# Evaluation and Comparison of Computational Models

By

Jay Myung, Yun Tang, and Mark A. Pitt

Ohio State University

September 16, 2008

Submitted for M. L. Johnson and L. Brand (eds.), *Computer Methods* (A volume of *Methods in Enzymology*)

Running Head: Methods of Model Evaluation and Comparison

All correspondence to

Dr. Jay I. Myung
Department of Psychology
Ohio State University
225 Psychology Building
1835 Neil Avenue Mall
Columbus, Ohio 43210-1351

Voice: 614-292-1862
Fax: 614-292-5601
Email: myung.1@osu.edu

## Abstract

Computational models are powerful tools that can enhance understanding of scientific phenomena. The enterprise of modeling is most productive when the reasons underlying a model’s adequacy, and possibly its superiority to other models, are understood. This article begins with an overview of the main criteria that must be considered in model evaluation and selection, in particular explaining why generalizability is the preferred criterion for model selection. This is followed by a review of measures of generalizability. In the final section, we demonstrate the use of five versatile and easy-to-use selection methods for choosing between two mathematical models of protein folding.

## Introduction

How does one evaluate the quality of a computational model of enzyme kinetics? The answer to this question is important and complicated. It is important because mathematics makes it possible to formalize the reaction, providing a precise description of how the factors affecting it interact. Study of the model can lead to significant understanding of the reaction, so much so that the model can serve not merely as a description of the reaction, but can contribute to explaining its role in metabolism. Model evaluation is complicated because it involves subjectivity, which can be difficult to quantify.

We begin this paper with a conceptual overview of some of the central issues in model evaluation and selection, with an emphasis on those pertinent to the comparison of two or more models. This is followed by a selective survey of model comparison methods, and then an application example that demonstrates the use of five simple yet informative model comparison methods.

Criteria on which models are evaluated can be grouped into those that are difficult to quantify and those for which it is easier to do so (Jacobs and Grainger, 1994). Criteria such as *explanatory adequacy* (whether the theoretical account of the model helps to make sense of observed data) and *interpretability* (whether the components of the model, especially its parameters, are understandable and are linked to known processes) rely on the knowledge, experience, and preferences of the modeler. Although the use of these criteria may favor one model over another, they do not lend themselves to quantification because of their complexity and qualitative properties. Model evaluation criteria for which there are quantitative measures include *descriptive adequacy* (whether the model fits the observed data), *complexity* or *simplicity* (whether the model's description of observed data is achieved in the simplest possible manner), and *generalizability* (whether the model provides a good predictor of future observations). Although each criterion identifies a property of a model that can be evaluated on its own, in practice they are rarely independent of one another. Consideration of all three simultaneously is necessary to assess fully the adequacy of a model.

**Conceptual Overview of Model Evaluation and Comparison**

Before discussing the three quantitative criteria in more depth, we highlight some of the key challenges of modeling. Models are mathematical representations of the phenomenon under study. They are meant to capture patterns or regularities in empirical data by altering parameters that correspond to variables that are thought to affect the phenomenon. Model specification is difficult because our knowledge about the phenomenon being modeled is rarely complete. That is, the empirical data obtained from studying the phenomenon are limited, providing only partial information (i.e., snapshots) about its properties and the variables that influence to it. With limited information, it is next to impossible to construct the "true" model. Furthermore, with only partial information, it is likely that multiple models are plausible; more than one model can provide a good account of the data. Given this situation, it is most productive to view models as approximations, which one seeks to improve through repeated testing.

Another reason models can be only approximations is that data are inherently noisy. There is always measurement error, however small, and there may also be other sources of uncontrolled variation introduced during the data collection process that amplifies this error. Error clouds the regularity in the data, increasing the difficulty of modeling. Because noise cannot be removed from the data, the researcher must be careful that the model is capturing the meaningful trends in the data and not error variation. As will be explained below, one reason why generalizability has become the preferred method of model comparison is how it tackles the problem of noise in data.

The descriptive adequacy of a model is assessed by measuring how well it fits a set of empirical data. A number of goodness-of-fit (GOF) measures are in use, including sum of squared errors (SSE), percent variance accounted for (PVA), and maximum likelihood (ML; e.g., Myung, 2003)). Although their origins differ, they measure the discrepancy between the empirical data and a model's ability to reproduce those data. GOF measures are popular because they are relatively easy to compute and the measures are versatile, being applicable to many types of models and types of data. Perhaps most of all, a good fit is an almost irresistible piece of evidence in favor of a model's adequacy. The model appears to do just what one wants it to -- mimic the process that generated the data. This reasoning is often taken a step further by suggesting that the better the fit, the more accurate the model. When comparing competing models, then, the one that provides the best fit should be preferred.

GOF would be suitable for model evaluation and comparison if it were not for the fact that data are noisy. As described above, a data set contains the regularity that is presumed to reflect the phenomenon of interest plus noise. GOF does not distinguish between the two, providing a single measure of a model's fit to both (i.e., GOF = fit to regularity + fit to noise). As this conceptual equation shows, a good fit can be achieved for the wrong reasons, by fitting noise well instead of the regularity. In fact, the better a model is at fitting noise, the more likely it will provide a superior fit than a competing model, possibly resulting in the selection of a model that in actuality bears little resemblance to the process being modeled. GOF alone is a poor criterion for model selection because of the potential to yield misleading information.

This is not to say that GOF should be abandoned. On the contrary, a model's fit to data is a crucial piece of information. Data are the only link to the process being modeled, and a good fit can indicate that the model mimics the process well. Rather, what is a needed is a means of ensuring that a model does not provide a good fit for the wrong reason.

