore, both parametric and structural components of complexity are considered. A major additional advantage of NML (which it shares with AIC and BIC) is that it does not require rival models to be nested. Thus, if two competing theories posit different patterns
of constraints, such models can be directly compared using criteria derived from information theory.

### 6.7 Applying MDL in Practice

To illustrate how the MDL principle may be employed in practice, we present two brief examples from the applied literature. In both examples we compute NML; in the second, we supplement NML with computation of SIC because original data were available with which to compute the $|nI(\theta)|$ term. Neither the denominator term in NML (see Equation [A1]) nor the structural complexity term in FIA (see Equation [A2]) can be computed directly in the context of CSM. Numerical integration techniques are typically applied instead. To facilitate computation of NML, we simulated the data space by generating large numbers of random uniform correlation matrices (R) using Markov chain Monte Carlo (MCMC) methods. These matrices were uniform in the sense that all possible R matrices had equal *apriori* probabilities of being generated. All models were fit to all simulated matrices, and the likelihoods were averaged to form the denominator of the NML formula. The numerators were supplied by simply noting the likelihood value associated with the converged solution for each model applied to real data.

**Example 1.** Our second example draws on three covariance structure models compared by Larose, Guay, and Boivin (2002). The authors were primarily interested in comparing the Cognitive Bias Model and Social Network Model, two models proposed to explain variability in a Loneliness latent variable using
Attachment Security, Emotional Support, and Social Support. These two models (which we denote $L_1$ and $L_2$) are presented in the first two panels of Figure 3.3. Based on results indicating that both models fit the data well and were thus viable explanations for the observed pattern of effects, the authors devised a third model combining features of the first two, dubbed the Cognitive-Network Model ($L_3$ in Figure 3.3).

All three models were found to fit the data well using self-report measures ($N = 125$), and to fit even better using friend-report measures. In both cases, the Cognitive-Network Model was found to fit the data significantly better than either the Cognitive Bias Model or the Social Network Model. Following procedures already described, we reevaluated Larose et al.’s models (fit to self report data) using NML. Results are reported in Table 3.2. Because raw data were available in their article, we are also able to provide estimates of SIC.

Contrary to the authors’ findings, both NML and SIC indicate that the Cognitive Bias Model performs better than either the Social Networks Model or the proposed Cognitive-Network Model in terms of generalizability. Combining features of two already well-fitting models does not necessarily grant a scientific advantage when the resulting model is more complex than either of its
FIGURE 3.2
Rival models investigated by Ingram et al. (2001).
FIGURE 3.3
Rival models investigated by Larose et al. (2002).

Model L1

Model L2

Model L3
competitors. In this instance, as in the previous example, the chosen model was selected primarily because it showed better absolute fit; this better fit was due in part to the fact that the Cognitive-Network Model was more complex than its competitors. An implication of this finding is that, whereas the Cognitive-Network Model may fit the given data set better than the Cognitive Bias Model and the Social Networks Model in absolute terms, it has a lower likelihood of generalizing well to future data.

6.8 Summary

Like other information-theoretic selection criteria, MDL does not require rival models to be parametrically nested. Nor does its use require the assumption that a true model exists. Furthermore, MDL considers more sources of complexity than simply a model’s number of parameters. In sum, we feel that the MDL principle has great potential for use in model comparisons in CSM.

6.9 Limitations

Of course, NML is not a panacea. Three limitations of NML are that it is difficult to compute, it relies on the assumptions of maximum likelihood, and it involves often arbitrary bounds on the data space. The first limitation will be overcome as processor speeds increase and as NML becomes included in standard model estimation packages. In the meantime, the more tractable MDL approximation, SIC (Rissanen, 1989), can be used if the numerical integration
necessary for NML proves too time-intensive. As for the second limitation, it is unknown how robust MDL methods are to violations of ML assumptions. This would be a fruitful avenue for future research.

The third limitation is more challenging because it requires the researcher to make a subjective decision regarding boundaries on the data space. We restricted attention to correlation matrices for simplicity. We recognize that many modeling applications require covariance matrices rather than correlation matrices (and sometimes also mean vectors). For example, virtually any application in which models are fit to multiple groups simultaneously, such as in factorial invariance studies, requires the use of covariance matrices. Growth curve modeling requires covariance matrices and mean vectors. Lower and upper boundaries must be imposed on generated means and variances if such data are required, and these choices constitute even more subjective input. It is generally agreed that data generated for the purpose of quantifying model complexity should be uniformly representative of the data space (Dunn, 2000), yet choices regarding the range of data generation may exert great influence on the ranking of competing models. It is thus important that reasonable bounds be investigated to ensure reasonable and stable model rankings. A discussion of the implications for arbitrary integration ranges can be found in Lanterman (2005).

### 6.10 DISCUSSION

We have proposed two alternatives to traditional methods of comparing covariance structure models. Both alternatives were suggested in response to
limitations of the popular LR test; the most severe limitation is that the hypothesis
tested by the LR test (that two models have identical fit) is never true in practice,
so investigating its truth or falsity would seem to be a questionable undertaking
(MacCallum et al., 2006). The first alternative procedure posits a modified null
hypothesis such that the difference in fit between two nested models is within
tolerable limits. The second alternative we discuss is to compare rival (not
necessarily nested) models in terms of relative generalizability using selection
indices based on the MDL principle. Both methods encourage a model comparison
approach to science that is likely to move the field in the direction of successively
better models.

There are interesting parallels between the strategies proposed here and a
framework for model assessment proposed by Linhart and Zucchini (1986) and
elaborated upon by Cudeck and Henly (1991) in the context of CSM. Because it
relies on RMSEA to specify null and alternative hypotheses, the first approach
(using RMSEA to specify hypotheses of close fit) can be seen as way to compare
nested models in terms of their approximation discrepancy, or lack of fit in the
population. In other words, this method is a way to gauge models’ relative
nearness to the objectively true data-generating process, or their relative
verisimilitudes. The second method of model comparison makes use of the MDL
principle to facilitate comparison of models in terms of their relative
generalizabilities, or abilities to predict future data arising from the same
generating process. This strategy can be seen as a way to compare models (nested
or non-nested) in terms of their overall discrepancy, tempering information about
lack of fit with lack of confidence due to sampling error. When $N$ is large, enough information is available to support highly complex models if such models are appropriate. When $N$ is small, uncertainty obliges us to conservatively select less complex models until more information becomes available (Cudeck & Henly, 1991). Thus, NML and similar criteria are direct applications of the parsimony principle, or Occam’s razor.

The parallels between the measures of verisimilitude and generalizability on one hand, and the Linhart–Zucchini and Cudeck–Henly frameworks on the other, perhaps deserve more attention in future research. High verisimilitude and high generalizability are both desirable characteristics for models to possess, but selecting the most generalizable model does not necessarily imply that the selected model is also closest to the objective truth. Therefore we do not advocate choosing one approach or the other, or even limiting attention to these two strategies. Rather, we suggest combining these strategies with existing model evaluation and selection techniques so that judgments may be based on as much information as possible. Regardless of what strategy the researcher chooses, the strongest recommendation we can make is that researchers should, whenever circumstances permit it, adopt a model selection strategy rather than to evaluate single models in isolation. The methods illustrated here are viable alternatives to the standard approach, and can be applied easily in many modeling settings involving longitudinal and/or ecological data.