Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Mathematical Psychology 50 (2006) 193–202
Journal of Mathematical Psychology
www.elsevier.com/locate/jmp

Model selection for the rate problem: A comparison of significance testing, Bayesian, and minimum description length statistical inference

Michael D. Lee[a,*], Kenneth J. Pope[b]

[a]Department of Psychology, University of Adelaide, SA 5005, Australia
[b]School of Informatics and Engineering, Flinders University of South Australia, Australia

Received 13 December 2004; received in revised form 10 November 2005
Available online 27 January 2006

Abstract

One particularly useful but under-explored area for applying model selection in psychology is in basic data analysis. Many problems of deciding whether data have "significant differences" can profitably be viewed as model selection problems. We consider significance testing, Bayesian and minimum description length (MDL) model selection on a common data analysis problem known as the rate problem. In the rate problem, the question is whether or not the underlying rate of some phenomenon is the same in two populations, based on finite samples from each population that count the number of "successes" from the total number of observations. We develop optimal Bayesian and MDL statistical criteria for making this decision, and compare their performance to the standard significance testing approach. A series of Monte-Carlo evaluations, using different realistic assumptions about the availability of data in rate problems, show that the Bayesian and MDL criteria perform extremely similarly, and perform at least as well as the significance testing approach.

Keywords: Bayesian statistics; Minimum Description Length; Significance testing; Rate Problem

1. Model selection for the rate problem

Recent psychological interest in modern model selection has largely focused on the comparison and evaluation of cognitive models (e.g., Lee, 2004; Myung & Pitt, 1997; Navarro & Lee, 2004; Pitt, Myung, & Zhang, 2002). This is a worthwhile pursuit, because improving cognitive models is central to progress in psychology: models provide the formal expressions of theoretical ideas that can be evaluated against empirical observation. It is also true, however, that many routine data analysis problems in experimental psychology can profitably be viewed as model selection problems. Deciding whether data have "significant differences" can be conceived as deciding whether a simple model that assumes no differences is adequate, or whether a more complicated model allowing for differences is required. The basic goal—making inferences under uncertainty from limited and noisy data—is the same as for choosing between cognitive models. The differences are mostly just ones of degree. The models involved in data analysis usually have a smaller number of parameters than cognitive models, and generally have a simpler relationship to the observed data, often taking the form of well known statistical distributions.

In this paper, we compare the significance testing approach widely used in psychology with Bayesian and minimum description length (MDL) approaches for a data analysis problem known as the rate problem. In the rate problem, the question is whether or not the underlying rate of some observable phenomenon is the same in two populations, given a sample from each. Formally, if a phenomenon occurs k_1 times out of n_1 observations in one sample, and k_2 times out of n_2 observations in another, the rate problem is to determine whether the underlying rate of the phenomenon occurring is the same in both populations. The rate problem can be viewed as a model (or hypothesis) selection problem between a "same rate" and a "different rate" model. The "same rate" model M_s assumes that both populations have the same underlying rate θ. The

*Corresponding author. Fax: +61 8 8303 3770.
E-mail address: michael.lee@adelaide.edu.au (M.D. Lee).

0022-2496/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2005.11.010

"different rates" model M_d assumes that the first population has rate θ_1 while the second population has a potentially different rate θ_2.

There are many important real-world problems where the ability to make good statistical decisions for rate problems is (or historically has been) fundamental. For example, early investigations into the relationship between smoking and lung cancer (e.g., Wynder, 1954) relied on counts of non-smokers in samples of people with and without lung cancer. The Salk polio field vaccine trials relied on a comparison of the rates of evidence for the disease in large treatment and control groups (e.g., Francis et al., 1955). Part of the debate over the effectiveness of capital punishment examines whether there are differences in homicide rates for jurisdictions that do and do not have the death penalty (e.g., Sellin, 1980). Decisions in legal cases alleging discrimination can hinge on whether or not there are different rates of promotion for different demographic groups (e.g., DeGroot, Fienberg, & Kaldane, 1986, p. 9). More mundanely, rate problems occur in comparing the failure rates of different student groups, the levels of response to different advertising campaigns, the relative preferences for two products, and a range of other medical, biological, behavioral, social, and other issues in the empirical sciences.

In this paper, we develop significance testing, Bayesian and MDL criteria for choosing between the "same rate" and "different rates" models. After examining differences in the way the three criteria make decisions, we compare their performance in a range of realistic situations, including fixed sample sizes, sequential data gathering scenarios, and in cases where the data are generated by unknown processes. Based on these evaluations, we conclude by making some recommendations about the relative merits of the three criteria for rate problems, and discuss the implications of our results for data analysis in psychology more broadly.