What allows a model to fit noisy data better than its competitors is that it is the most complex. *Complexity* refers to the inherent flexibility of a model that allows it to fit diverse data patterns (Myung and Pitt, 1997). By varying the values of its parameters, a model will produce different data patterns. What distinguishes a simple model from a complex one is the model's sensitivity to parameter variation. For a simple model, parameter variation will produce small and gradual changes in model performance. For a complex model, small parameter changes can result in dramatically different data patterns. It is this flexibility in producing a wide range of data patterns that makes a model complex. For example, the cubic model y $= ax^2 + bx + c$ is more complex than the linear model $y = ax + b$. As will be shown in the next section, model selection methods such as AIC and BIC include terms that penalize model complexity, thereby neutralizing complexity differences among models.

Underlying the introduction of these more sophisticated methods is an important conceptual shift in the goal of model selection. Instead of choosing the model that provides the best fit to a single set of data, choose the model that, *with its parameters held constant*, provides the best fit to the data if the experiment were repeated again and again. That is, choose the model that generalizes best to replications of the same experiment. Across replications, the noise in the data will change, but the regularity of interest should not. The more noise that the model captures when fit to the first data set, the poorer its measure of fit will be when fitting the data in replications of that experiment because the noise will have changed. If a model captures mostly the regularity, then its fits will be consistently good across replications. The problem of distinguishing regularity from noise is solved by focusing on generalizability. A model is of questionable worth if it does not have good predictive accuracy in the same experimental setting. Generalizability evaluates exactly this, and it is why many consider generalizability to be the best criterion on which models should be compared (Grunwald, Myung and Pitt, 2005).

------------------------

Figure 1 about here

------------------------

The graphs in Figure 1 summarizes the relationship between the three quantitative criteria of model evaluation and selection: GOF, complexity, and generalizability. Model complexity is along the x axis and model fit along the y axis. GOF and generalizability are represented as curves whose performance can be compared as a function of complexity. The three smaller graphs contain the same data set (dots) and the fits to these data by increasingly more complex models (lines). The leftmost model underfits the data. The data are curvilinear whereas the model is linear. In this case, GOF and generalizability produce similar outcomes because the model is not complex enough to capture the bowed shape of the data. The model in the middle graph is a bit more complex and does a good job of fitting only the regularity in the data. Because of this, the GOF and generalizability measures are higher and also similar. Where the two functions diverge is when the model is more complex than is necessary to capture the main trend. The model in the right-most graph captures the experiment-specific noise, fitting every data point perfectly. GOF rewards this behavior by yielding an even higher fit score, whereas generalizability does just the opposite, penalizing the model for its excess complexity.

The problem of overfitting is the scourge of GOF. It is easy to see when overfitting occurs in Figure 1, but in practice it is difficult to know when and by how much a model overfits a data set, which is why generalizability is the preferred means of model evaluation and comparison. By using generalizability, we evaluate a model based on how well it predicts the statistics of future samples from the same underlying processes that generated an observed data sample.

## Model Comparison Methods

In this section we review measures of generalizability that are currently in use, touching on their theoretical foundations and discussing the pros and cons of their implementation. Readers interested in more detailed presentations are directed to two special issues on model selection in the *Journal of Mathematical Psychology* (Myung, Forster and Browne, 2000; Wagenmakers and Waldorp, 2006).

### Akaike Information Criterion and Bayesian Information Criterion

As illustrated in Figure 1, good generalizability is achieved by trading off GOF with model complexity. This idea can be formalized to derive model comparison criteria. That is, one way of estimating a model's generalizability is by appropriately discounting the model's goodness of fit relative to its complexity. In so doing, the aim is to identify the model that is sufficiently complex to capture the underlying regularities in the data but not unnecessarily complex to capitalize on random noise in the data, thereby formalizing the principle of Occam's razor.

The Akaike Information Criterion (AIC; Akaike, 1973; Bozdogan, 2000), its variation called the second-order AIC (AICc; Sugiura, 1978; Burnham and Anderson, 2002), and the Bayesian Information Criterion (BIC; Schwartz. 1978) exemplify this approach and are defined as

$$\begin{aligned} AIC &= -2\ln f(y\,|\,w^*) + 2k \\ AICc &= -2\ln f(y\,|\,w^*) + 2k + \frac{2k(k+1)}{n-k-1} \\ BIC &= -2\ln f(y\,|\,w^*) + k\ln(n) \end{aligned} \tag{1}$$

where $y$ denotes the observed data vector, $ln\ f(y|w^*)$ is the natural logarithm of the model's maximized likelihood calculated at the parameter vector $w^*$, $k$ is the number of parameters of the model, and $n$ is the sample size. The first term of each comparison criterion represents a model's lack of fit measure (i.e., inverse GOF), with the remaining terms representing the model's complexity measure. Combined, they estimate the model's generalizability such that the lower the criterion value, the better the model is expected to generalize.

AIC is derived as an asymptotic (i.e., large sample size) approximation to an information theoretic distance between two probability distributions, one representing the model under consideration and the other representing the "true" model (i.e., data-generating model). As such, the smaller the AIC value, the closer the model is to the "truth." AICc represents a small sample size version of AIC and is recommended for data with relatively small $n$ with respect to k, say $n/k < 40$ (Burnham and Anderson, 2002, p. 66). BIC, which is a Bayesian criterion as the name implies, is derived as an asymptotic expression of the minus two log marginal likelihood, which is described later in this article.

