ARTICLE

Applying Occam's razor in modeling cognition: A Bayesian approach

IN JAE MYUNG and MARK A. PITT
Ohio State University, Columbus, Ohio

In mathematical modeling of cognition, it is important to have well-justified criteria for choosing among differing explanations (i.e., models) of observed data. This paper introduces a Bayesian model selection approach that formalizes Occam's razor, choosing the simplest model that describes the data well. The choice of a model is carried out by taking into account not only the traditional model selection criteria (i.e., a model's fit to the data and the number of parameters) but also the extension of the parameter space, and, most importantly, the functional form of the model (i.e., the way in which the parameters are combined in the model's equation). An advantage of the approach is that it can be applied to the comparison of non-nested models as well as nested ones. Application examples are presented and implications of the results for evaluating models of cognition are discussed.

A goal of research in psychology, as in other behavioral sciences, is to infer the underlying process that generated observed data. The use of sophisticated mathematical models to describe these processes has grown considerably, especially in cognitive psychology (see, e.g., J. R. Anderson & Sheu, 1995; N. H. Anderson, 1981; Ashby & Townsend, 1986; Busemeyer & Townsend, 1993; Gillund & Shiffrin, 1984; Green & Swets, 1966; Hintzman, 1986; Kruschke, 1992; Massaro & Friedman, 1990; Medin & Schaffer, 1978; Murdock, 1982; Nosofsky, 1986; Oden & Massaro, 1978; Reed, 1972; van Zandt & Ratcliff, 1995). Yet, the development of equally sophisticated and well-justified methods for evaluating the adequacy of the models themselves has lagged behind. Jacobs and Grainger (1994) recently summarized a number of criteria for choosing among models: (1) generality (does the model generalize well across different experimental settings?); (2) explanatory adequacy (are the assumptions of the model plausible and compatible with established findings?); (3) descriptive adequacy (does the model fit the pattern of data well?); and (4) complexity (is the formulation [e.g., the number of parameters] of the model simple?). Among these, descriptive adequacy and complexity have been used most frequently, probably because they are easier to quantify than the other two. Together they embody the principle of Occam's razor, which states that "entities should not be multiplied beyond necessity" (William of Occam, ca. 1290–1349). The goal of model selection is to choose the *simplest* (i.e., least complex) model that describes the data *well* (i.e., descriptive adequacy).

In this paper, we introduce a new Bayesian method of formalizing Occam's razor in model selection. It goes beyond current selection methods by taking into account dimensions of complexity that are not captured by its predecessors. We begin with a tutorial on model selection methods that are currently in use. Next, fundamentals of the Bayesian model selection approach are described, and its desirable properties are discussed. The utility of the Bayesian approach is then demonstrated using concrete examples with simulated data. Finally, the merits and shortcomings of the Bayesian approach are discussed and contrasted with traditional approaches.

MODEL SELECTION CRITERIA

Descriptive Adequacy

The goal of mathematical modeling in cognitive psychology is straightforward: Given observed data, identify the underlying processes that generated the data. Because a model is defined as a set of assumptions about underlying processes, the goal of the researcher is to determine the viability of the model. There are, however, at least two obstacles to such an endeavor. First, given the nearly infinite number of distinct models that can be defined by combining different assumptions, the true model might

not be one of a particular set of models that is being tested. Second, random noise in data can obscure model identification. Consequently, a realistic goal of mathematical modeling is to choose the model that represents the closest approximation to the "true" model.

How can the closest approximation be identified when the true model has yet to be discovered? Consider the situation in which the true model is included in the set of models being tested and further, data are noise free. In this ideal situation, the true model must fit the data perfectly (e.g., as measured using a metric such as sum of squared errors). Note that this is a *necessary*—but not sufficient—condition, for there could be more than one model that fits the data perfectly. By extending this logic to less ideal situations (e.g., noisy data), the following model selection rule is obtained: Choose the model that provides the best fit to the data. Accordingly, the foremost criterion of model selection, *descriptive adequacy*, is born. Examples of descriptive adequacy measures that are in use include the percent variance accounted for by the model (i.e., coefficient of determination), the sum of squared errors (SSE) between observed and predicted outcomes, and the maximum likelihood, in which the probability of obtaining the observed data is maximized with respect to the model's possible parameter values (see Bickel & Doksum, 1977).

For a model to be considered true, it must satisfy the minimal condition of sufficiency in fitting data well. A failure to do so invalidates the model. An illustrative example is shown in Figure 1. Each of the four solid lines in the figure represents a model's best fit to the same data set (solid dots) using the least squares estimation method. Model 1 is a two-parameter linear model. As can be seen, it fits the data poorly, with only 79.5% of the variance accounted for. Systematic deviations from the line are evident at the endpoints and in the middle of the range.

Clearly, Model 1 fails the test of descriptive adequacy and can be dropped from further consideration. In contrast, Model 2, a three-parameter exponential model, fits the data fairly well, accounting for 96% of the variance with no systematic deviation from the fit. Between the first two models, Model 2 would be chosen as the preferred description of the data. Models 3 and 4 are discussed in the next section.

Model Complexity

It is important to note that descriptive adequacy is a heuristic. Selection of the best fitting model may be useful in identifying the true model or closest approximation, but the rule's accuracy is not guaranteed. This is because model fit can be improved by increasing model complexity. *Complexity* refers to the flexibility inherent in a model that enables it to fit diverse patterns of data.[1] It can be understood by contrasting the data-fitting capabilities of simple and complex models. A simple model is one that assumes that a specific pattern will be found in the data. If this pattern occurs, the model will fit the data well. Simple models make clear and falsifiable predictions precisely because a specific pattern is assumed to be present. In terms of actually testing the model, what this means is that the model's fit will be good over a sizable range of parameter values. A complex model, on the other hand, is more flexible than a simple model, providing good fits to a wide range of data patterns. To do so, however, the complex model's parameters must be finely tuned. This is because as the model's parameters change, even slightly, the postulated data pattern also changes.

There are at least three dimensions of a model that contribute to its complexity, thereby significantly affecting model fit: the number of parameters, the model's functional form, and the extension of the parameter space. In
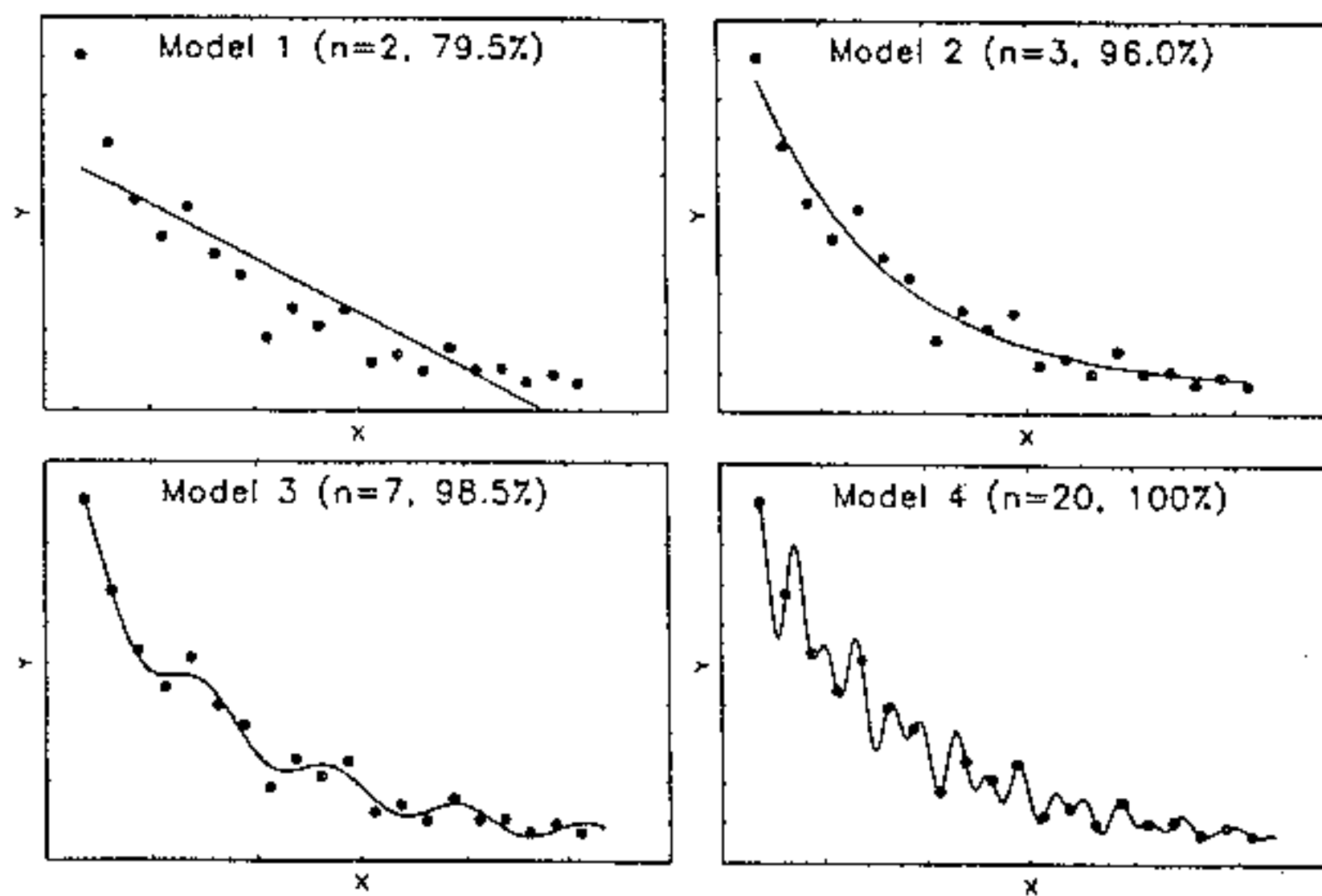

Figure 1. The effect of the number of parameters in a model on the model's ability to fit data. A single data set (dots) was fitted to four models (lines) differing in the number of parameters (*n*). Percentage of variance accounted for by each model is shown in parentheses.

the following subsections, we discuss the implications of each of these for model selection.

Number of parameters. In general, a model with many parameters fits data better than a model with few parameters, even if the latter generated the data (Collyer, 1985). The effect of excessive parameters on model fit is illustrated in the bottom panels in Figure 1. Model 3 was created from Model 2 by introducing an additional cyclic component with four new parameters. Its fit is improved over Model 2, but only by a meager 2.5%. The extra parameters seem unnecessary, capturing what appear to be a few idiosyncrasies in the data. Model 4 is a more dramatic example of a model with excessive parameters. It fits the data perfectly and even tells us more than the data do. Model 4 is an interpolation model with 20 parameters, the same number as data points. Such a model can always be found and by design, it will fit the data perfectly. A problem with Models 3 and 4 is that they generalize poorly to other data because they precisely fit only one data set. Thus it seems unlikely that these models accurately reflect the mental processes responsible for generating the data.

The preceding examples should make it clear that there is a tradeoff between fit and generalizability as the number of parameters increases. The best fit should be preferred when it is not achieved at the expense of additional parameters. (For a discussion of the related issue of the theoretical justification of the extra parameters, see Jacobs & Grainger, 1994, and Cudeck & Henly, 1991.) Three model selection methods were proposed that adjust for variation in the number of parameters among models. They do so by penalizing more heavily models with many parameters as opposed to those with few parameters.

Akaike (1973, 1983) introduced the Akaike information criterion (AIC), which is defined as

$$AIC_i = -2\ln(ML_i) + 2n_i. \qquad (1)$$

In this equation, ML_i is the maximum likelihood for Model i and n_i is the number of free parameters in the model. The criterion prescribes that the model that minimizes the AIC should be chosen. If the choice is between two models with equal maximum likelihood values but different numbers of parameters, AIC favors the model with fewer parameters. The AIC has been employed in time series analysis (e.g., Cryer, 1986), psychometric data analysis (e.g., Bozdogan, 1987), and mathematical modeling of categorization (e.g., Maddox & Ashby, 1993; Takane & Shibayama, 1992).