2. Three decision criteria

2.1. Significance testing approach

2.1.1. Model selection theory

The significance testing approach is based on a null hypothesis that explains differences in observations by chance variation. Statistical decisions are made by measuring, via a test statistic, the probability that differences as extreme or more extreme than those observed would arise under the sampling distributions prescribed by the null hypothesis. If this probability is smaller than some fixed critical value α, the null hypothesis is rejected in favor of a (possibly unstated) alternative hypothesis that assumes some substantive account of the observed differences.

2.1.2. Application to the rate problem

The significance testing approach to the rate problem is to treat the "same rate" model as the null hypothesis, since it assumes there is no difference between the populations being sampled. The most commonly used test statistic (Fleiss, 1981, pp. 29–30) uses frequentist maximum likelihood estimators k_1/n_1 and k_2/n_2 for the population rates, and a Gaussian sampling distribution with a pooled variance estimate, as follows:

$$Z = \frac{k_1/n_1 - k_2/n_2}{\sqrt{(1/n_1 + 1/n_2)(k_1+k_2)/(n_1+n_2)(1-(k_1+k_2)/(n_1+n_2))}}.$$

2.2. Bayesian approach

2.2.1. Model selection theory

Bayesian statistics differs from the significance testing approach by using probability distributions, rather than estimators, to represent uncertainty. This means that the Bayesian approach to model selection is based on the posterior odds, given by

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)}{p(D|M_2)}\frac{p(M_1)}{p(M_2)},$$

where D are the data and M_1 and M_2 are competing models. The ratio $p(M_1)/p(M_2)$ gives the prior odds of the two models, and the ratio $p(D|M_1)/p(D|M_2)$, which is usually called the Bayes Factor, gives the relative evidence that the data provide for the models. Together these give the posterior odds for models based on the available data, and M_1 or M_2 is chosen depending on whether these odds are greater than or less than one.

For parameterized models, the Bayes Factor involves the marginal probability

$$p(D|M) = \int p(D|\theta, M)p(\theta|M)\,\mathrm{d}\theta,$$

and so requires the prior distribution of the parameters under the model. Under an "objective" Bayesian approach, the natural initial representation for Bayesian inference is one corresponding to complete ignorance. That is, the starting point of an analysis is one where nothing is known, and analysis proceeds by incorporating information as it becomes available (see Lee & Wagenmakers, 2005, for an overview).

The appropriateness of various methods for determining "non-informative" priors has been a matter of considerable debate (e.g., Kass & Wasserman, 1996). We follow Jaynes (2003, Chapter 12) in adopting transformational invariance methods for defining prior distributions corresponding to complete ignorance. This method relies on using the information inherent in the statement of a problem to constrain the choice of prior distribution. Intuitively, the idea is to consider ways in which a problem could be restated, so that it remains fundamentally the same problem, but is expressed in a different formal way. Prior distributions must necessarily be invariant under these transformations, since otherwise different ways of stating the same problem would lead to different inferences being

drawn. In general, the requirement of invariance from information inherent in a problem provides strong constraints on the choice of prior distributions, and often determines them uniquely.

2.2.2. Application to the rate problem

For the rate problem, complete ignorance about the rate of success is expressed by the prior distribution known as Haldane's prior. A rigorous derivation of this prior using transformational invariance is given by Jaynes (2003, pp. 382–385). The form of the prior arises because, consistent with the assumption of *complete* ignorance, it is not known whether successes and failures can both actually be observed. This leads to the extreme possibilities, with success rates of zero or one, being more probable, while still allowing for the possibility that the true rate is somewhere between zero and one. If, however, we do know that both success and failure are possible, Jaynes (2003, p. 385) shows that this additional piece of information leads, using maximum entropy principles, to the uniform distribution being the only possible objective prior. We think the second case is more realistic for most rate problems (i.e., most studies of rates already know both outcomes are possible), and so we use uniform priors on rate parameters throughout this paper.[1]

The "same rate" model has likelihood function

$$p(k_1,n_1,k_2,n_2|\theta,M_s) = \binom{n_1}{k_1}\theta^{k_1}(1-\theta)^{n_1-k_1}\binom{n_2}{k_2}\theta^{k_2}(1-\theta)^{n_2-k_2},$$

and, with the uniform prior, the marginal probability is

$$\begin{aligned} p(k_1,n_1,k_2,n_2|M_s) &= \int_0^1 \binom{n_1}{k_1}\theta^{k_1}(1-\theta)^{n_1-k_1}\binom{n_2}{k_2}\theta^{k_2}(1-\theta)^{n_2-k_2}\,\mathrm{d}\theta \\ &= \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n_1+n_2}{k_1+k_2}}\frac{1}{n_1+n_2+1}. \end{aligned}$$