The above three criteria differ from one another in how model complexity is conceptualized and measured. The complexity term in AIC depends on only the number of parameters, *k*, whereas both AICc and BIC consider the sample size (*n*) as well, although in different ways. These two dimensions of a model are not the only ones relevant to complexity, however. Functional form, which refers to the way the parameters are entered in a model's equation, is another dimension of complexity that can also affect a model's data fitting capability (Myung and Pitt, 1997). For example, two models, $y = ax^b + e$ and $y = ax + b + e$, with a normal error *e* of constant variance, are likely to differ in complexity, despite the fact that they both assume the same number of parameters. For models such as these, the above criteria are not recommended because they are insensitive to the functional form dimension of complexity. Instead, we recommend the use of the comparison methods, described next, which are sensitive to all three dimensions of complexity.

**Cross-validation and Accumulative Prediction Error**

Cross-validation (CV; Stone, 1974; Browne, 2000) and the Accumulative Prediction Error (APE: Dawid, 1984; Wagenmakers, Grunwald and Steyvers, 2006) are sampling-based methods for estimating generalizability from the data, without relying on explicit, complexity-based penalty terms as in AIC and BIC. This is done by artificially simulating the data collection and prediction steps using the observed data in the experiment.

CV and APE are applied by following a three-step procedure: (1) divide the observed data into two sub-samples, the calibration sample, $y_{cal}$, simulating the "current" observations and the validation sample, $y_{val}$, simulating "future" observations; (2) fit the model to $y_{cal}$ and obtain the best-fitting parameter values, denoted by $w^*(y_{cal})$; and (3) with the parameter values fixed, the model is fitted to $y_{val}$. The resulting prediction error is taken as the model's generalizability estimate.

The two comparison methods differ from each other in how the data are divided into calibration and validation samples. In CV, each set of *n-1* observations in a data set serves as the calibration sample, with the remaining observation treated as the validation sample on which the prediction error is calculated. Generalizability is estimated as the average of *n* such prediction errors, each calculated according to the above three-step procedure. This particular method of splitting the data into calibration and validation samples is known as the leave-one-out CV in statistics. Other methods of splitting data into two sub-samples can also be used. For example, the data can be split into two equal halves, or into two sub-samples of different sizes. In the remainder of this article, CV refers to the leave-one-out cross validation procedure.

In contrast to CV, in APE the size of the calibration sample increases successively by one observation at a time for each calculation of prediction error. To illustrate, consider a model with *k* parameters. We would use the first *k+1* observations as the calibration sample so as to make the model identifiable, and the *(k+2)-th* observation as the validation sample, with the remaining observations not being used. The prediction error for the validation sample is then calculated following the three-step procedure. This process is then repeated by expanding the calibration sample to include the *(k+2)-th* observation, with the validation sample now being the *(k+3)-th* observation, and so on. Generalizability is estimated as the average prediction error over the *(n-k-1)* validation samples. Time series data are naturally arranged in an ordered list, but for data that have no natural order, APE can be estimated as the mean over all orders (in theory), or over a few randomly selected orders (in practice). Figure 2 illustrates how CV and APE are estimated.

------------------------

Figure 2 about here

------------------------

Formally, CV and APE are defined as

$$\begin{aligned} CV &= -\sum_{i=1}^{n} \ln f\left(y_i \mid w^*(y_{\neq i})\right) \\ APE &= -\sum_{i=k+2}^{n} \ln f\left(y_i \mid w^*(y_{1,2,\ldots,i-1})\right) \end{aligned} \qquad (2)$$

In the above equation for CV, $-\ln f\left(y_i \mid w^*(y_{\neq i})\right)$, is the minus log likelihood for the calibration sample $y_i$ evaluated at the best-fitting parameter values $w^*(y_{\neq i})$, obtained from the validation sample $y_{\neq i}$. The subscript signifies "all observations except for the *i-th* observation." APE is defined similarly. Both methods prescribe that the model with the smallest value of the given criterion should be preferred.

The attractions of CV and APE are the intuitive appeal of the procedures and the computational ease of their implementation. Further, unlike AIC and BIC, both methods consider, albeit implicitly, all three factors that affect model complexity: functional form, number of parameters, and sample size. Accordingly, CV and APE should perform better than AIC and BIC, in particular when comparing models with the same number of parameters. Interestingly, theoretical connections exit between AIC and CV, and BIC and APE. Stone (1977) showed that under certain regularity conditions, model choice under CV is asymptotically equivalent to that under AIC. Likewise, Barron, Rissanen and Yu (1998) showed that APE is asymptotically equivalent to BIC.

## Bayesian Model Selection and Stochastic Complexity

Bayesian Model Selection (BMS; Kass and Raftery, 1995; Wasserman, 2000) and Stochastic Complexity (SC; Rissanen, 1996 and 2001; Grunwald, Myung and Pitt, 2005; Myung. Navarro and Pitt, 2006) are the current state-of-the-art methods of model comparison. Both methods are rooted on firm theoretical foundations, are non-asymptotic in that they can be used for data of all sample sizes, small or large, and finally, are sensitive to all dimensions of complexity. The price to pay for this generality is computational cost. Implementation of the methods can be non-trivial because they usually involve evaluating high-dimensional integrals numerically.