Schwarz (1978) introduced another criterion called the Bayesian information criterion (BIC), defined as

$$BIC_i = -2\ln(ML_i) + 2n_i \ln s \qquad (2)$$

where $\ln(s)$ is the natural log of sample size (s: number of observations per stimulus) of the data. The model that minimizes BIC should be chosen. Note its similarity to AIC. Only the penalty term for excessive parameters is slightly different ($2n_i$ vs. $n_i \ln(s)$). A comparison of both criteria shows that AIC favors more complex models than BIC for large sample sizes when $\ln(s) > 2$ (i.e., $s > 8$).

Steiger and colleagues (Steiger, 1990; Steiger & Lind, 1980; see also Browne & Cudeck, 1992, Cudeck & Henly, 1991) introduced yet another criterion, called the root mean square error of approximation (RMSEA), defined as

$$RMSEA = \sqrt{\frac{F_i}{N - n_i}}. \qquad (3)$$

In this equation, the function F_i is a measure of the lack of fit for model i (e.g., using SSE) and N is the data size.[2] RMSEA penalizes complex models by subtracting the number of parameters (n_i) from the divisor, N. The RMSEA closely approximates the root mean squared deviation (RMSD), which is a sample statistic often used in mathematical modeling (see, e.g., Massaro & Friedman, 1990).

Functional form. The number of parameters is the only dimension of model complexity that AIC, BIC, and RMSEA consider. An often unrecognized dimension of model complexity that can significantly affect model fit by simply capturing irrelevant patterns of data is *functional form*, which can be defined as the way in which parameters are combined in the model equation.

Figure 2 illustrates the effect that functional form can have on model fit. There is a universe (U) of possible data, subsets of which are consistent with different models. The set M_a denotes a data region that Model M_a fits "well" in an appropriately defined sense (e.g., maximum likelihood). Model M_b, which has more parameters than M_a but the same functional form, accounts for a wider range of data. Now consider the third model, M_c, which has the same number of parameters as Model M_a but assumes a functional form different from that of M_a. Not only is the M_c region larger than that of M_a (i.e., more data are accounted for by M_c) but also, most of the data that M_a fits can also be fitted by M_c. Importantly, the reverse is not true. Thus the functional form of Model M_c makes it a more flexible model than M_a.

Another example of the importance of functional form in model behavior is a comparison of two psychophysical models. Townsend (1975) pointed out that Stevens's law ($\Psi(x) = k \cdot x^a$) is more complex, and thus less falsifiable, than Fechner's law ($\Psi(x) = k \cdot \ln(x + \beta)$), even though both have the same number of parameters. This is
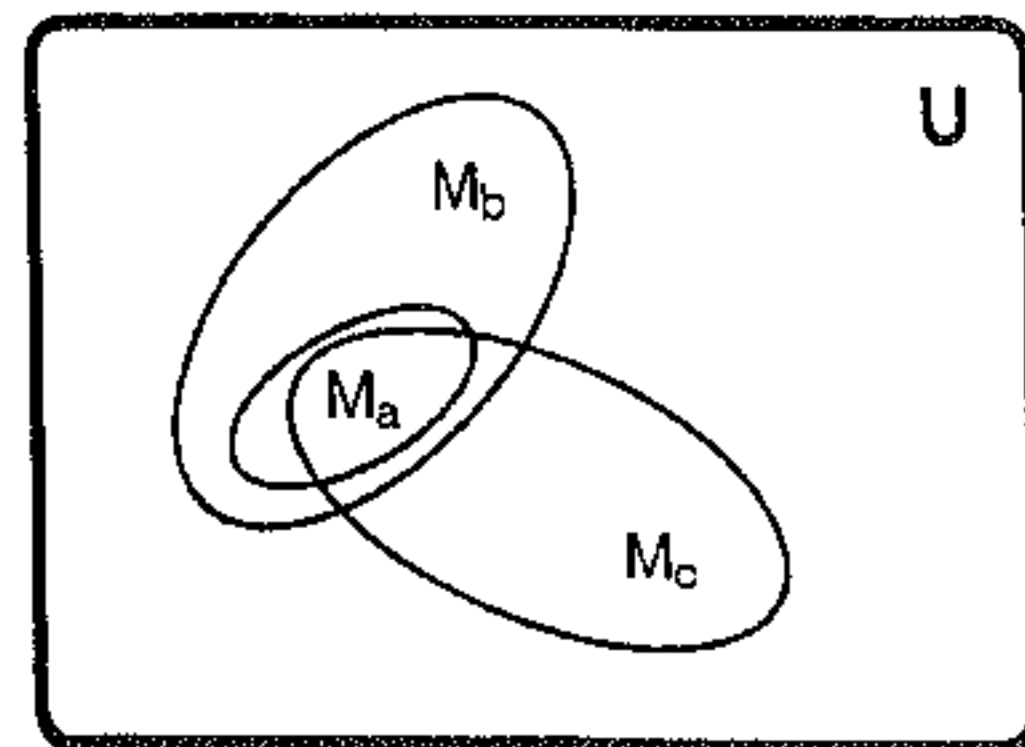

Figure 2. Data-fitting capabilities of various models.

because psychological and physical dimensions are assumed to be related by a power function in Stevens's law, making it capable of fitting data that have negative, positive, and zero curvature. Fechner's law assumes a logarithmic relationship, which can fit data patterns with a negative curvature only.

More recently, the superfluous effects of functional form in model fitting were shown in an insightful simulation study by Cutting, Bruno, Brady, and Moore (1992). They compared two *non-nested* models of perception, the linear integration model (LIM; Anderson, 1981) and the fuzzy logic model of perception (FLMP; Oden & Massaro, 1978).[3] The two models have the same number of parameters but different functional forms, a linear additive function in LIM and a nonlinear multiplicative function in FLMP. In the first simulation, data patterns were generated that spanned the range of the parameter space. Both models were then fitted to all data patterns. Results showed that FLMP was more flexible than LIM, providing a superior fit (i.e., smaller RMSD) for 80.3% of the data patterns with exponential functions and 95.7% with logistic functions. If the number of parameters were the only factor affecting model complexity, the percentage of superior fits should have been about 50%.

A more convincing demonstration of FLMP's superior functional form can be found in the results from another simulation, in which the models were fitted to a set of random numbers (dependent variable) generated from a uniform distribution ranging from 0 to 1. Again, FLMP performed at better than chance level (60.8%). Further exploring this issue, Cutting et al. (1992) also fitted the models to noisy data whose means were generated by LIM. Again, FLMP fitted the data better than LIM itself about 60% of the time, and did so even more decisively when the added noise was relatively large. Given that LIM produced the data, the percentage of superior fits by FLMP should have been at least below the chance level of 50%, though ideally it should have been zero.

Massaro and Cohen (1993) argued that Cutting et al.'s (1992) simulations distorted the true data-fitting abilities of FLMP. Massaro and Cohen stated that in the simulations in which the FLMP and LIM were fitted to a range of exponential and logistic functions, the functions were ones that favored FLMP, and that a set of plausible functions could be created that could just as easily cause the LIM to provide a superior fit. This criticism would have been more convincing if supporting evidence had been provided. Massaro and Cohen also claimed that the simulations in which the models were fitted to random data provided an unfair test of the models, because both models fit the data so poorly (very large RMSDs) that FLMP's superior fits are not meaningful. This reasoning ignores the consistency with which FLMP provided better fits than LIM and places an arbitrary limit on the interpretation of goodness-of-fit measures. Regardless of the quality of the fit, the simulations clearly show FLMP's superior ability at fitting data. Finally, Massaro and Cohen argued that the simulations using noisy data were invalid because Cutting et al. had used an artificially narrow range of data values (between 0.3 and 0.7, whereas observed values often span from 0 to 1). Our own simulations in the present paper go a long way toward dispelling this criticism. Data values ranged between 0.1 and 0.9, and we consistently found that FLMP provided superior fits. Furthermore, we also found that LIM *never* provided superior fits to those provided by FLMP. If the two models could mimic each other equally well (i.e., if they were equally flexible), then LIM should have fit FLMP data as well as FLMP fit LIM data. Without assuming that FLMP is more flexible than LIM, it is difficult to explain the simulation results, in particular the asymmetric pattern of model mimicry.

The point of this discussion is not to argue against the psychological validity of FLMP or LIM, but rather to make clear the importance of functional form in model selection. Because both models have the same number of parameters, FLMP's advantage must be due to the functional form of the model equation. The nonlinear function of FLMP seems more flexible than the linear function of LIM in fitting noisy data, thus improving model fit.

Extension of parameter space. Another dimension of model complexity that can also influence model fit is the extension of the parameter space. To illustrate the point, consider two one-parameter models that share the same functional form of $y = 1/(1 + e^{-\theta x})$, but assume different ranges of the parameter θ. In Model 1, θ ranges from $-R$ to R, where R is a constant. In Model 2, θ's range is cut in half, spanning from 0 to R. Model 1 allows the parameter to be either positive or negative, whereas Model 2 allows it to be only positive. The parameter range of Model 1 is twice that of Model 2. As shown in Figure 3, this difference in parameter space means that Model 1 can fit decreasing ($\theta < 0$) and increasing ($\theta > 0$) patterns of data. Model 2 can fit only the latter. Solely by virtue of the larger parameter space, Model 1 is guaranteed to fit data showing a decreasing pattern better than Model 2.

To summarize, mathematical models can vary along a number of dimensions. It is necessary to incorporate all of them into the model selection procedure to maximize its accuracy. Selection of a model based solely on descriptive adequacy can lead to incorrect conclusions unless appropriate adjustments are made for the three dimensions of model complexity (number of parameters, functional form, and extension of parameter space), which can significantly and independently affect model fit. Standard model selection criteria such as AIC and BIC consider only the first of these.[4] Such criteria are appropriate to use when comparing nested models, which tends to be done in psychometric research (e.g., analysis of variance models, regression models, latent variable models). However, their use in comparing non-nested models, which is virtually always the case in models of cognition, is not justified and can lead to erroneous conclusions, as we demonstrate in the section Application Example of Bayesian Model Selection.