The "different rates" model has likelihood function

$$p(k_1,n_1,k_2,n_2|\theta_1,\theta_2,M_d) = \binom{n_1}{k_1}\theta_1^{k_1}(1-\theta_1)^{n_1-k_1}\binom{n_2}{k_2}\theta_2^{k_2}(1-\theta_2)^{n_2-k_2},$$

with the marginal probability

$$\begin{aligned} p(k_1,n_1,k_2,n_2|M_d) &= \int_0^1\int_0^1 \binom{n_1}{k_1}\theta_1^{k_1}(1-\theta_1)^{n_1-k_1} \\ &\quad\times\binom{n_2}{k_2}\theta_2^{k_2}(1-\theta_2)^{n_2-k_2}\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2 \\ &= \binom{n_1}{k_1}\binom{n_2}{k_2}\int_0^1\int_0^1 \theta_1^{k_1}(1-\theta_1)^{n_1-k_1} \\ &\quad\times\theta_2^{k_2}(1-\theta_2)^{n_2-k_2}\,\mathrm{d}\theta_1\,\mathrm{d}\theta_2 \\ &= \frac{1}{(n_1+1)(n_2+1)}. \end{aligned}$$

The Bayes Factor comparing the models is

$$B = \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n_1+n_2}{k_1+k_2}}\frac{(n_1+1)(n_2+1)}{n_1+n_2+1},$$

and is equal to the posterior odds under the assumption of equal prior probabilities for the models. This ratio provides a Bayes criterion for making decisions with rate problems, choosing the "same rate" model when it is greater than one, and the "different rates" model when it is less than one.

2.3. MDL approach

2.3.1. Model selection theory

The MDL approach views models as fixed codes, and chooses the one that best compresses the data. A series of successively more exact and general stochastic complexity criteria that implement this principle have been developed by Rissanen (1978, 1987, 1996, 2001). The most recent of these criteria, the normalized maximum likelihood (NML), solves the minimax problem

$$\inf_q \sup_{g\in G} E_g \ln\frac{p(D|\theta^*(D),M)}{q(D)}$$

of finding the code that has the minimum worst-case increase in coding the data over the optimal code length. Here $p(D|\theta^*(D),M)$ is the probability of the data under the maximum likelihood estimate of the model parameters, $q(\cdot)$ is an ideal code, and G ranges over *all* distributions[2] over D of the given sample size n.

[1] We are aware some readers may be concerned that the uniform prior we use is not reparameterization invariant. It is true that, for some modeling problems, reparameterization invariance is a fundamental requirement. These are problems where the statistical model is essentially a "black box", and various parameterizations are possible to provide different (but equivalent) formalisms for indexing the distributions over data that constitute the model. The rate problem we are considering is not of this type. Our problem contains *additional* information about the parameters, by identifying them as rates that have specific quantitative meaning. It is the invariances associated with this meaning that Jaynes (2003) uses to derive Haldane and uniform priors for rate parameters. Of course, it would be possible to redo our analyses using Jeffreys' priors, which would take the form of $\mathrm{Beta}(\theta|0.5,0.5)\propto 1/\sqrt{\theta(1-\theta)}$ distributions, and achieve reparameterization invariance. For even a small number of data, the similarity of the Jeffreys' priors and the uniform priors we use means the results would be almost identical. But, an analysis based solely on reparameterization invariance would be sub-optimal, because it would ignore some of the available information.

[2] This means that G includes, for example, the Markov generating processes consider later, not just the set of all distributions corresponding to the "same rate" model. We are grateful to a reviewer for making this point clear.

The NML solution to the minimax problem, given by

$$\mathrm{NML} = \frac{p(D|\theta^*(D), M)}{\int_{\theta^*(D')\in\Omega} p(D'|\theta^*(D'), M)\,\mathrm{d}D'},$$

normalizes the maximum likelihood of the observed data D by the maximum likelihood of the model over all possible data D' that are indexed by the parameter space Ω under $\theta^*(\cdot)$. For model selection, the model with the maximal NML value is chosen.