BMS and SC are defined as

$$\begin{aligned} BMS &= -\ln \int f(y \mid w)\pi(w)dw \\ SC &= -\ln f(y \mid w^*) + \ln \int f(z \mid w^*(z))dz \end{aligned} \tag{3}$$

BMS is defined as the minus logarithm of the marginal likelihood, which is nothing but the mean likelihood of the data averaged across parameters and weighted by the parameter prior *π(w)*. The first term of SC is the minus log maximized likelihood of the observed data *y*. It is a lack of fit measure, as in AIC. The second terms is a complexity measure, with the symbol *z* denoting the *potential data* that could be observed in an experiment, not the actually observed data. Both methods prescribe that the model that minimizes the given criterion value is to be chosen.

BMS is related to the Bayes factor, the gold standard of model comparison in Bayesian statistics, such that the Bayes factor is a ratio of two marginal likelihoods between a pair of models. BMS does not yield an explicit measure of complexity but complexity is taken into account implicitly through the integral and thus avoids overfitting. To see this, an asymptotic expansion of BMS under Jeffrey's prior for *π(w)* yields the following large sample approximation (Balasubramanian, 1997)

$$BMS \approx -\ln f(y \mid w^{*}) + \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) + \ln \int \sqrt{\det(I(w))}\, dw \qquad (4)$$

where *I(w)* is the Fisher information matrix of sample size 1 (e.g., Schervish, 1995). The second and third terms on the right hand side of the expression represent a complexity measure. It is through the Fisher information in the third term that BMS reflects the functional form dimension of model complexity. For instance, the two models mentioned earlier, $y = ax^b + e$ and $y = ax + b + e$, would have different values of the Fisher information, though they both have the same number of parameters. The Fisher information term is independent of sample size *n*, with its relative contribution to that of the second term becoming negligible for large *n*. Under this condition, the above expression reduces to another asymptotic expression, which is essentially one-half of BIC in Eq. (1).

SC is a formal implementation of the principle of minimum description length that is rooted in algorithmic coding theory in computer science. According to the principle, a model is viewed as a code with which data can be compressed, and the best model is the one that provides maximal compression of the data. The idea behind this principle is that regularities in data necessarily imply the presence of statistical redundancy, which a model is designed to capture, and therefore, the model can be used to compress the data. That is, the data are re-expressed, with the help of the model, in a coded format that provides a shorter description than when the data are expressed in an uncompressed format. The SC criterion value in Eq. (3) represents the overall description length in bits of the maximally compressed data and the model itself, derived for parametric model classes under certain statistical regularity conditions (Rissanen, 2001).

The second (complexity) term of SC deserves special attention because it provides a unique conceptualization of model complexity. In this formulation, complexity is defined as the logarithm of the sum of maximized likelihoods that the model yields collectively for all *potential* data sets that could be observed in an experiment. This formalization captures nicely our intuitive notion of complexity. A model that fits well a wide range of data patterns, actual or hypothetical, should be more complex than a model that fits well only a few data patterns, but does poorly otherwise. A serious drawback of this complexity measure is that it can be highly non-trivial to compute the quantity because it entails numerically integrating the maximized likelihood over the entire data space. This integration in SC is even more difficult than in BMS, because the data space is generally of much higher dimension than the parameter space.

Interestingly, a large-sample approximation of SC yields Eq. (4) (Rissanen, 1996), which itself is an approximation of BMS. More specifically, under Jeffrey's prior, SC and BMS become asymptotically equivalent. Obviously, this equivalence does not extend to other priors and does not hold if the sample size is not large enough to justify the asymptotic expression.

## Model Comparison at Work: Choosing between Protein Folding Models

In this section we apply five model comparison methods to discriminating two protein folding models.

In the modern theory of protein folding, the biochemical processes responsible for the unfolding of helical peptides is of interest to researchers. The Zimm-Bragg theory provides a general framework under which one can quantify the helix-coil transition behavior of polymer chains (Zimm and Bragg, 1959). Scholtz, Barrick, York, Stewart and Baldwin (1995) applied the theory “to examine how the α-helix to random coil transition depends on urea molarity for a homologous series of peptides.” (p. 185). The theory predicts that the observed mean residue ellipticity *q* as a function of the length of a peptide chain and the urea molarity is given by

$$q = f_H \cdot (g_H - g_C) + g_C \tag{5}$$

In the above equation, $f_H$ is the fractional helicity and $g_H$ and $g_C$ are the mean residue ellipticities for helix and coil, respectively, defined as

$$\begin{aligned} f_H &= \frac{rs}{(s-1)^3}\left(\frac{n\cdot s^{n+2}-(n+2)s^{n+1}+(n+2)s-n}{n\left(1+\left[rs/(s-1)^2\right]\left[s^{n+1}+n-(n+1)s\right]\right)}\right) \\ g_H &= H_0\left(1-\frac{2.5}{n}\right)+H_U\cdot[urea] \\ g_C &= C_0 + C_U\cdot[urea] \end{aligned} \tag{6}$$

where *r* is the helix nucleation parameter, *s* is the propagation parameter, *n* is the number of amide groups in the peptide, $H_0$ and $C_0$ are the ellipticities of the helix and coil, respectively, at 0°C in the absence of urea, and finally, $H_U$ and $C_U$ are the coefficients that represent the urea dependency of the ellipticities of the helix and coil (Scholtz et al, 1995; Greenfield, 2004).