The Bayesian method described next provides a way to overcome the limitation of the standard model selection
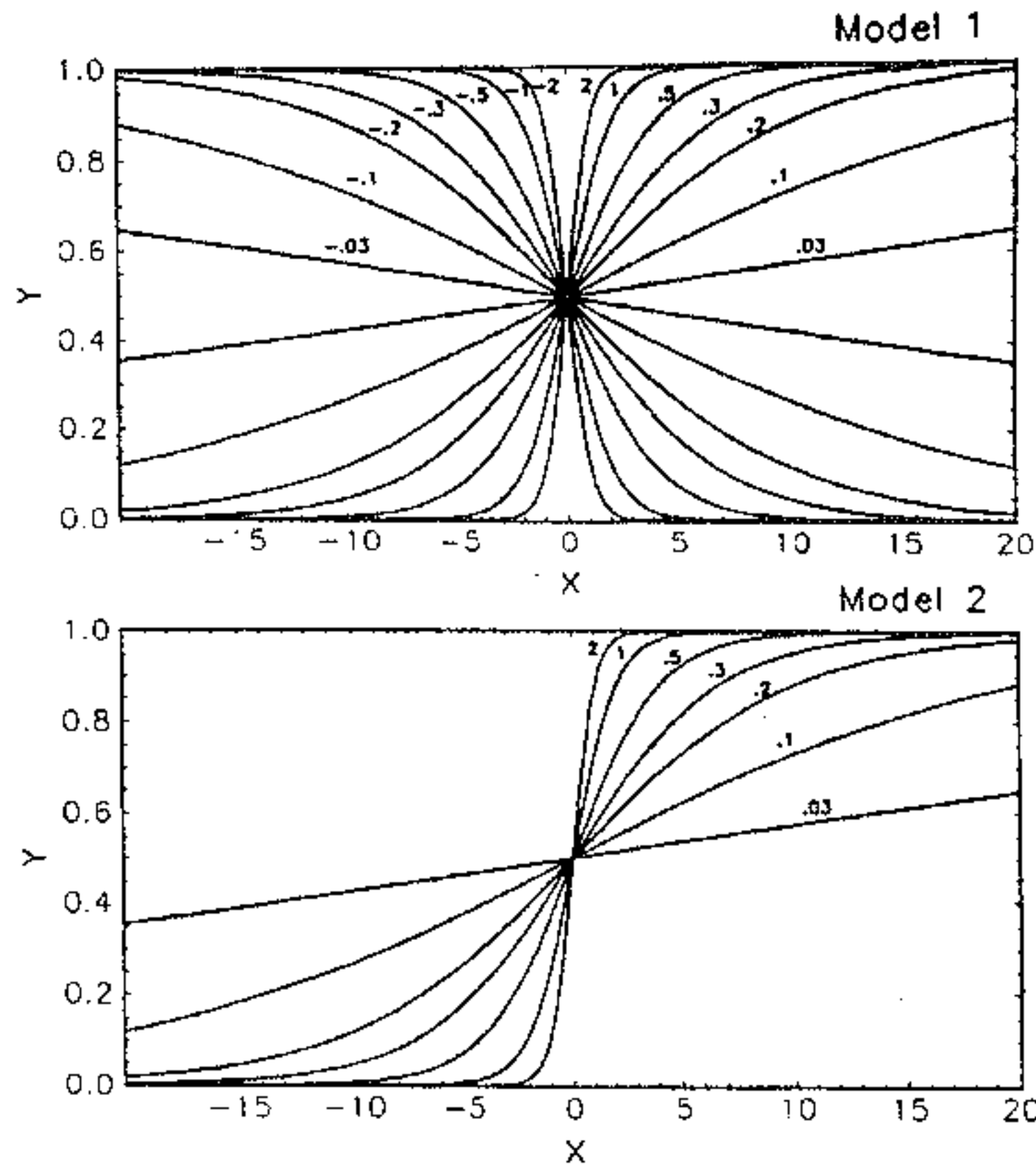
Figure 3. An example of how the extension of the parameter space can affect model fit. A range of possible data patterns that each of two different models can fit is shown. The two models share the same functional form of $y = 1/(1 + e^{-\theta x})$ but assume different ranges of the parameter θ: $-R < \theta < R < \infty$ for Model 1; $0 < \theta < R < \infty$ for Model 2. Values of R are listed next to the corresponding function.

criteria. The method combines *all three* dimensions of model complexity, as well as descriptive adequacy, into a single measure, thus making it particularly suitable for testing models of cognition.[5]

BAYESIAN MODEL SELECTION

We begin this section by describing the basics of the Bayesian model selection approach, followed by a consideration of the computational issues surrounding its implementation.

Fundamentals of Bayesian Model Selection

Recently there has been a surge of interest in Bayesian methods of model selection among statisticians, mathematicians, physicists, chemists, and neural network researchers (e.g., Berger & Perrichi, 1996; Bretthorst, 1989; Carlin & Chib, 1995; Gelfand & Dey, 1994; Gregory & Loredo, 1992; W. H. Jeffreys & Berger, 1992; Le & Raftery, 1996; MacKay, 1992; Raftery, 1993, 1994; Rissanen, 1986, 1990; Smith & Roberts, 1993). In this section, we provide a comprehensive overview of Bayesian model selection methods, emphasizing their utility in evaluating models of cognition and discussing their implications for mathematical modeling of psychological data (see Kass & Raftery, 1995, for a more technically rigorous presentation of the material).

Application of the Bayesian approach requires a statistical formulation of a model that specifies the probability density function (i.e., distribution) of the data. For example, a model of semantic priming may assume that response times in the lexical decision task follow a normal probability density with a certain mean and standard deviation. Moreover, the mean may increase or decrease depending on theoretical predictions, whereas the standard deviation is an unknown but fixed quantity.

Model selection is carried out by computing the posterior probability that each of the models is correct (i.e., true) given a particular data set. This process is formally stated as follows:

For each model M_i and the data D, compute the posterior probability, $P(M_i|D)$, from the prior probability, $P(M_i)$, and the evidence for M_i, $P(D|M_i)$, using the Bayes rule, and then choose the model that maximizes the posterior probability.

For simplicity, we will consider only the two-model case of models M_1 and M_2, although extension to multimodel cases is straightforward. From the Bayes rule, we obtain

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(M_1)}{P(M_2)} \cdot \frac{P(D|M_1)}{P(D|M_2)} . \quad (4)$$

The first term, $P(M_1|D)/P(M_2|D)$, is called the posterior odds ratio, and represents the ratio of the model likelihoods given the data. The second term, $P(M_1)/P(M_2)$, is called the prior odds ratio; it represents the ratio of the model likelihoods before evaluating the data. Model priors are determined independently of the data. The third term, $P(D|M_1)/P(D|M_2)$ is called the Bayes factor or evidence ratio. The numerator and denominator each represent the probabilities of the data given each model. Model selection is achieved by computing the prior odds ratio and the Bayes factor. In other words, the equation above can be rewritten as

posterior odds ratio = (prior odds ratio) × (Bayes factor).

Prior odds $P(M_i)$. $P(M_i)$ represents the probability that the model M_i is a true description of the events under study *before* data are collected. For real-world problems, this probability may not exist, or not be known even if it does. We assume equal model priors; that is, $P(M_1)/P(M_2) = 1$. This assumption is often made by researchers in Bayesian model selection (Berger & Perrichi, 1996; Carlin & Chib, 1995; Gregory & Loredo, 1992; Raftery, 1993), and is an unavoidable simplification of a problem for which there is not as yet a satisfactory solution. An implication of the assumption is that model selection is based solely on the Bayes factor. The decision to ignore model priors does not undermine use of the Bayesian approach. Rather, it represents a calculated approximation to the true Bayesian model selection criterion.[6]

Bayes factor $P(D|M_1)/P(D|M_2)$. The Bayes factor is defined as a ratio of two *marginal likelihoods*, which can best be understood by noting their relationship to likelihood functions. A likelihood function is the probability of the given data computed for a particular value of the parameter of a model, $P(D|\theta,M_i)$, and thus is a function of the parameter θ. In essence, the likelihood function is a goodness-of-fit measure of a model for a given parameter value. If the value of the parameter changes, the fit might also change. The marginal likelihood of a model, $P(D|M_i)$, is the probability of the data as a whole, independent of parameter values. It is an average of likelihoods under a prior distribution of the parameter. The marginal likelihood is expressed in the following integral form:

$$P(D|M_i) = \int P(D|\theta,M_i)P(\theta|M_i)d\theta \quad (i = 1,2). \quad (5)$$

where θ is a parameter vector under Model i, $P(D|\theta,M_i)$ is the likelihood function, and $P(\theta|M_i)$ is the prior density of θ for Model i.

Under the particular assumption of a uniform prior density function of the parameter [i.e., $P(\theta|M_i) = $ constant], the marginal likelihood is proportional to the area
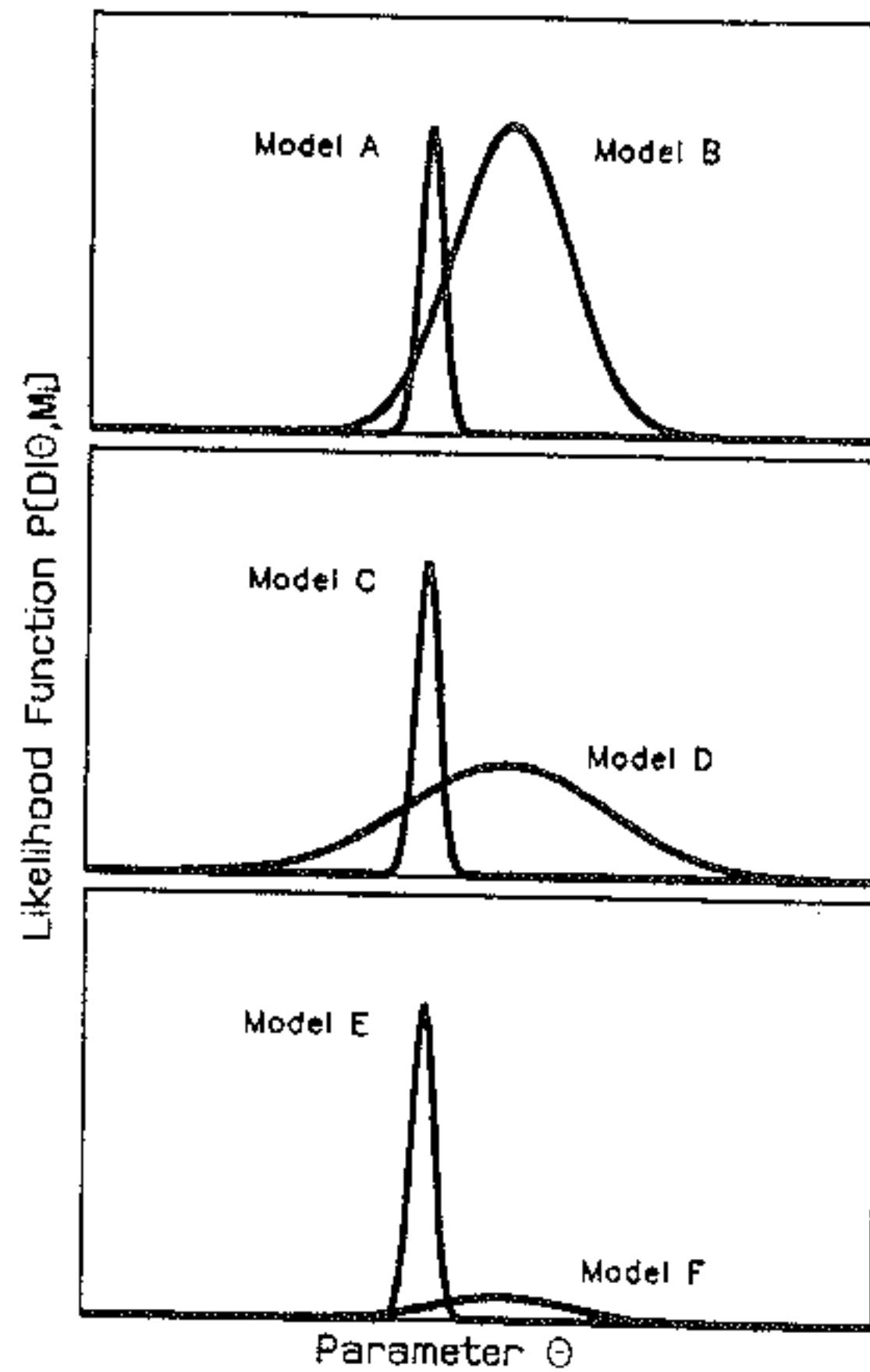

Figure 4. Maximum likelihood functions for three pairs of models as a function of the parameter θ. Models A, C, and E are more complex (more peaked functions) than are models B, D, and F. See text for details.

under the curve representing the likelihood function. To illustrate, likelihood functions for two models, A and B, are shown in the top panel of Figure 4. Note that the abscissa of the figure represents parameter value, not data value. Although the maximum likelihoods (i.e., the highest values) are the same for both models, the marginal likelihood of Model B is larger because the area under its function is larger. Note that the functions should be thought of as actual representations of the models. Their peakedness is determined by model complexity, to which we now turn.