2.3.2. Application to the rate problem

The partial derivative of the likelihood function for the "same rate" model, with respect to the rate parameter θ, is

$$\frac{\partial \ln p(i_1, n_1, i_2, n_2|\theta)}{\partial\theta} = \frac{i_1 + i_2}{\theta} - \frac{n_1 + n_2 - i_1 - i_2}{1-\theta},$$

and so the maximum likelihood function is given by

$$\theta^*(i_1, i_2) = \frac{i_1 + i_2}{n_1 + n_2}.$$

This means the NML for the "same rate" model is

$$\frac{p(k_1, n_1, k_2, n_2|\theta^*(k_1, k_2), M_\mathrm{s})}{\sum_{i_1=0}^{n_1}\sum_{i_2=0}^{n_2} p(k_1, n_1, k_2, n_2|\theta^*(i_1, i_2), M_\mathrm{s})} = \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}((k_1+k_2)/(n_1+n_2))^{k_1+k_2}(1-(k_1+k_2/n_1+n_2))^{n_1+n_2-k_1-k_2}}{\sum_{i_1=0}^{n_1}\sum_{i_2=0}^{n_2}\binom{n_1}{i_1}\binom{n_2}{i_2}((i_1+i_2)/(n_1+n_2))^{i_1+i_2}(1-(i_1+i_2/n_1+n_2))^{n_1+n_2-i_1-i_2}}.$$

For the "different rates" model, the maximum likelihood parameter estimates are

$$\theta_1^*(i_1, i_2) = \frac{i_1}{n_1}, \quad \theta_2^*(i_1, i_2) = \frac{i_2}{n_2},$$

and so the NML for the "different rates" model is

$$\frac{p(k_1, n_1, k_2, n_2|\theta_1^*(k_1, k_2), \theta_2^*(k_1, k_2), M_\mathrm{d})}{\sum_{i_1=0}^{n_1}\sum_{i_2=0}^{n_2} p(k_1, n_1, k_2, n_2|\theta_1^*(i_1, i_2)\theta_2^*(i_1, i_2), M_\mathrm{d})} = \frac{\binom{n_1}{k_1}(k_1/n_1)^{k_1}(1-(k_1/n_1))^{n_1-k_1}\binom{n_2}{k_2}(k_2/n_2)^{k_2}(1-(k_2/n_2))^{n_2-k_2}}{\sum_{i_1=0}^{n_1}\sum_{i_2=0}^{n_2}\binom{n_1}{i_1}(i_1/n_1)^{i_1}(1-(i_1/n_1))^{n_1-i_1}\binom{n_2}{i_2}(i_2/n_2)^{i_2}(1-(i_2/n_2))^{n_2-i_2}}.$$

To make a decision for the rate problem, the "same rate" or "different rates" model is chosen according to which has the greater NML value.

3. Comparison

To compare the significance testing, Bayesian, and MDL approaches to the rate problem, we restrict ourselves to choosing one of the models under the assumption that both are a priori equally likely. We also assume that the utility of decision making equally weights both correct and incorrect decisions for both models. For the standard Z statistic, we consider critical values corresponding to the widely used α levels of 0.01, 0.05, and 0.10.

Fig. 1 shows the decision boundaries of the criteria for all possible counts k_1 and k_2 in samples of size $n_1 = 20$ and $n_2 = 10$, respectively. Every point correspond to a possible pair of sample counts, and the decision bounds for the criteria show their model selection behavior. Counts falling inside the decision bound of a criterion result in the "same rate" model being chosen; counts falling outside result in the "different rates" model being chosen. Each of the four panels compares the Bayes criterion decision bound with one other criterion. In this way, the Bayes criterion provides a visual standard reference to compare all of the decision bounds.

Fig. 1 shows that the Bayes and NML criteria agree for all but 14 of the possible decisions, with the NML being more conservative in choosing the "same model rate" in 12 of these cases. The significance testing approach, on the other hand, is much more conservative than the Bayes and NML criteria when $\alpha = 0.01$, but makes progressively more similar decisions as α increases to 0.05 and 0.10.

Fig. 2 shows the decision boundaries for the "same rate" criteria for all possible counts k_1 and k_2 in larger samples of size $n_1 = 100$ and $n_2 = 50$, respectively. The Bayes and NML criteria now make extremely similar decisions that cannot be distinguished visually. The significance testing approach is now more conservative when $\alpha = 0.01$, very similar when $\alpha = 0.05$ and less conservative when $\alpha = 0.10$.