We consider two statistical models for urea-induced protein denaturation that determine the urea dependency of the propagation parameter *s*. One is the linear extrapolation method model (LEM; Pace and Vanderburg, 1979) , and the other is what is the binding-site model (BIND; Pace, 1986). Each expresses the propagation parameter *s* in the following form

$$\begin{aligned} &\text{LEM:}\quad \ln s = \ln s_0 - \frac{m\cdot[urea]}{R\cdot T} \\ &\text{BIND:}\ \ln s = \ln s_0 \text{-} d\cdot\ln\left(1\text{+k}\cdot(0.9815\cdot[urea]\ \text{-}\ 0.02978\cdot[urea]^2 + 0.00308\cdot[urea]^3)\right) \end{aligned} \tag{7}$$

where $s_0$ is the *s* value for the homopolymer in the absence of urea, *m* is the change in the Gibbs energy of helix propagation per residue, $R$ = 1.987 cal mol $^{-1}$ K$^{-1}$, *T* is the absolute temperature, *d* is the parameter characterizing the difference in the number of binding sites between the coil and helix forms of a residue, and *k* is the binding constant for urea.

Both models share four parameters: $H_{0}$,, $C_{0}$,, $H_U$, $C_U$. LEM has two parameters of its own ($s_0$, *m*), yielding a total of six parameters to be estimated from the data. BIND has three unique parameters ($s_0$, *d, k*). Both models are designed to predict the mean residue ellipticity denoted *q* in terms of the chain length *n* and the urea molarity *[urea]*. The helix nucleation parameter *r* is assumed to be fixed to the previously determined value of 0.0030 (Scholtz, Qian, York, Stewart and Baldwin, 1991).

------------------------

Figure 3 about here

------------------------

Figure 3 shows simulated data (symbols) and best-fit curves for the two models (LEM in solid lines and BIND in dotted lines). The data were generated from LEM for a set of parameter values with normal random noise of zero mean and one standard deviation added to the ellipticity prediction in Eq. (5) (see the figure caption for details). Note how closely both models fit the data. By visual inspection, one cannot tell which of the two models generated the data. As a matter of fact, BIND, with one extra parameter than LEM, provides a better fit to the data than LEM (SSE = 12.59 vs. 14.83), event though LEM generated the data. This outcome is an example of the over-fitting that can emerge with complex models, as depicted in Figure 1. To appropriately filter out the noise-capturing effect of overly complex models, and thereby put both models on an equal footing, we need the help of statistical model comparison methods that neutralize complexity differences.

We conducted a model recovery simulation to demonstrate the relative performance of five model comparison methods (AIC, AICc, BIC, CV and APE) in choosing between the two models. BMS and SC were not included because of the difficulty in computing them for these models. A thousand data sets of twenty-seven observations each were generated from each of the two models, using same nine points of urea molarity (0, 1, 2, ..., 8) for three different chain lengths of n =13, 20, 50. The parameter values used to generated the simulated data were taken from Table I of Scholtz et al (1995) and were as follows: ($H_0$ = *-44,000,* $C_0$ = *4,400,* $H_U$ = *320,* $C_U$ = *340* , $s_0$ = *1.34,* $m$ = *23.0*) and temperature $T$ = *273.15* for LEM; and ($H_0$ = *-42,500,* $C_0$ = *5,090,* $H_U$ = *-620,* $C_U$ = *280* , $s_0$ = *1.39,* $d$ = *0.52,* $k$ = *0.14*) for BIND. Normal random errors of zero mean and standard deviation of 1 were added to the ellipticity prediction in Eq. (5).

The five model comparison methods were compared on their ability to recover the model that generated the data. A good method should be able to identify the true model (i.e., the one that generated the data) 100% of the time. Deviations from perfect recovery reveal a bias in the selection method. (The Matlab code that implements the simulations can be obtained from the first author.)

------------------------

Table I about here

------------------------

The simulation results are reported in Table I. Values in the cells represent the percentage of samples in which a particular model (e.g., LEM) fitted best data sets generated by one of the models (LEM or BIND). A perfect selection method would yield values of 100% along the diagonal. The top 2 x 2 matrix shows model recovery performance under ML, a purely goodness of fit measure. It is included as a reference against which to compare performance when measures of model complexity are included in the selection method. How much does model recovery improve when the number of parameters, sample size, and functional form are taken into account?

With ML, there is a strong bias toward BIND. The result in the first column of the matrix shows that BIND was chosen more often than the true data-generating model, LEM (53% vs. 47%). This bias is not surprising given that BIND, with one more parameter than LEM, can capture random noise better than LEM. Consequently, BIND tends to be selected more often than LEM under a goodness-of fit-selection method such as ML, which ignores complexity differences. The results from using AIC show that when the difference in complexity due to the number of parameters is taken into account, the bias is largely corrected (19% vs. 81%), and even more so under AICc and BIC, both of which consider sample size as well (7% vs. 93% and 9% vs. 91%, respectively). When CV and APE were used, which are supposed to be sensitive to all dimensions of complexity, the results show that the bias was also corrected, although the recovery rate under these criteria was about equal to or slightly lower than that under AIC. When the data were generated from BIND (right column of values), the data generating model was selected more often than the competing model under all selection methods, including ML.

To summarize, the above simulation results demonstrate the importance of considering model complexity in model comparison. All five model selection methods performed reasonably well by compensating for differences in complexity between models, and thus identifying the data-generating model. It is interesting to note that Scholtz et al (1995) evaluated the viability of the same two models plus a third, seven-parameter model, using goodness-of-fit, and found that all three models provided nearly identical fits to their empirical data. Had they compared the models using one of the selection methods discussed in this article, it might have been possible to obtain a more definitive answer.

We conclude this section with the following cautionary note regarding the performance of the five selection methods in Table I: The better model recovery performance of AIC, AICc and BIC over CV and APE should not be taken as indicative of how the methods will generally perform in other settings (Myung & Pitt, 2004). There are very likely other situations in which the relative performance of the selection methods reverses.