Model Complexity: A Bayesian View

A model is selected in the Bayesian approach by maximizing the marginal likelihood. Under some simplifying assumptions, the marginal likelihood can be expressed in words as

$$\text{marginal likelihood} = \frac{\text{goodness of fit}}{\text{model complexity}} . \quad (6)$$

The numerator, goodness of fit, is the maximum likelihood of the data given the model, $P(D|\theta_o,M_i)$, where θ_o is the parameter value that maximizes the likelihood

function. The denominator term represents a model complexity measure that embodies all three dimensions of complexity: number of parameters, functional form, and extension of parameter space. A simplified mathematical function depicting their relationship is shown below. (A technical discussion of the mathematics is presented in Appendix A.)

$$\text{model complexity} = g(n_i, F_i(\theta_o), R_i) \quad (7a)$$

$$g(n_i, F_i(\theta_o), R_i) \approx \beta F_i(\theta_o)(R_{il} \cdots R_{in_i}) \quad (7b)$$

In the equation above, β is a positive scaling factor, n_i is the number of parameters of Model i, and $R_i = (R_{il}, \ldots, R_{in_i})$ is a vector representing the range of the parameter vector θ of Model i (e.g., R_{ik} equals 2 if the parameter θ_k is defined on $-1 \leq \theta_k \leq +1$). $F_i(\theta_o)$ is a "functional-form" factor whose value depends on how the parameters are combined mathematically in the model. Because this factor is evaluated at θ_o, which is determined by the data, the value of the functional-form factor may depend on the data.[7] Roughly speaking, the peakedness of the likelihood function near θ_o is positively correlated with the value of $F_i(\theta_o)$. For example, in Figure 3, $F_i(\theta_o)$ is higher for Model A than for Model B. As $F_i(\theta_o)$, R_{ik}, and n_i increase, so does model complexity. According to Equation 6, a more complex model yields a smaller marginal likelihood; consequently, it is less likely to be selected as the best fitting model. Model B would be chosen over Model A using the Bayesian method if other things were equal (i.e., same R_{ik} and n_i for both models).

The middle and lower panels of Figure 4 are examples of other model relationships that illustrate further how model complexity affects model selection. For the pairs of models, assume that they have the same number of parameters and the same extension of the parameter space, and further, that the prior density of the parameter follows a uniform distribution. In the lower panel, both selection methods (Bayesian and standard) would likely reach the same conclusion when evaluating Models E and F. Not only does Model F provide a much poorer fit than Model E, but also its marginal likelihood is smaller. Whether the maximum likelihood or the marginal likelihood were used to select the most appropriate model, Model E would be the clear winner.

In the middle panel, Model C would be chosen over Model D using the standard method because C's maximum likelihood is larger than D's. Bayesian model selection is less straightforward in this case. Even though D's fit is not as good, it would be chosen because its marginal likelihood is larger than C's. Maximization of the marginal likelihood favors a model that not only fits data well, but also does so over a wide range of parameters. Model D does so to a far greater extent than C. Comparisons such as this, in which goodness of fit and model complexity differ greatly between models, should benefit from application of the Bayesian approach.

In this last example, the Bayesian method might seem to yield counterintuitive results because the model that provides the best fit to the data is not chosen as the pre-

ferred model. This is because model complexity is just as important as model fit in influencing model selection. In the Bayesian approach, the ability of the model to fit the data reasonably well over a range of parameter values is what is important because it indicates that the model captures the structure in the data with a minimum of complexity (e.g., minimal reliance on functional form, number of parameters, etc.). Models in which parameters must be finely tuned to achieve a good fit (Models A and C) are valued less, precisely because the fit is achieved by relying on the complexity of the model. The contribution of complexity to model fit should not be underestimated. By finely tuning parameters, a complex model can fit a range of data patterns, not because it is the "true" model, but because it is the most flexible, easily adjusting to variations in the data. Taken to an extreme, one could imagine a highly complex model fitting almost any data set very well.

The preceding examples should make it clear that the difference in the pattern of data to be maximized by the likelihood function versus the marginal likelihood can have nontrivial implications for model selection and is at the heart of the Bayesian approach. Maximization of the marginal likelihood is accomplished by pitting maximization of the likelihood function against complexity minimization. A more complex model will be favored only if its goodness of fit is large enough to *justify* the additional complexity of the model. It is in this sense that Bayesian model selection is a quantitative implementation of Occam's razor.

Computation of the Bayes Factor

Although the simplified expression in Equation 6 is useful to illustrate basic ideas of Bayesian model selection, it is derived under some restrictive assumptions about the shape of the likelihood function. Goodness of fit and model complexity are not, in general, separable from each other, but instead, the two work together implicitly within a single measure, which is the marginal likelihood. Computation of the Bayes factor $P(D|M_1)/P(D|M_2)$ then requires evaluating the multiple integral in Equation 5. Readers who are not interested in the mathematical details of computing the Bayes factor should skip to the section, Application Example of Bayesian Model Selection.

Priors $P(\theta|M_i)$. To evaluate the integral, the prior density of the parameter, $P(\theta|M_i)$, must be specified. Two straightforward methods for obtaining the prior are described below. Others are discussed in Appendix B. Whatever method is used, it is very important to perform some form of sensitivity analysis to ensure that the resulting Bayes factor is reasonably stable over a range of values. For a more comprehensive treatment of this crucial topic, see Berger (1985) and Kass and Raftery (1995).

A simple-minded method for choosing priors is to use a noninformative prior, which by definition assumes no information about the parameter θ. For example, the uniform density function $\pi(\theta) = 1$ on $0 \leq \theta \leq 1$ is a noninformative prior. Although a noninformative prior would be an obvious choice when absolutely no information is

available, often the noninformative prior is an *improper* prior having infinite mass (i.e., $\int \pi(\theta) d\theta = \infty$). One example is the uniform density $\pi(\theta) = c$ on $0 \leq \theta < \infty$ where $c > 0$. Another reasonable choice of a noninformative prior for $0 \leq \theta < \infty$ that has the desirable scale-invariance property [i.e., $\pi(\theta) = \pi(\theta/\alpha)/\alpha$ for any $\alpha > 0$] is $\pi(\theta) = 1/\theta$; this is also an improper prior. Although the justification and interpretation of noninformative improper priors is disputed (see Berger, 1985, pp. 89–90), they are routinely used in Bayesian inference (e.g., Gelfand & Dey, 1994; H. Jeffreys, 1961).

Informative priors can be used when information about θ is available. One way to do this is to obtain an informative prior from an experimental setup. To illustrate, consider a signal detection experiment consisting of n conditions in which the signal-to-noise ratio is systematically varied from lowest in Condition 1 to highest in Condition n. Assume that the number of correct responses out of a total of s trials in Condition k ($k = 1, \ldots, n$) is binomially distributed with $Bin(s, \theta_k)$ where θ_k ($0 \leq \theta_k \leq 1$) is the binomial probability parameter. A reasonable way to obtain the prior $\pi(\theta_1, \ldots, \theta_n)$ is to assume that θ_ks are the order statistics (i.e., $\theta_i \leq \theta_j$ for any $i < j$), resulting from an independent random sample from the uniform distribution on $[0,1]$. For example, given a particular random sample of three ($n = 3$) observations, $(.53, .29, .87)$, from the uniform distribution, the desired order statistics are obtained as $\theta_1 = .29$, $\theta_2 = .53$, and $\theta_3 = .87$. For other similar cases, by being creative and reasonable, one can find priors that best capture the information available in the experimental design as well as in the data.

Numerical methods for computing the Bayes factor. A closed-form solution to the integral in Equation 5 is the preferred method for computing the Bayes factor. Most often, however, the integral must be evaluated numerically. A simple Monte Carlo integration method (Thisted, 1988, p. 302) can be used for integrals involving fewer than 10 parameters. More efficient methods such as Markov chain Monte Carlo (MCMC) methods should be used for integrals involving large numbers of parameters. For comprehensive treatments of MCMC methods and numerical integration methods in general, readers are directed to Smith (1991), Smith and Roberts (1993), and Thisted (1988, sec. 5.6).

The logic of the MCMC method is to obtain a sufficiently large sample of random observations generated from a probability density function in a *prespecified* manner, and then to use the sample to evaluate a desired integral. More specifically, the integral in Equation 5 is approximated by the following time average:

$$\frac{1}{T} \sum_{i=1}^{T} P(D|\theta^t, M_i) \qquad (8)$$

In this equation, $\{\theta^1, \theta^2, \ldots, \theta^T\}$ is a random sample of T vectors drawn from the prior density function $\pi(\theta)$ of Model i. Various versions of MCMC, such as the Gibbs sampler (Gelfand & Smith, 1990; Geman & Geman, 1984; Wakefield, Smith, Racine-Poon, & Gelfand, 1994) and the Metropolis-Hastings algorithm (Hastings, 1970), have been proposed. The Gibbs sampler will be described here, primarily because it is easy to apply once the set of fully conditional distributions of the parameters (the probability distributions of each parameter conditional on all remaining parameters) is identified, at least up to proportionality.

The iteration procedure for generating a random sample of any size by the Gibbs sampler is summarized in the following steps. A hypothetical example of the Gibbs sampler is shown in square brackets, [...], for a three-parameter case with the parameters $(\theta_1, \theta_2, \theta_3)$ defined between 0 and 1:

Initialization:

Determine T;

Pick an arbitrary starting vector $\theta^0 = (\theta_1^0, \ldots, \theta_n^0)$
$[\theta^0 = (.52, .19, .34)]$;

Define and set a temporary vector $\theta^{temp} := \theta^0$
$[\theta^{temp} = (.52, .19, .34)]$;

Let $\theta^{temp}\{-k\}$ denote a set of the current values in θ^{temp} without the kth element

$$[\text{e.g., for } k = 2, \theta^{temp}\{-2\} = (\theta_1^{temp} = 52,$$
$$\theta_3^{temp} = .34).]$$

Let SUM $= 0$;

Let $t = 1$.

Step 1: Take a random sample of θ_1^t from the conditional distribution $\pi(\theta_1 | \theta^{temp}\{-1\})$

$$[\theta_1^t = .91];$$

Replace θ_1^{t-1} of θ^{temp} with the new
θ_1^t $[\theta^{temp} = (.91, .19, .34)]$.

Step 2: Take a random sample of θ_2^t from the conditional distribution $\pi(\theta_2 | \theta^{temp}\{-2\})$

$$[\theta_2^t = .48];$$

Replace θ_2^{t-1} of θ^{temp} with the new θ_2^t
$[\theta^{temp} = (.91, .48, .34)]$.

...

Step n: Take a random sample of θ_n^t from the conditional distribution $\pi(\theta_n | \theta^{temp}\{-n\})$

$$[\theta_n^t = .20];$$

Replace θ_n^{t-1} of θ^{temp} with the new θ_n^t
$[\theta^{temp} = (.91, .48, .20)]$.

Evaluation:

Set $\theta^t = \theta^{temp}$ and evaluate $P(D|\theta^t, M_i)$ at θ^t
$[\theta^t = (.91, .48, .20)]$;

SUM $= $ SUM $+ P(D|\theta^t, M_i)$;

If $t < T$ then let $t = t+1$ and go to Step 1 or else go to End.

End:

SUM $= $ SUM / T.

The value of the variable SUM represents a numerical solution to the integral in Equation 5 for the specified number of iterations (T). The number of iterations nec-

essary to obtain an accurate approximation of the integral is not easy to determine beforehand. Raftery and Lewis (1991) have provided useful guidelines for estimating T, which depends on the type of problem, the model equation, and the specific priors used. Probably the most practical method is to examine the converging pattern of the sum by plotting it against the number of iterations to identify where the curve becomes stationary. A few runs usually give a good estimate for a given problem.

APPLICATION EXAMPLE OF BAYESIAN MODEL SELECTION

In this section, an implementation of the Bayesian model selection method is illustrated through numerical examples with simulated data.

Models of Information Integration

A long-standing question that continues to receive considerable attention in cognitive psychology is, How is information from different sources (e.g., sensory and contextual) integrated during perception? The answer to this question has broad implications for theory development because information integration is a common denominator among models: Models must specify when and how independent sources of information are combined during processing. Entire classes of models will have to be modified if they are shown to have the wrong functional architecture.

Interest in information integration spawned a variety of mathematical models of integration (see N. H. Anderson, 1981; Massaro & Friedman, 1990). Their number and relative simplicity provided an appropriate context in which to test the Bayesian selection method. In a typical experiment on information integration, two or more stimulus dimensions are factorially manipulated and presented to one or two modalities (e.g., auditory, visual). Participants identify stimuli along one dimension. Of interest is the influence of the orthogonal dimension on perception. For example, context effects in speech perception are investigated by measuring categorization of perceptually ambiguous phonemes embedded in words and nonwords.