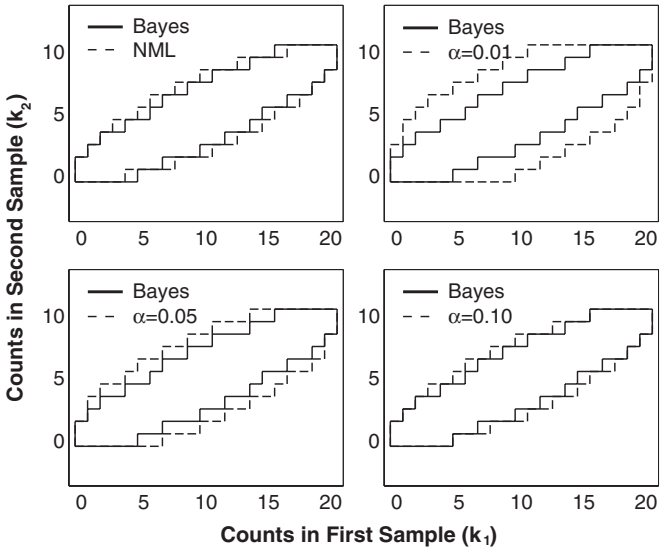

Fig. 1. The decision boundaries for the Bayes, NML and significance testing criteria, for all possible counts k_1 and k_2 in samples of size $n_1 = 20$ and $n_2 = 10$, respectively.

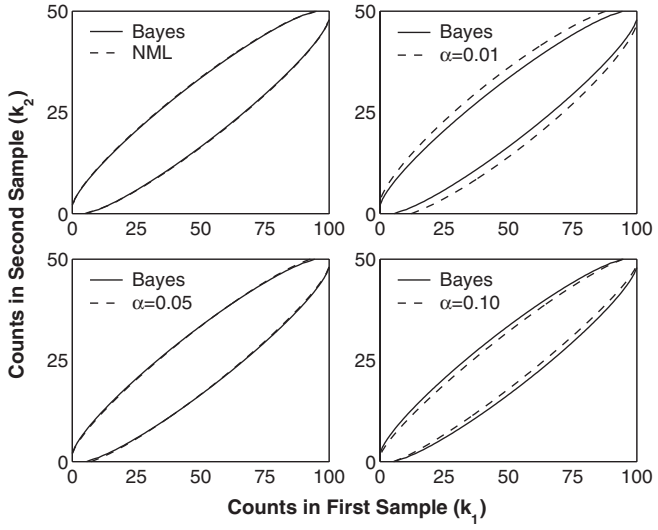

Fig. 2. The decision boundaries for the Bayes, NML and significance testing criteria, for all possible counts k_1 and k_2 in samples of size $n_1 = 100$ and $n_2 = 50$, respectively.

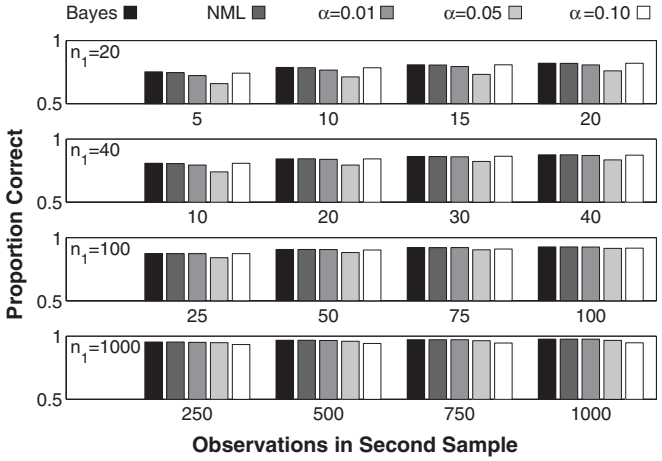

Fig. 3. Accuracy of Bayesian, NML, and significance testing decision criteria for 16 rate problems. The four panels (top to bottom) correspond to the cases where the first sample has $n_1 = 20$, 40, 100, and 1000 observations. Within each panel, four different possibilities for the number of observations in the second sample n_2 are shown.

Together, the results in Figs. 1 and 2 suggest two important conclusions. The first is that the Bayes and NML criteria make very similar decisions, especially as the sample sizes increase. The second is that the relationship between the significance testing approach and the Bayes and NML criteria depends on an interaction between the critical value and the sample sizes. In particular, the Bayes and NML criteria make decisions consistent with large critical values for small sample sizes, but with progressively smaller critical values for larger sample sizes.

4. Evaluation

Given the existence of differences in the decisions made by the Bayes and NML criteria on the one hand, and the significance testing criterion at different critical levels on the other hand, we undertook a series of evaluations as to which makes more accurate decisions.