## Conclusions

We began this article by discussing several issues a modeler should be aware of when evaluating computational models. They include the notion of model complexity, the triangular relationship among goodness of fit, complexity and generalizability, and generalizability as the ultimate yardstick of model comparison. We then introduced several model comparison methods that can be used to determine the "best-generalizing" model among a set of competing models, discussing the pros and cons of each method. Finally, we demonstrated the application of some of the comparison methods using simulated data for the problem of choosing between biochemical models of protein folding.

Measures of generalizability are not without their own drawbacks, however. One is that they can be applied only to statistical models defined as a parametric family of probability distributions. This restriction leaves one with few options when wanting to compare non-statistical models, such as verbal models and computer simulation models. Often times, researchers are interested in testing qualitative (e.g., ordinal) relations in data (e.g., condition $A < $ condition $B$), and comparing models on their ability to predict qualitative patterns of data, but not quantitative ones.

Another limitation of measures of generalizability is that they summarize the potentially intricate relationships between model and data into a single real number. After applying CV or BMS, the results can sometimes raise more questions than answers. For example, what aspects of a model's formulation makes it superior to its competitors? How representative is a particular data pattern of a model's performance? If it is typical, the model provides a much more satisfying account of the process than if the pattern is generated by the model using a small range of unusual parameter settings. Answers to these questions also contribute to the evaluation of model quality.

We have begun developing methods to address questions such as these. The most well developed method thus far is a global qualitative model analysis technique dubbed *parameter space partitioning* (PSP; Pitt, Kim, Navarro and Myung, 2006; Pitt, Myung and Altieri, 2007). In PSP, a model's parameter space is partitioned into disjoint regions, each of which corresponds to a qualitatively different data pattern. Among other things, using PSP, one can use PSP to identify all data patterns a model can generate by varying its parameter values. With information such as this in hand, one can learn a great deal about the relationship between the model and its behavior, including understanding the reason for the model's ability or inability to account for empirical data.

In closing, statistical techniques, when applied with discretion, can be useful for identifying sensible models for further consideration, thereby aiding the scientific inference process (Myung and Pitt, 1997). We cannot over-emphasize the importance of using non-statistical criteria such as explanatory adequacy, interpretability, and plausibility of the models under consideration, though they have yet to be formalized in quantitative terms and subsequently incorporated into the model evaluation and comparison methods. Blind reliance on statistical means is a mistake. On this point we agree with Browne and Cudeck (1992), who said "Fit indices [statistical model evaluation criteria] should not be regarded as a measure of usefulness of a model...they should not be used in a mechanical decision process for selecting a model. Model selection has to be a subjective process involving the use of judgement" (p. 253).

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrox and F. Caski, *Second International Symposium on Information Theory* (pp. 267-281). Akademia Kiado, Budapest.

Balasubramanian, V. (1997). Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation, 9*, 349-368.

Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE: Transactions on Information Theory, 44*, 2743-2760.

Berger, J. O. and Berry, D. A. (1998). Statistical analysis and the illusion of objectivity. *American Scientist, 76*, 159-165.

Bozdogan, H. (2000). Akaike information criterion and recent developments in information complexity. *Journal of Mathematical Psychology, 44*, 62-91.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*, 108-132.

Browne, M. W. and Cudeck, R. C. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*, 230-258.

Burnham, L. S., and Anderson, D. R. (2002). *Model selection and inference: A practical information-theoretic approach* (2nd edition). Springer-Verlag. New York.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A, 147, 278-292.*

Greenfield, N. J. (2004). Analysis of circular dichroism data. *Methods in Enzymology, 383*, 282 - 317.

Grunwald, P., Myung, I.J., and Pitt, M.A. (2005). *Advances in Minimum Description Length: Theory and Application*. Cambridge, MA: MIT Press.

Jacobs, A. M. and Grainger, J. (1994). Models of visual word recognition–sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 1311-1334.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Myung (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 44*, 190-204.

Myung, I. J., Forster, M., and Browne, M. W., eds. (2000). Special issue on model selection. *Journal of Mathematical Psychology, 44*, 1-2.

Myung, I. J., Navarro, D. J. and Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology, 50*, 167-179.

Myung, I. J. and Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*, 79-95.

Myung, I. J. and Pitt, M. A. (2004). Model comparison methods. In L. Brand & M. L. Johnson (Eds.), *Numerical Computer Methods, Part D (A volume of Methods in Enzymology, evaluation, testing and selection, vol. 383, pp.351-366).*

Pace, C. N. (1986). Determination and analysis of urea and guanidine hydrochloride denatiration curves. *Methods in Enzymology, 131*, 266-280.

Pace, C. N. and Vanderburg, K. E. (1979). *Biochemistry, 18*, 288-292.

Pitt, M.A., Kim, W., Navarro, D.J., and Myung, J.I. (2006). Global model analysis by parameter space partitioning. *Psychological Review,* **113**, 57-83.

Pitt, M. A. and Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*, 421-425.

Pitt, M. A., Myung, I. J., and Altieri , N. (2007). Modeling the word recognition data of

Vitevitch and Luce (1998): Is it ARTful? *Psychonomic Bulletin & Review, 14*, 442 - 448.

Rissanen, J, (1996). Fisher information and stochastic complexity. *IEEE Trans. Information Theory 42*, 40-47.

Rissanen, J, (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Trans. Information Theory 47*, 1712-1717.