For simplicity, only data from a two-factor, two-response-category experiment will be considered. Extension of the approach to more general cases is straightforward. Suppose that the first factor has n_1 levels and the second has n_2 levels, so that S_{ij} $(i = 1, \ldots, n_1; j = 1, \ldots, n_2)$ denotes the stimulus constructed with Level i of the first factor and Level j of the second. The stimuli are presented to participants in a random order and equally often over s independent trials. Participants categorize the stimuli as one of two possibilities, A or B. Let a random variable X_{ij} denote the number of Category A responses participants made when Stimulus S_{ij} was presented. Then, X_{ij} will follow a binomial probability distribution with parameters p_{ij} (probability of Category A response) and s (number of independent observations).

Three non-nested models of information integration were chosen for evaluation on the basis of their similarities in implementation. They were the fuzzy logic model of perception (FLMP) by Oden and Massaro (1978), the linear integration model (LIM) by N. H. Anderson (1981), and the theory of signal detection (TSD) by Green and Swets (1966). All three assume that the probability of classifying a stimulus as a member of Category A is a function of the extent to which the two feature dimensions of the stimulus (i and j) support the category response (Massaro & Friedman, 1990). Specifically, the response probability, p_{ij}, is assumed to be a function of two independent parameters, θ_i and λ_j, each of which represents the degree of support for a Category A response given the specific i and j feature dimensions of a stimulus. The three models, however, differ in how the two parameters are combined to produce p_{ij}.

According to the FLMP, p_{ij} takes the following nonlinear form (see Massaro & Friedman, 1990):

$$p_{ij, \text{FLMP}} = \frac{\theta_i \lambda_j}{\theta_i \lambda_j + (1 - \theta_i)(1 - \lambda_j)}. \quad (9a)$$

The LIM assumes a linear combination rule for p_{ij} as follows:

$$p_{ij, \text{LIM}} = \frac{\theta_i + \lambda_j}{2}. \quad (9b)$$

For the TSD, the response probability is given by

$$P_{ij, \text{TSD}} = \Phi\left[s_{ij} \sqrt{|\Phi^{-1}(v_i)^2 + v_{ij} \Phi^{-1}(\lambda_j)^2|}\right], \quad (9c)$$

where $\Phi()$ and $\Phi^{-1}()$ are the cumulative and inverse cumulative normal functions, respectively. The s_{ij} is the sign (± 1) of $(\Phi^{-1}(\theta_i) + \Phi^{-1}(\lambda_j))$ and v_{ij} is the sign (± 1) of $(\Phi^{-1}(\theta_i) \cdot \Phi^{-1}(\lambda_j))$.

Note that all three models possess the same number of parameters (two) as well as the same extension of the parameter space (i.e., $0 < \theta_i, \lambda_j < 1$), but differ in their functional form. Model selection using a standard method, such as AIC, would be decided solely on goodness of fit. Numerical examples in the following section demonstrate the improvement in model selection that is achieved with the Bayesian method.

Specification of Priors

To apply the Bayesian method, a probability density distribution for the two parameters of the model must be specified. We assumed that the distributions were the same for all three models, and further, that the parameters belonging to the first experimental variable were independent of those belonging to the second. In determining the characteristics of the distributions, a sensible choice is informative priors, because *prior* information about the parameters is available from the experimental setup that the present examples are intended to simulate. The method in which levels of the independent variable are created frequently ensures that ordinal information about the parameters is available, even before data are collected. For example, in speech perception experi-

ments, the levels of one of the variables (e.g., steps along a /ba/–/da/ phonetic continuum) are varied incrementally so that the probability of a Category A response at level i is greater than or equal to that at level i' if $i < i'$, and vice versa. The levels of the other variable can be manipulated in a similar fashion. Details on how this ordinal information was incorporated in the prior probabilities are provided in Appendix C.

Simulated Data and Model Fitting

Data simulating three response patterns in a 2×8 factorial design (i.e., $n_1 = 2$ and $n_2 = 8$) were created. Three distinct parameter sets, chosen to represent a range of response patterns in categorization experiments, were used to generate the simulated data.[8] Each set contained 10 parameter values (2 + 8). For each set, 16 ($N = 16$) binomial response probabilities (p_{ij}) were then computed using one of the three model equations in Equation 9. Each of the nine panels in Figure 5 shows a plot of the 16 probabilities for a given model and parameter set. Note that these probabilities represent ideal, error-free performance.

To simulate actual performance by human participants, samples were created using each parameter set in Figure 5 according to the binomial probability distribution as fol-

lows. For each of the 16 response probabilities (p_{ij}), a series of 20 ($s = 20$) independent binary outcomes (0 or 1) were generated in such a way that the probability of 1 was p_{ij}. Next, the number of 1s in the series was summed and divided by s to obtain an observed proportion for the particular combination of i and j. Finally, the above procedure was repeated for the remaining 15 values to obtain 16 observed proportions, which together constituted a single sample. One hundred samples were created for each of the nine panels in Figure 5.

Each of the three models was fitted to each simulated sample separately. This was done using the standard method and the Bayesian method to compare their abilities to recover (i.e., identify) the model that generated the original data. For the standard method, a nonlinear optimization routine (Marquardt compromise method; Marquardt, 1963) was used to find the least squares estimates that minimized the sum of squared errors between the simulated and predicted data. Because the number of parameters was the same across models, this procedure was equivalent to other methods, such as AIC. For the Bayesian method, the integral in Equation 5 was evaluated numerically using an extended series of simple Monte Carlo simulations.[9]
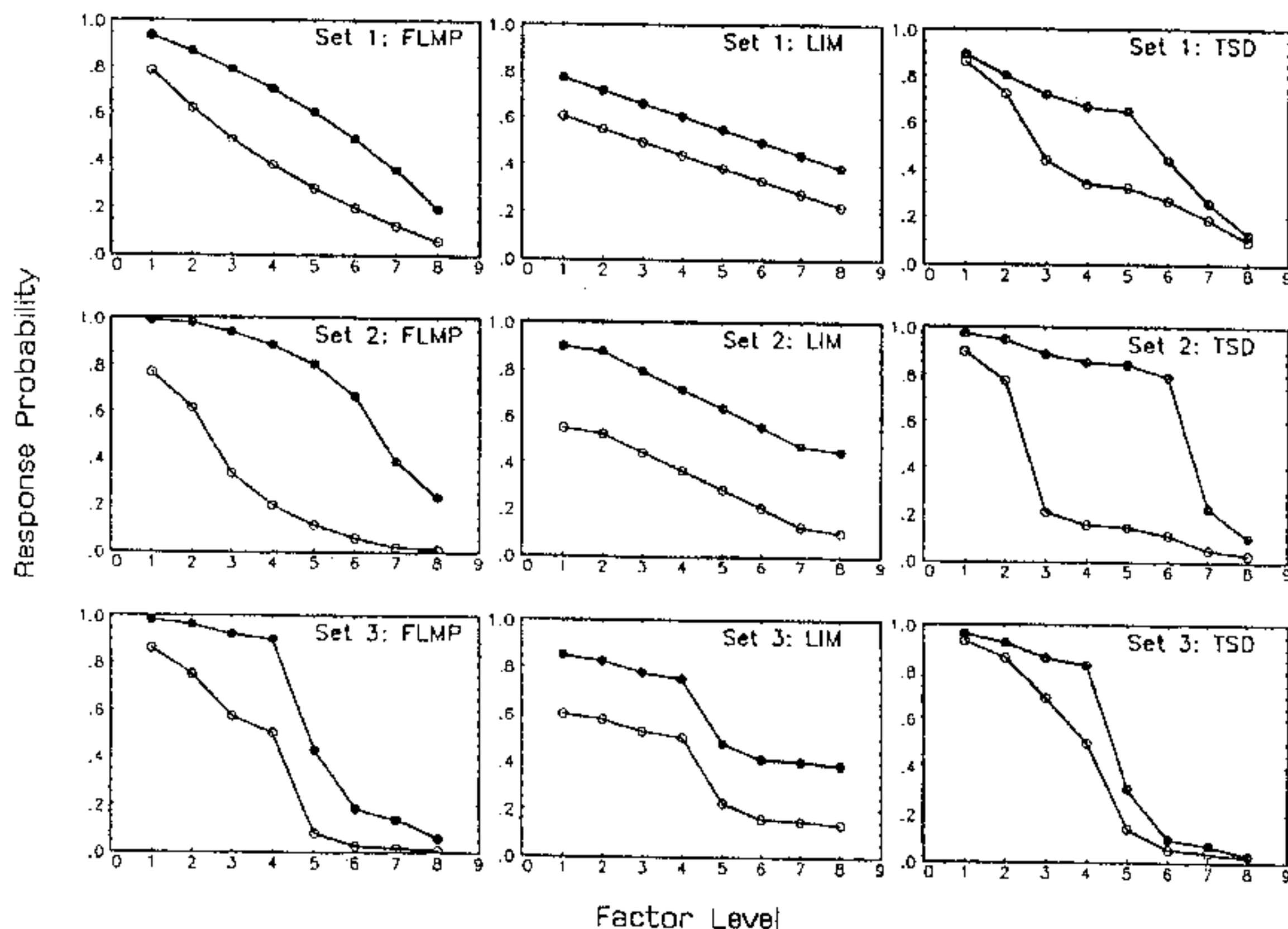
Figure 5. Binary response probabilities used to create the simulated data. The three panels in each row represent three different response patterns generated from the fuzzy logic model of perception (FLMP), the linear integration model (LIM), and the theory of signal detection (TSD) using Equations 9a–9c with a single set of 10 parameter values. Three sets of parameter values, corresponding to the three rows in the figure, were used.

Table 1
Summary Fits of the Simulated Data for the FLMP, LIM, and TSD Models
in the Standard Method

Data Generated From:		Set 1			Set 2			Set 3		
		FLMP	LIM	TSD	FLMP	LIM	TSD	FLMP	LIM	TSD
Model fitted:										
FLMP	RMSD	0.093	0.106	0.109	0.046	0.089	0.065	0.045	0.103	0.065
	% variance	95.63	85.44	93.71	99.30	94.84	98.82	99.40	93.09	98.85
	% win	62	14	20	65	47	20	54	33	26
LIM	RMSD	0.156	0.108	0.181	0.215	0.098	0.311	0.252	0.114	0.270
	% variance	88.44	85.05	82.72	86.96	94.00	75.81	83.71	91.75	81.96
	% win	0	36	0	0	31	0	0	17	0
TSD	RMSD	0.100	0.105	0.092	0.051	0.099	0.060	0.048	0.102	0.059
	% variance	94.94	85.79	95.52	99.18	93.61	99.02	99.31	93.23	99.06
	% win	38	50	80	35	22	80	46	50	74

Note—For each model fitted, the first row shows the arithmetic mean root mean squared deviation (RMSD) for the model, averaged across 100 samples, the second row shows the arithmetic mean percent variance accounted for, and the third row shows the percent of samples in which the model fits better than the other two models. FLMP, fuzzy logic model of perception; LIM, linear integration model; TSD, theory of signal detection.

Results

The results are summarized in Table 1 for the standard method. For the simulated data generated from the FLMP, model recovery rate—the percent of samples in which the model that generated the samples was correctly identified as the best fitting model—was 50%–60%, essentially uniform across the three parameter sets. When the original FLMP model was incorrectly identified (30%–50% of the time), it was always identified as the TSD model, never as the LIM model. Note that even when the FLMP was correctly identified, the winning margin was miniscule, rarely besting the TSD by more than a 10% difference in RMSD. The mirror image of this outcome was obtained for the simulated data generated from the TSD. The TSD was recovered approximately 78% of the time and was incorrectly identified as the FLMP around 22% of the time. It was never recovered as the LIM. These results suggest that the FLMP and TSD are close competitors. Each is flexible enough to mimic the other reasonably well. LIM is a lone loser that consistently failed to beat the data-fitting abilities of FLMP and TSD.