4.1. Evaluation with fixed sample sizes

The first evaluation examines the accuracy of the criteria in a range of fixed sample size pairings. This evaluation corresponds to situations where limited data can be collected, and so assumes the largest possible samples are available for analysis.

Fig. 3 summarizes the accuracy of the criteria on 16 rate problems with different sample sizes. The first sample has a size of 20, 40, 100 or 1000 observations, while the second sample has a size beginning at one quarter this number of observations and increasing to the same size as the first sample. For each of these possible combinations, 10^5 trials,[3] giving data for both samples, were generated as follows. On "same rate" trials, a rate parameter was randomly chosen from the uniform distribution on the interval (0,1). Two independent counts were then drawn from a binomial distribution with this rate parameter, using the appropriate sample size as the number parameter. On "different rates" trials, two different rate parameters were independently chosen, under the restriction that they differed by at least 0.1, and counts were generated from the appropriate binomial distributions.[4] For both "same rate" and "different rates" trials, each of the model selection criteria was applied to the data, and was evaluated as being either correct or incorrect. Whether a individual trial was a "same rate" or "different rates" trial was independently and randomly determined, with equal probability given to both possibilities.

Fig. 3 shows the mean accuracy of all criteria, for all of the rate problems considered. It can be seen that the Bayes and NML criteria are approximately equally accurate for all problems, and that they often more accurate, and never less accurate, than the significance testing approach at its various α values. The superiority of the Bayes and NML criteria is particularly evident for small sample sizes.[5] The other result worth noting in Fig. 3 is that the critical level

[3]This number of trials is large enough that standard error bars would be indistinguishable visually from the mean accuracies shown in Fig. 3.

[4]Ensuring the rates differed by at least 0.1 was intended to make them "meaningfully" different, in the context of scientific modeling. Intuitively, for example, if two phenomena occur with underlying rates of 0.3 and 0.4, it is important to identify this difference to understand the phenomena and make useful predictions. If, on the other hand, the two rates are 0.31 and 0.32, the same model could be regarded as more useful than the different model. The choice of 0.1 was a subjective one, which seemed to us to capture a minimum difference that would be meaningful for rate problems.

[5]Advocates of the significance testing approach might argue that more exact tests ought to be applied for the case $n_1 = 20$, $n_2 = 5$. This is probably wise, but raises the problem of deciding when to change method, and highlights that the Bayes and NML criteria have the advantage of being applicable for all sample sizes.

corresponding to the greatest accuracy for the significance testing approach varies across the problems. For small sample sizes, the choice $\alpha = 0.10$ gives the greatest accuracy, but for larger sample sizes, the choice $\alpha = 0.01$ leads to better performance. In the context of the earlier comparisons in Figs. 1 and 2, this suggests the greatest significance testing accuracy is achieved by setting critical levels that align its decision boundaries with those of the Bayes and NML criteria.

4.2. Evaluation with increasing number of data

Our second evaluation examines the pattern of increase in accuracy of the criteria as the number of data available increases. This evaluation corresponds to situations where additional data are available, but may require additional resources, and so the interest is in the trade-off between accuracy and the number of data.

Fig. 4 summarizes the results of evaluating the accuracy of all criteria with increasing sample sizes across 10^5 independent trials. As before, each trial was independently and randomly chosen to be either a "same rate" or "different rates" trial with equal probability. On "same rate" trials, a rate parameter for both samples was randomly chosen from the uniform distribution on the interval (0,1). On "different rates" trials, two different rate parameters were chosen independently, again under the restriction that differed by at least 0.1. A datum was then generated using the Bernoulli distribution with the first rate parameter. The currently available data, corresponding to counts of successes k_1 and k_2 from total observations n_1 and n_2 for both populations, were then used to make a decision by all of the criteria, and the accuracy of this decision was evaluated. An additional datum was then generated from the second population, all of the available data used to make decisions, and those decisions evaluated.