Schervish, M. J. (1995). *The Theory of Statistics*. New York: Springer-Verlag.

Scholtz, J. M., Barrick, D., York, E. J., Stewart, J. M. and Balding, R. L. (1995). Urea unfolding of peptide helices as a model for interpreting protein unfolding. *Proceedings of the National Academy of Sciences USA, 92*, 185-189.

Scholtz, J. M., Qian, H., York, E. J., Stewart, J. M. and Balding, R. L. (1991). *Biopolymers, 31*, 1463-1470.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with Discussion). *Journal of the Royal Statistical Society, Series B, 36*, 111-147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B, 39*, 44-47.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics: Theory and Methods, A7*, 13-26.

Wagenmakers, E.-J., Grunwald, P., and Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology, 50*, 149-166.

Wagenmakers, E.-J. and Waldorp, L. (2006). Editors' introduction. *Journal of Mathematical Psychology, 50*, 99-100.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44*, 92-107.

Zimm, B, H, and Bragg, J. K. (1959). *Journal of Chemical Physics, 34*, 1963-1974.

## Author Notes

This work was supported by research grant R01-MH57472 from the National Institute of Health to JIM and MAP. This article is an updated version Myung and Pitt (2004). There is some overlap in content. Correspondence should be addressed to Jay Myung or Mark Pitt, Department of Psychology, Ohio State University, 1835 Neil Avenue, Columbus, OH, 43210. Electronic mail should be sent to Myung.1@osu.edu or Pitt.2@osu.edu.

Table I. Model Recovery Performance of Five Model Comparison Methods

| Model comparison method | Model fitted: | Data were generated from: | |
|---|---|---|---|
| | | LEM | BIND |
| ML | LEM | 47 | 4 |
| | BIND | 53 | 96 |
| AIC | LEM | 81 | 16 |
| | BIND | 19 | 84 |
| AICc | LEM | 93 | 32 |
| | BIND | 7 | 68 |
| BIC | LEM | 91 | 28 |
| | BIND | 9 | 72 |
| CV | LEM | 77 | 26 |
| | BIND | 23 | 74 |
| APE | LEM | 75 | 45 |
| | BIND | 25 | 55 |

*Note:* The two models, LEM and BIND, are defined in Eq. (7). APE was estimated after randomly ordering the twenty-seven data points of each data set.

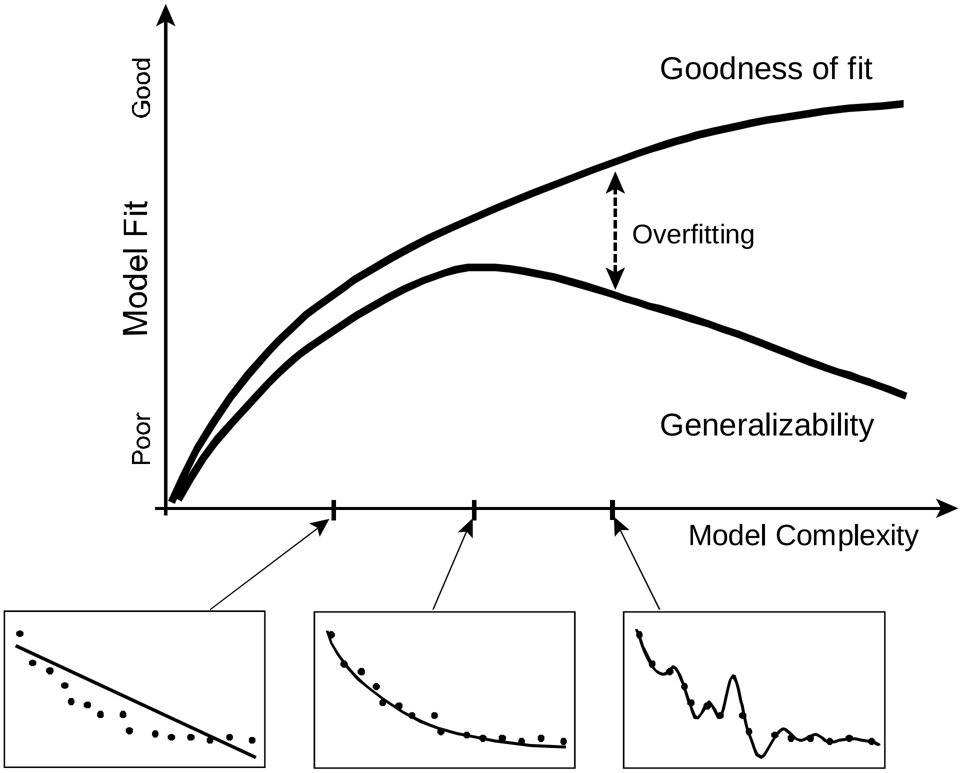

Figure 1. An illustration of the relationship between goodness of fit and generalizability as a function of model complexity. The y axis represents any fit index, where a larger value indicates a better fit (e.g., maximum likelihood). The three smaller graphs provide a concrete example of how fit improves as complexity increases. In the left graph, the model (line) is not complex enough to match the complexity of the data (dots). The two are well matched in complexity in the middle graph, which is why this occurs at the peak of the generalizability function. In the right graph, the model is more complex than the data, capturing microvariation due to random error. Reprinted from Pitt and Myung (2002).