The above interpretation is further supported by the model fitting results for the LIM. What is most striking is the failure of LIM to beat out the competitors in fitting its own data. Model recovery rate for the LIM was only 28% across the three parameter sets.[10] The FLMP and TSD bested the LIM in two of the three parameter sets. Comparison of the RMSDs, however, reveals that the fits for all models are fairly similar. Competitors were never far behind the best fitting model. This is shown graphically in the three left panels of Figure 6, which are frequency distributions of RMSD ratios fitted to the LIM data by the LIM and FLMP models. The FLMP usually bested the LIM by less than a 10% difference in RMSD. This asymmetry in data fitting between FLMP and LIM is just what Cutting et al. (1992) found.

The model recovery results from application of the Bayesian method are summarized in Table 2. Unlike RMSD scores, which can be difficult to interpret because they are relative values, the marginal likelihoods in Table 2 are meaningfully interpretable. For each model, they represent the *actual* probabilities of observing the particular pattern of data for the given model. Because of the extremely large number of possible data patterns, these probabilities should be very small (10^{-20} to 10^{-16}). They should not be interpreted as being insignificant. Recall that the Bayes factor is a ratio of two marginal likelihoods. Bayes factors greater than 20 should be con-

Table 2
Summary Fits of the Simulated Data for the FLMP, LIM, and TSD Models in the Bayesian Method

Data Generated From:		Set 1			Set 2			Set 3		
		FLMP	LIM	TSD	FLMP	LIM	TSD	FLMP	LIM	TSD
Model fitted:										
FLMP	$P(D|M)$	6.7×10^{-17}	7.0×10^{-19}	9.9×10^{-18}	8.4×10^{-15}	1.0×10^{-17}	2.4×10^{-17}	3.4×10^{-15}	9.2×10^{-19}	2.1×10^{-15}
	% Win	86	0	25	98	3	6	91	11	38
LIM	$P(D|M)$	7.3×10^{-19}	7.7×10^{-17}	2.6×10^{-20}	4.9×10^{-21}	1.8×10^{-16}	4.3×10^{-28}	1.7×10^{-27}	1.1×10^{-17}	4.8×10^{-30}
	% Win	5	100	0	0	97	0	0	89	0
TSD	$P(D|M)$	7.8×10^{-19}	2.2×10^{-19}	2.8×10^{-17}	2.3×10^{-16}	1.9×10^{-19}	1.3×10^{-15}	2.0×10^{-16}	3.6×10^{-20}	2.9×10^{-15}
	% Win	9	0	75	2	0	94	9	0	62

Note—For each model fitted, the first row shows the geometric mean marginal likelihood, $P(D|M)$, averaged across 100 samples, and the second row shows the percent of samples in which the model fits better than the other two models. FLMP, fuzzy logic model of perception; LIM, linear integration method; TSD, theory of signal detection.

sidered strong evidence for the favored model. Values greater than 150 should be considered very strong evidence (Kass & Raftery, 1995).

Overall, the data show a marked improvement in data recovery over the standard method. For the FLMP and TSD models, recovery rate between the Bayesian and standard methods improved in four of the six cells (86 vs. 62, 98 vs. 65, 94 vs. 80, 91 vs. 54). There was one tie (75 vs. 80) and one case in which it did worse (62 vs. 74). In addition, the winning model won more decisively with the Bayesian method. For the four cases in which there was an improvement, the marginal likelihood was 17 to 86 times that of the losing model.

Even more impressive are the recovery results for the LIM model. The Bayesian method recovered the LIM model from 89% to 100% of the time. This outcome is in sharp contrast to the standard method, which produced recovery rates barely one third as good (17%–36%). When the LIM did win, it was an overwhelming victor a majority of the time. This is shown graphically in the right three panels of Figure 6, which are frequency distributions of the log ratios of marginal likelihoods (i.e., Bayes factors) fitted to the LIM data by the LIM and FLMP models. As can be seen across all three para-

meter sets, the degree of support for the LIM is at quite a comfortable level, the Bayes factors ranging from about 10 to 1,000.

The results in Tables 1 and 2 together clearly demonstrate the superiority of the Bayesian method in model selection. By penalizing models for the complexity of their functional form, the Bayesian method improves model recovery rate significantly over the standard method. The functional forms of the three models are indeed markedly different from one another. The LIM assumes a linear form with respect to the parameters, whereas both the FLMP and TSD assume sophisticated nonlinear forms. Unless this component of model complexity is incorporated into the selection process, a complex model will be mistakenly declared as the best fitting model for a given parameter set more often than a simple model, as shown in Table 1. Note that this will occur not because the complex model is more plausible, but because the extra flexibility built into the complex model enables it to capture more of the random, idiosyncratic variation in the data than the simpler model, thus improving goodness of fit.

However, the Bayesian method failed to recover the correct model some of the time. This occurred primarily when the FLMP was fitted to TSD data using parameter
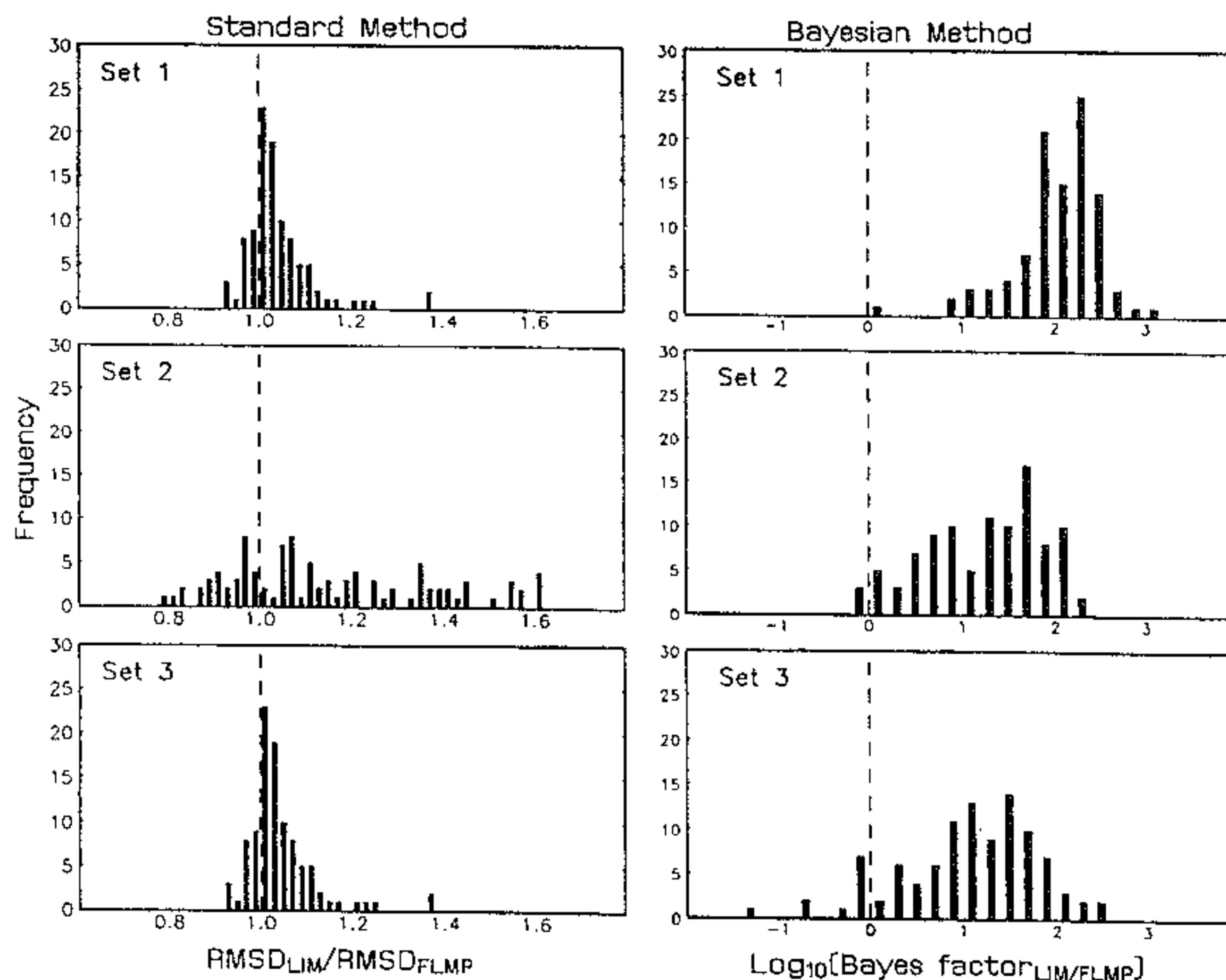

Figure 6. Frequency distributions of root mean squared deviation (RMSD) ratios for the standard method (left three panels) and of the Bayes factor, $P(D|LIM)/P(D|FLMP)$, for the Bayesian method (right three panels). The LIM data were fitted to both the linear integration model (LIM) and the fuzzy logic model of perception (FLMP). Note that the smaller the RMSD of a model, the better the model, whereas the larger the marginal likelihood of a model, the better the model.

sets 1 and 3 (top right corner of each 3×3 matrix). On trials in which a model recovery error was made, the FLMP beat the TSD by very little, with the Bayes factor rarely exceeding 5. Note also how localized model recovery errors were across the three data sets. Fewer errors were found with parameter set 2 or in the opposite corner of each matrix, where they might be expected, when the TSD model was fitted to FLMP data (2%, 9% error rates).

Because the errors occurred with specific parameter sets and the Bayes factor was usually small in these cases, we attribute the high errors to FLMP providing exceptionally good fits to the regions in data space occupied by parameter sets 1 and 3 (Figure 5). That is, FLMP's goodness of fit was so good in this region that it partially offset the complexity penalty that the model is normally assessed, thereby besting the TSD by a hair. This superior data-fitting ability does not extend to the data space occupied by parameter set 2, enabling the correct model to be recovered most of the time. Systematic studies are planned to test the viability of this speculative account.

DISCUSSION

In this final section, we discuss the pros and cons of using the Bayesian and standard methods. An advantage of the standard method is that it is easy to use. To choose among models, one need obtain only two easy-to-compute indices for each model: a model fit index, such as RMSD or maximum likelihood, and a complexity index, a count of the number of free parameters. These are combined to yield an overall measure of model fit using a simple algebraic formula, the exact form of which depends on the particular criterion used (e.g., AIC, BIC, RMSEA). The model that minimizes the overall fit is chosen.

The present findings demonstrate that despite the simplicity and convenience of the standard method, critical dimensions of complexity, such as functional form, are ignored when it is used. Influences of functional form can be ignored only when two rather restricting assumptions are met: The sample size is sufficiently large (i.e., theoretically infinite) and the models being compared are nested. Indeed, it is these assumptions from which many of the selection criteria for standard models (e.g., AIC and BIC) were derived and their desirable properties were shown (e.g., asymptotic optimality; Schwarz, 1978). Strictly speaking, whenever these assumptions are violated, use of the standard method is not justified.

This does not mean that the standard method should be abandoned. It is useful for discriminating among nested models that have the same functional form, thereby canceling out effects due to this dimension of complexity. This might often be the case in psychometric data analysis in which models being compared tend to be nested (e.g., regression models, latent variable models). However, as we have seen, use of the standard method can lead to erroneous decisions, especially for data of small or moderate sample sizes.

In contrast to the standard method, the Bayesian method does not require the frequently unrealistic assumption of a large sample size. It is safe to apply the method to comparisons among models with data of *any* sample size. Nor is the Bayesian method confined to evaluating nested models. This is because it appropriately adjusts the model complexity measure for differences in functional form. This latter feature makes the method particularly well suited for comparing mathematical models of cognition, of which many are non-nested.