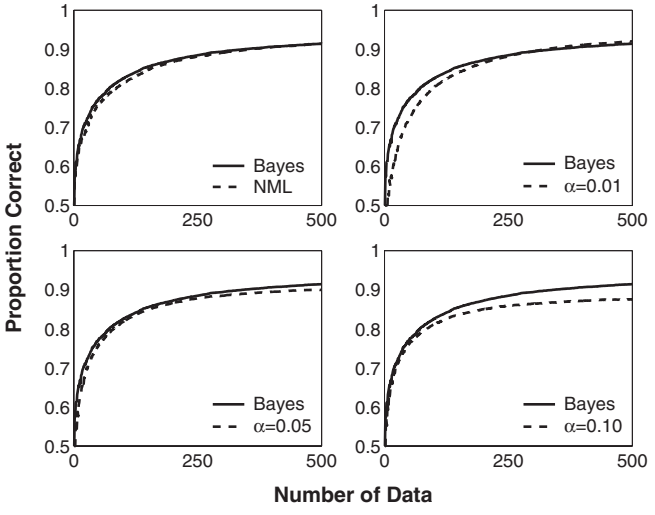

Fig. 4. The accuracy of Bayesian, NML, and significance testing decision criteria for rate problems as a function of the number of data available.

This process was iterated 250 times, giving a total of 500 data at the end of each trial.

Fig. 4 shows the pattern in change in mean accuracy for each of the criteria, as a function of the number of available data. As before, the performance of the Bayes criterion is shown in all four panels to provide a standardized reference for the other four curves. It can be seen that, once again, the Bayes and NML criteria perform extremely similarly. In addition, these criteria are always more accurate than the significance testing approach for some range of the number of data. They are more accurate than the significance testing approach with $\alpha = 0.01$ and 0.05 when few data are available, and more accurate than significance testing criteria with $\alpha = 0.05$ and 0.10 when many data are available.

4.3. Evaluation against significance testing performance measures

A valid criticism of the preceding evaluations is that they rely on a 'proportion correct' accuracy measure that significance testing methods do not seek to optimize. Rather, significance testing methods are designed to minimize so-called "Type II" errors (i.e., the probability of retaining the null when it is false) at an α criterion that explicitly fixes Type I errors (i.e., the probability of rejecting the null when it is true) at a constant value.

To address this criticism, we compared the Types I and II errors for the Bayes and significance testing criteria on four different rate problems with fixed sample sizes. We considered every possible sample that could be observed for each problem, and calculated the probability of each of these data sets under the "same rate" and "different rates" modeling assumptions made previously. We then applied both the significance testing and Bayesian criteria in the way they could be applied in practice.

For significance testing, we considered fixed α values of 0.001, 0.01, 0.05, and 0.10, which correspond to those most commonly used in the psychological literature. For the Bayesian criterion, we considered progressively more stringent levels of evidence for deciding that the "different rates" model should be inferred, which provides the Bayesian analogue of setting progressively more conservative α values. Following the well-known suggested standards for Bayes Factors proposed by Kass and Raftery (1995, p. 777), we considered decision thresholds on the (natural) log-odds scale at 0, 2, 6 and 10, corresponding to increasingly strong evidence being required before deciding the rates are different.

The quality of the decisions made by both of these approaches, for the problems we considered, is shown in Fig. 5. In each case, the downward pointing triangles correspond, from left to right, to 0.001, 0.01, 0.05 and 0.10 significance testing levels, while the upward pointing triangles correspond, from left to right, to the application of the 10, 6, 2 and 0 Bayesian levels. Also shown in Fig. 5 by the two lines is the performance that would be achieved

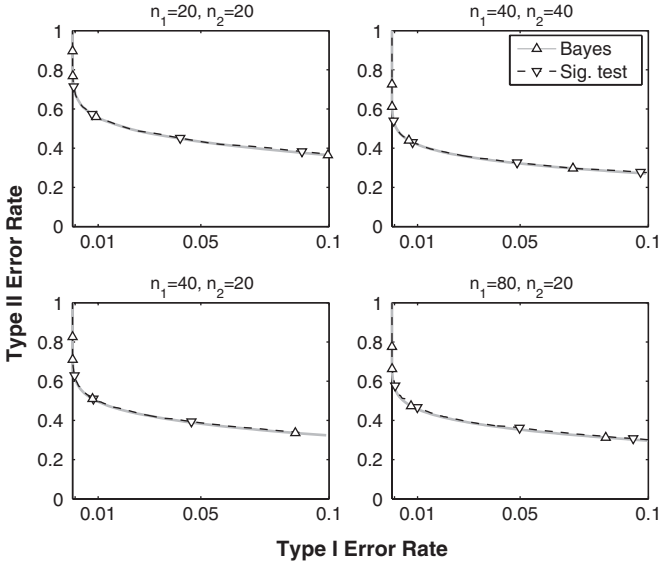

Fig. 5. Relationship between Types I and II errors for the Bayes and significance testing criteria on four rate problems.