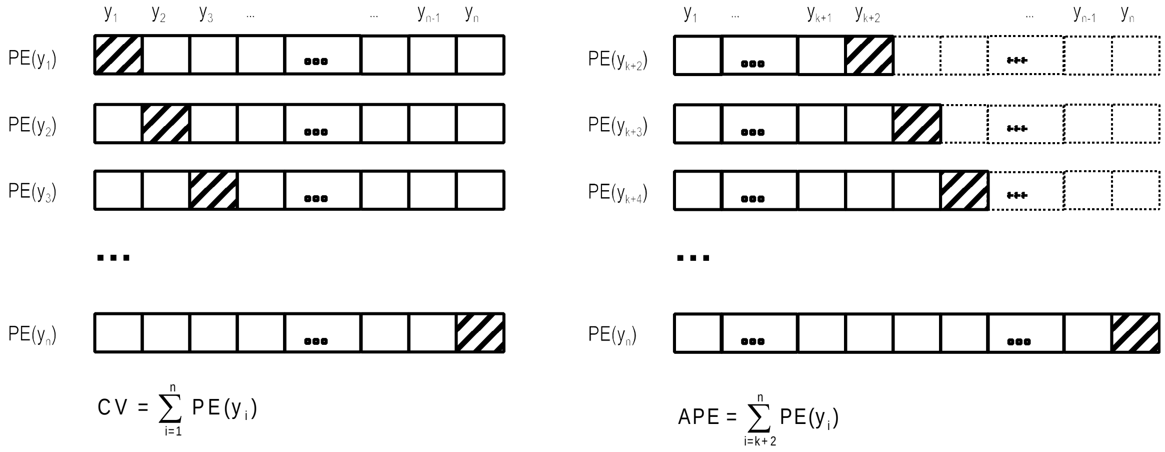
$y_1$ $y_2$ $y_3$ ... ... $y_{n-1}$ $y_n$
PE($y_1$)
PE($y_2$)
PE($y_3$)
...
PE($y_n$)
$CV = \sum_{i=1}^{n} PE(y_i)$
$y_1$ ... $y_{k+1}$ $y_{k+2}$ ... $y_{n-1}$ $y_n$
PE($y_{k+2}$)
PE($y_{k+3}$)
PE($y_{k+4}$)
...
PE($y_n$)
$APE = \sum_{i=k+2}^{n} PE(y_i)$

Figure 2. The difference between the two sampling-based methods of model comparison, cross-validation (CV) and Accumulative Prediction Error (APE), is illustrated. Each chain of boxes represents a data set with each data point represented by a box. The slant-lined box is a validation sample and the plain boxes with the bold outline represent the calibration sample. The plain boxes with the dotted outline in the right panel are not being used as part of the calibration or validation sample. The symbol *PE($y_i$), i = 1,2, ...n*, stands for the prediction error for the *i-th* validation data point.

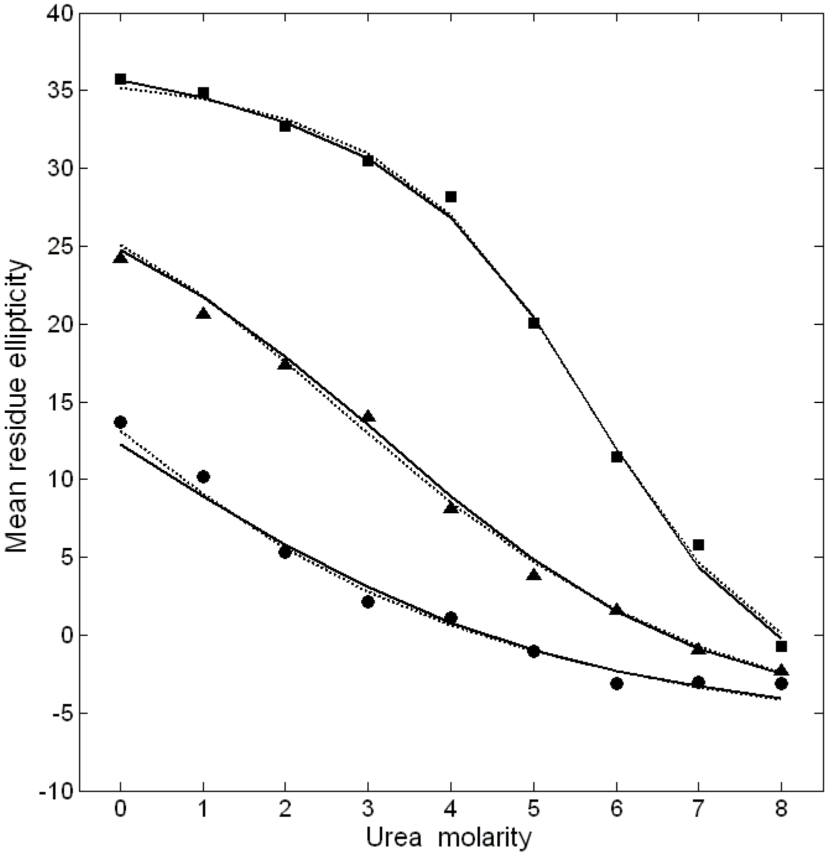

Figure 3. Best fits of the LEM (solid lines) and BIND (dotted lines) models to data generated from LEM using the nine points of urea molarity (0, 1, 2, ..., 8) for three different chain lengths of n =13 (circles), 20 (triangles), and 50 (squares). Data fitting was done first by deriving model predictions using Eqs. (5) - (7) based on the parameter values of $H_0 = -44{,}000$, $C_0 = 4{,}400$, $H_U = 320$, $C_U = 340$ , $s_0 = 1.34$, $m = 23.0$ reported in Scholtz et al (1995, Table 1). See text for further details.