Areas of inquiry in which the Bayesian method could be fruitfully applied include categorization models, such as decision-bound models (Ashby, 1992; Ashby & Gott, 1988; Ashby & Townsend, 1986), prototype models (Reed, 1972), and context models (Medin & Schaffer, 1978; Nosofsky, 1986); memory models, such as search-of-associative-memory (SAM) models (Gillund & Shiffrin, 1984), multiple trace models (Hintzman, 1986, 1988), and holographic memory models (Metcalfe-Eich, 1982; Murdock, 1982); and causal inference models, such as the linear model (Schustack & Sternberg, 1981), the weighted ΔP model (Allan, 1980, 1993), and the Bayesian inference model (J. R. Anderson, 1990; J. R. Anderson & Sheu, 1995). Like the standard method, the Bayesian method is not limited to testing models of cognition and could be used to test quantitative models in any area of psychology.

The Bayesian method, however, has its drawbacks. One is that parameter priors are required to compute the marginal likelihoods. The standard method, on the other hand, ignores information about priors, even when available. Although this is a rational strategy in an exploratory phase of research (e.g., when no or little reliable information about parameter priors can be obtained), it may not be justifiable when building mathematical models, which are usually developed at advanced stages of a research program. In this case, it makes sense to capitalize on the available information in the data or in the experimental setup to compute priors. We have suggested several ways of determining priors by utilizing such information (Appendix B).

Another drawback of the Bayesian method is its heavy demand for computational resources. Unlike the simple formulas prescribed in the standard method, the Bayesian method usually requires numerical computations of integral form to evaluate the posterior probability distribution. This task has been simplified greatly by recent advances in the area of Bayesian computation (e.g., the Gibbs sampler, Geman & Geman, 1984). Nevertheless, it may still be viewed as an obstacle by those who are not familiar with numerical computation methods.[11]

Despite the apparent differences between the Bayesian and standard methods, there are cases in which both will perform similarly for large sample sizes. For example, the BIC can be viewed as a crude approximation of the Bayes factor that ignores priors (see Appendix D).

Given the advantages and disadvantages of the Bayesian method, it might be best to apply the method to comparisons among well-developed mathematical models.

Without sufficiently detailed models, the construction of reliable priors may not be possible, or if accomplished, might increase the chances of unwarranted decisions. In addition, if a myriad of mathematical models have to be evaluated, a nontrivial amount of computing time will be needed. We recommend that the simple, easy-to-use standard method be applied first to eliminate grossly unfit models in order to narrow the field to a few highly competitive ones. Then apply the Bayesian method to further differentiate among the survivors. This technique balances computational costs with model selection costs. There is no reason for the Bayesian method to be the single, universal model selection tool.

In conclusion, we have introduced a Bayesian model selection method that embodies the principle of Occam's razor. We have tried to present a balanced view of the method, pointing out its desirable features (i.e., sensitivity to functional form and applicability to comparison among non-nested models and data of any sample size) as well as some current drawbacks (i.e., computational cost and specification of parameter priors). When applied with discretion, it can be a useful tool for distinguishing sensible models from superfluous ones, thus contributing to theoretical advancement in the discipline.

One final note. Quantitative methods must not be used as the final decision makers when comparing models. Bayesian or non-Bayesian, quantitative model selection methods cannot replace other selection criteria (e.g., explanatory power, plausibility, internal consistency). Mathematical modeling will best serve the discipline when used in conjunction with these criteria. In this regard, we concur with the word of caution offered by Browne and Cudeck (1992, p. 253), "Fit indices should not be regarded as a measure of usefulness of a model. ... Consequently, 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 judgment."

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox & F. Caski (Eds.), *Second International Symposium on Information Theory* (p. 267). Budapest: Akademiai Kiado.

AKAIKE, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute*, 50, 277-290.

ALLAN, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, 15, 147-149.

ALLAN, L. G. (1993). Human contingency judgments: Rule based or associativity? *Psychological Bulletin*, 114, 435-448.

ANDERSON, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

ANDERSON, J. R., & SHEU, C.-F. (1995). Causal inferences as perceptual judgments. *Memory & Cognition*, 23, 510-524.

ANDERSON, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

ASHBY, F. G. (1992). Multidimensional models of categorization. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 449-483). Hillsdale, NJ: Erlbaum.

ASHBY, F. G., & GOTT, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 14, 33-53.

ASHBY, F. G., & TOWNSEND, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, 93, 154-179.

BALAKRISHNAN, N., & COHEN, A. C. (1991). *Order statistics and inference: Estimation methods*. New York: Academic Press.

BAMBER, D., & VAN SANTEN, J. P. H. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology*, 29, 443-473.

BERGER, J. O. (1985). Statistical decision theory and Bayesian analysis (2nd ed.). New York: Springer-Verlag.

BERGER, J. O., & PERRICHI, L. R. (1996). The intrinsic Bayes factor for model selection. *Journal of the American Statistical Association*, 91, 109-122.

BICKEL, P. J., & DOKSUM, K. A. (1977). *Mathematical statistics*. Oakland, CA: Holden-Day.

BOZDOGAN, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.

BRETTHORST, G. L. (1989). Bayesian model selection: Examples relevant to NMR. In J. Skilling (Ed.), *Maximum entropy and Bayesian methods* (pp. 377-388). Amsterdam: Kluwer.

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

BUSEMEYER, J. R., & TOWNSEND, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100, 432-459.

CARLIN, B. P., & CHIB, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 3, 473-484.

CHAITIN, G. J. (1966). On the length of programs for computing binary sequences. *Journal of the Association for Computing Machinery*, 13, 547-569.

COLLYER, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception & Psychophysics*, 38, 476-481.

COVER, T. M., & THOMAS, J. A. (1991). *Elements of information theory*. New York: Wiley.

CRYER, J. D. (1986). *Time series analysis*. Boston: PWS-Kent.

CUDECK, R., & HENLY, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, 109, 512-519.

CUTTING, J. E., BRUNO, N., BRADY, N. P., & MOORE, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, 121, 364-381.

DE BRUIJN, N. G. (1958). *Asymptotic methods in analysis*. Amsterdam: North-Holland.

GELFAND, A. E., & DEY, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B*, 56, 501-514.

GELFAND, A. E., & SMITH, A. E. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

GEMAN, S., & GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 6, 721-741.

GILLUND, G., & SHIFFRIN, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91, 1-67.

GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

GREGORY, P. C., & LOREDO, T. J. (1992). A new method for the detection of a periodic signal of unknown shape and period. *Astrophysical Journal*, 398, 146-168.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chain and their applications. *Biometrika*, 57, 97-109.

HINTZMAN, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, 93, 411-428.

HINTZMAN, D. L. (1988). Judgments of frequency and recognition in a multiple-trace memory model. *Psychological Review*, 84, 260-278.

JACOBS, A. M., & GRAINGER, J. (1994). Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception & Performance*, 20, 1311-1334.

JAYNES, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106, 620-630; 108, 171-190.

JEFFREYS, H. (1961). *Theory of probability* (3rd ed.). New York: Oxford University Press.

JEFFREYS, W. H., & BERGER, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, 80, 64-72.

KAPUR, J. N., & KESAVAN, H. K. (1992). *Entropy optimization principles with applications*. New York: Academic Press.

KASS, R. E., & RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

KOLMOGOROV, A. N. (1968). Logical basis for information theory and probability theory. *IEEE Transactions on Information Theory*, 14, 662-664.

KRUSCHKE, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.

KULLBACK, S., & LEIBLER, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79-86.

LE, N. D., & RAFTERY, A. E. (1996). Robust Bayesian model selection for autoregressive processes with additive outliers. *Journal of the American Statistical Association*, 91, 123-131.

LI, M., & VITANYI, P. (1993). *An introduction to Kolmogorov complexity and its applications*. New York: Springer-Verlag.

MACKAY, D. J. C. (1992). *Bayesian methods for adaptive models*. Unpublished doctoral dissertation, California Institute of Technology, Pasadena.

MADDOX, W. T., & ASHBY, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, 53, 49-70.

MARQUARDT, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics*, 11, 431-441.

MASSARO, D. W., & COHEN, M. M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, 122, 115-124.

MASSARO, D. W., & FRIEDMAN, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, 97, 225-252.

MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207-238.

METCALFE-EICH, J. (1982). A complete holographic associative recall model. *Psychological Review*, 89, 627-661.

MURDOCK, B. B., JR. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89, 609-626.

NOSOFSKY, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115, 39-57.

ODEN, G. C., & MASSARO, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, 85, 172-191.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B*, 57, 99-138.

RAFTERY, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 163-180). Thousand Oaks, CA: Sage.

RAFTERY, A. E. (1994). *Approximate Bayes factors and accounting for model uncertainty in generalized linear models* (Tech. Rep. 255). Seattle: University of Washington, Department of Statistics.

RAFTERY, A. E., & LEWIS, S. (1991). How many iterations in the Gibbs sampler? *Bayesian Statistics*, 4, 763-773.

REED, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3, 382-407.

RISSANEN, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, 14, 1080-1100.

RISSANEN, J. (1990). Complexity of models. In W. H. Zurek (Ed.), *Complexity, entropy, and the physics of information* (pp. 117-125). Reading, MA: Addison-Wesley.

ROBERTS, F. S. (1979). *Measurement theory*. Reading, MA: Addison-Wesley.

SCHUSTACK, M. W., & STERNBERG, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General*, 110, 101-120.

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

SMITH, A. F. M. (1991). Bayesian computational methods. *Philosophical Transactions of the Royal Society of London: Series A*, 337, 369-386.

SMITH, A. F. M., & ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 55, 3-23.

SOLOMONOFF, R. J. (1964). A formal theory of inductive inference. *Information Control*, 7, 1-22, 224-254.

STEIGER, J. H. (1990). Structural model evaulation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173-180.

STEIGER, J. H., & LIND, J. C. (1980, November). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City.

TAKANE, Y., & SHIBAYAMA, T. (1992). Structure in stimulus identification data. In F. G. Ashby (Ed.), *Multidimensional models of perception and cognition* (pp. 335-362). Hillsdale, NJ: Erlbaum.

THISTED, R. A. (1988). *Elements of statistical computing: Numerical computation*. New York: Chapman & Hall.

TIERNEY, L., & KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82-86.

TOWNSEND, J. T. (1975). The mind-body equation revisited. In C. Cheng (Ed.), *Philosophical aspects of the mind-body problem* (pp. 200-218). Honolulu: Honolulu University Press.

TRIBUS, M. (1969). *The principle of maximum entropy*. Elmsford, NY: Pergamon.

VAN ZANDT, T., & RATCLIFF, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin & Review*, 2, 20-54.

WAKEFIELD, J. C., SMITH, A. F. M., RACINE-POON, A., & GELFAND, A. E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics*, 43, 201-221.

NOTES

1. Note that the term *complexity* only partly corresponds to its usage in everyday language.

2. It is important to distinguish between data size (N) in Equation 3 and sample size (s: the number of observations per stimulus) in Equation 2. To illustrate, consider a categorization experiment in which each subject was asked to categorize 16 different stimuli generated from a 2×8 factorial manipulation of two stimulus dimensions. A total of 20 subjects were tested, with each responding to each stimulus only once. In this case, the data size is $N = 16$, the sample size is $s = 20$, and thus the total number of observations is $sN = 320$.

3. Two models are non-nested if the functional form of one model cannot be reduced to a special case of the other. For example, a polynomial model of the form $y = \theta_0 + \theta_1 x + \theta_2 x^2$ and an exponential model of the form $y = \theta_0 + \theta_1 \exp(-\theta_2 x)$ are non-nested. On the other hand, the polynomial model and another model of the form $y = \theta_0 + \theta_1 x$ are nested because the latter is obtained from the former as its special case by setting $\theta_2 = 0$.

4. Bozdogan (1987) extended the AIC to reflect functional form using an asymptotically consistent criterion called the consistent AIC with Fisher information (CAICF). Note that the CAICF and AIC were developed for comparison among nested models, in which effects due to functional form are at best minimal.

5. Bamber and van Santen (1985) argued that even when assessing the testability (whether there exist potential outcomes that are inconsistent with the model) and identifiability (whether the model's parameters are uniquely determined for a specific outcome) of a model, the same three dimensions of complexity must be considered.

6. From a computational complexity point of view, meaningful model priors *can* be defined under the assumption that structure in the world evolved out of randomness (Chaitin, 1966; Kolmogorov, 1968; Solomonoff, 1964; for reviews of the topic, see Cover & Thomas, 1991, and Li & Vitanyi, 1993). The idea is that if objects in the world were created randomly, the probability of observing a simple object should be much higher than that of observing a complex object. The utility of this approach for deriving meaningful priors in cognitive psychology has yet to be explored satisfactorily.

7. Although the functional form of a model remains fixed, model complexity due to functional form may not, and it can be dependent on the data itself. One situation in which complexity does remain fixed is testing normal probability models (see Appendix A).

8. The 10 parameter values assumed for each parameter set were as follows: for parameter set 1, $\theta = (0.66, 0.33)$, $\lambda = (0.88, 0.77, 0.66, 0.55, 0.44, 0.33, 0.22, 0.11)$; for parameter set 2, $\theta = (0.85, 0.15)$, $\lambda = (0.95, 0.90, 0.74, 0.58, 0.42, 0.26, 0.10, 0.05)$; and for parameter set 3, $\theta = (0.75, 0.25)$, $\lambda = (0.95, 0.90, 0.80, 0.75, 0.20, 0.07, 0.05, 0.02)$.

9. A typical run of the Monte Carlo simulation required about 50,000 iterations to obtain a reasonable level of accuracy for the marginal likelihood, though in some cases up to 500,000 iterations were needed.

10. This imperfect model recovery performance for the LIM persists for larger sample sizes (s). For example, simulations not reported here showed that model recovery rate for the LIM was still about 50% for $s = 100$ and 80% for $s = 1,000$ fairly uniformly across the three parameter sets.

11. C source code, which implements the Monte Carlo methods used in the application examples, is available from the authors.

12. This condition is known as the independence axiom in conjoint measurement theory and is widely assumed in models of choice behavior under uncertainty (Roberts, 1979).

APPENDIX A

This appendix describes details of the Laplace approximation of the marginal likelihood in Equation 5. Assume that the likelihood function $P(D|\theta,M_i)$ is highly peaked near its maximum θ_o (e.g., normal density) and further, that the prior density $P(\theta|M_i)$ is essentially flat in the neighborhood of θ_o. Under these assumptions, the following approximation to the marginal likelihood for large sample size (De Bruijn, 1958; Tierney & Kadane, 1986) is obtained:

$$P(D|M_i) \approx \frac{P(D|\theta_o,M_i)}{|H_i(\theta_o)|^{1/2} / [(2\pi)^{n_i/2} P(\theta_o|M_i)]}, \quad (i = 1,2). \quad (A1)$$

In this equation, n_i is the number of parameters for the Model i, and $H_i(\theta_o)$ is the negative Hessian matrix of the log likelihood, $H_i(\theta) = -\nabla^2 \log[P(D|\theta,M_i)P(\theta|M_i)]$ ($= -\nabla^2 \log[P(\theta|D,M_i)]$ from the Bayes rule), evaluated at θ_o. $P(D|\theta_o,M_i)$ is the maximum likelihood of the data given the model. As can be seen in the above equation, maximization of the marginal likelihood involves not only maximization of the numerator (the likelihood function) but also, importantly, minimization of the denominator term of the equation. Intuitively, the denominator can be understood as a measure of model complexity:

$$\text{Model Complexity} = \frac{|H_i(\theta_o)|^{1/2}}{(2\pi)^{n_i/2} P(\theta_o|M_i)}. \quad (A2)$$

This model complexity measure reflects the functional form of the model through $|H_i(\theta_o)|$. $H_i(\theta_o)$ will take on a different value depending on how the parameters are combined in the model. The functional-form factor, $F_i(\theta_o)$, in Equation 7a is defined as $F_i(\theta_o) = |H_i(\theta_o)|^{1/2}$. To give a concrete example for $F_i(\theta_o)$, consider a family of multivariate normal probability models $N(\mu, \Sigma)$ in which each model is defined by a distinct covariance matrix Σ and has the mean vector as a parameter (i.e., $\theta = \mu$). Assuming a uniform density for $P(\theta|M_i)$, it can be shown that $H_i(\theta_o)$ is equal to the inverse of the covariance matrix and is independent of the data D as well as the parameter μ, so $F_i(\theta_o) = |\Sigma|^{-1/2}$. Note that $|\Sigma|^{-1/2}$ becomes $1/\sigma$ for $n_i = 1$ of univariate normal probability models. Thus, the smaller the variance of a normal probability model, the larger the $F_i(\theta_o)$ and thus the more complex the model. Put another way, a model that is finely tuned to fit observed data (i.e., highly peaked around θ_o) will have a higher complexity measure than a model that can fit the data for a wide range of parameter values.

The number of parameters in a model is also represented in $H_i(\theta_o)$. For a sufficiently large sample size s, we obtain an exponential function of parameters (n_i) (see Raftery, 1993),

$$|H_i(\theta_o)|^{1/2} \approx (sN)^{n_i/2}, \quad (A3)$$

where N is the data size. The more the number of parameters in a model, the larger the complexity measure.

The complexity measure is also sensitive to the extension of the parameter space, which is reflected through $P(\theta_o|M_i)$. To see this, first note that the prior $P(\theta|M_i)$ denotes a regular probability density, the integral of which over the extension of the parameter space must be equal to 1. In general, $P(\theta_o|M_i)$, which refers to the value of the probability density evaluated at θ_o, will be smaller for a large extension of the parameter space than for a small extension. Because $P(\theta_o|M_i)$ is inversely related to the complexity measure, a model defined with a large extension of the parameter space is more complex than one with a small extension. As a concrete example, suppose that each element of the parameter vector θ for Model i is defined on a real line of finite size R_{ik} (e.g., $0 \leq \theta_k \leq R_{ik}$, $k = 1, \ldots, n_i$, for some constant $R_{ik} < \infty$), and further, that the prior density $P(\theta|M_i)$ for θ is an independent, uniform distribution. We then get

$$P(\theta_o|M_i) = 1 / \prod_{k=1}^{n_i} R_{ik}. \quad (A4)$$

If other things are equal, the wider the range R_{ik}, the more complex the model.

Finally, combing the result in Equation A4 with the definition $F_i(\theta_o) = |H_i(\theta_o)|^{1/2}$ in Equation A2 yields the approximation in Equation 7b.

APPENDIX B

This appendix describes two additional methods for obtaining informative priors for $P(\theta|M_i)$.

Maximum entropy priors. Informative priors can be used when information about θ is available from past data. If information is given in terms of moment constraints, such as mean and variance, then maximum entropy priors are a natural choice. The principle of maximum entropy (Jaynes, 1957) is an inference method for estimating a probability distribution used when limited information is available in the form of the moments of that distribution. For example, suppose that for a parameter of the range $0 \leq \theta < \infty$ the mean of the probability distribution is known and is equal to some positive value, λ. The maximum entropy prior is the exponential density $\pi(\theta) = e^{(-\theta/\lambda)}/\lambda$. Jaynes provided theoretical justification for such a maximum entropy prior as that which utilizes all information contained in the moment constraints but which is maximally noncommittal with respect to information not available. For more comprehensive treatments of maximum entropy inference, including solution methods, readers are directed to Berger (1985, sec. 3.4), Cover and Thomas (1991), and Tribus (1969). Kapur and Kesavan (1992, p. 359) have included a comprehensive list of maximum entropy priors given various moment constraints.

Partial and intrinsic priors. Another appealing method of obtaining informative priors is the *partial* Bayes factor method,

which uses existing data (Berger & Perrichi, 1996; O'Hagan, 1995). Half of the data are used as a training sample to obtain an informative posterior distribution of the parameters. This distribution is then used to compute the Bayes factor with the other half of the data. This method is particularly useful when noninformative priors are not well justified or difficult to obtain. One obvious problem is that the results depend on the choice of the training sample as well as the sample size. Berger and Perrichi have proposed a new criterion, the *intrinsic* Bayes factor method, to resolve the problem. It takes into account all possible training samples when computing the posterior distribution. That is, first an informative posterior distribution from each training sample is estimated. These values are then averaged (arithmetically or geometrically) to obtain the posterior distribution. This method is a sensible way to obtain "objective" priors, though it may be computationally intensive when numerical solutions are required.

APPENDIX C

This appendix describes a method for obtaining the informative prior used in the application example described in the present paper.

Assume that the prior density is the same for all three models (FLMP, LIM, and TSD), and further, that the parameters belonging to the first factor are independent of those belonging to the second factor. These assumptions lead to

$$\pi(\theta_1,\ldots,\theta_{n_1},\lambda_1,\ldots,\lambda_{n_2}) = \pi_1(\theta_1,\ldots,\theta_{n_1})\pi_2(\lambda_1,\ldots,\lambda_{n_2}). \quad \text{(C1)}$$

The two probability densities, π_1 and π_2, are determined utilizing ordinal information about the parameters, which is available from the experimental setup. That is, if the response probability assumed by each of the three models in Equations 9a, 9b, and 9c is a monotonically and independently increasing function of both parameters θ_i and λ_j, the following orderings among parameters will hold:[12]

$$\theta_i \geq \theta_{i'}, \text{ for } i < i'; \lambda_j \geq \lambda_{j'} \text{ for } j < j'. \quad \text{(C2)}$$

This ordinal information about the parameters was incorporated into priors using order statistics. They were constructed numerically as follows. To generate random samples of $\theta_1,\ldots,\theta_{n_1}$ that follow $\pi_1(\theta_1,\ldots,\theta_{n_1})$, first draw an independent ran-

dom sample of $V_1, V_2,\ldots, V_{n_1}$ from the uniform density on $[0,1]$. Then let $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_{n_1}$ be the n_1 order statistics resulting from the sample $\{V_i, i = 1,\ldots, n_1\}$. For instance, for a particular sample of $(V_1, V_2, V_3) = (0.43, 0.27, 0.81)$ where $n_1 = 3$, the order-statistic sample for the three parameters would be $(\theta_1, \theta_2, \theta_3) = (0.81, 0.43, 0.27)$. Similarly, $\pi_2(\lambda_1,\ldots, \lambda_{n_2})$ was constructed numerically from the n_2 order statistics resulting from an independent random sample $\{V_i, i = 1,\ldots, n_2\}$ drawn from the same uniform density. Alternatively and equivalently acceptable, the Gibbs sampler can be used here to generate V_is, since closed-form expressions for the multivariate priors, π_1 and π_2, are known (see, e.g., Balakrishnan & Cohen, 1991).

APPENDIX D

Recall from Appendix A that for a sufficiently large sample size, the Laplace method in Equation A1, along with the asymptotic form in Equation A3 for the negative Hessian matrix, provides an adequate approximation for the marginal likelihood $P(D|M_i)$. Combining these two, the following approximation for twice the logarithm of the Bayes factor is obtained (Raftery, 1993):

$$2 \ln \left[\frac{P(D|M_1)}{(P(D|M_2)} \right] \approx \text{BIC}(M_2) - \text{BIC}(M_1) + 2 \ln \left[\frac{P(\theta_o|M_1)}{P(\theta_o|M_2)} \right]. \quad \text{(D1)}$$

Therefore, model selection based on the BIC can be viewed as a simple but crude approximation to those based on the Bayes factor. An important theoretical difference is that BIC ignores priors; that is, $P(\theta_o|M_1)/P(\theta_o|M_2) = 1$.

As for the AIC, model selection using it is asymptotically equivalent to that based on the Bayes factor if the amount of information about θ in the Kullback-Leibler sense (Kullback & Leibler, 1951), supplied by the prior density $P(\theta|M_i)$, is comparable to that supplied by the data in the form of the likelihood function $P(D|\theta_o, M_i)$ (Kass & Raftery, 1995).