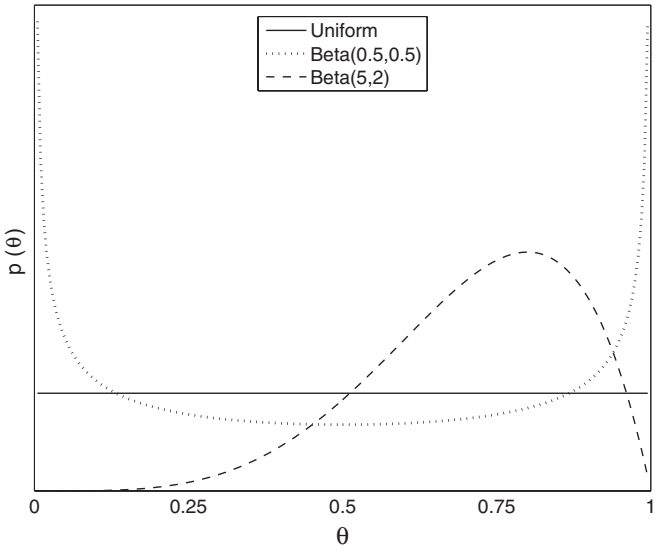

Fig. 6. The uniform, Beta($\theta|0.5, 0.5$), and Beta($\theta|5, 2$) distributions for a rate θ.

by applying other α levels for significance testing, or other evidence levels for Bayesian inference. It can be seen that, for all of the problems, the performance of the two criteria is virtually identical.

4.4. Evaluation against data generated from different distributions

Although our use of uniform priors in the Bayesian inference is justified by the available information, it is true that, in practice, it is unlikely that observed data will meet these distributional assumptions. In this sense, the preceding evaluations, by drawing simulated data from uniform distributions over rates, could be argued to favor the Bayesian criterion.

To address this criticism, we repeated the first two evaluations drawing simulated data from different distributions. In particular, we drew rates from the general Beta distribution $\text{Beta}(\theta|a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$, using the values $a = b = 0.5$ and $a = 5$, $b = 2$. These distributions are contrasted with the uniform prior assumed by the Bayesian analysis in Fig. 6. It is worth noting that both alternative distributions deviate from the uniform distribution in different and important ways. For example, the Beta($\theta|0.5, 0.5$) distribution might loosely be regarded as "multimodal", in the sense that it increases in two different regions, while the Beta($\theta|5, 2$) distribution is negatively skewed and gives almost no density to possible rates that are supported by the uniform prior.

For the Beta($\theta|5, 2$) case, the performance of the criteria for fixed sample sizes is shown in Fig. 7, and performance as the number of data available increase is shown in Fig. 8. In both cases, the results are extremely similar to those observed for the uniform generating distribution in Figs. 3 and 4 and the same comparative conclusions seem to be warranted.

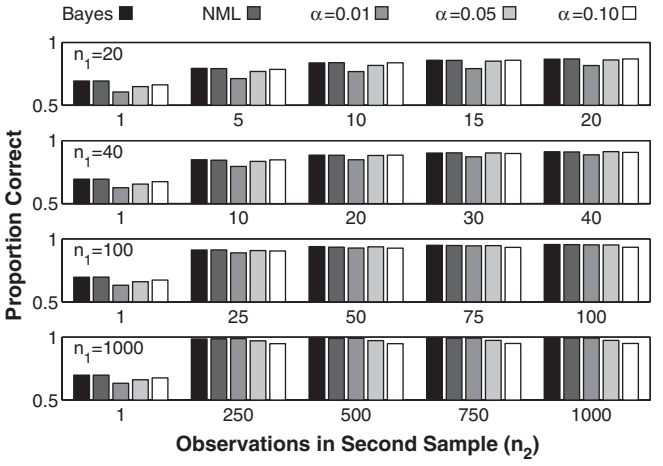

Fig. 7. The performance on the Bayesian, NML, and significance testing decision criteria for 20 rate problems with data drawn from a Beta($\theta|0.5, 0.5$) distribution.

For the Beta($\theta|5, 2$) case, the performance of the criteria for fixed sample sizes is shown in Fig. 9, and performance as the number of data available increase is shown in Fig. 10. In this case, the overall performance of all of the criteria is worse, but their relative levels of performance show the same patterns. Once again, the Bayes and NML criteria perform extremely similarly, and these criteria are almost always as accurate or more accurate than the significance testing approach.

4.5. Evaluation against data generated with sequential dependencies

Another potential criticism of our evaluations is that some advocates (e.g., Rissanen, 2001) argue MDL is

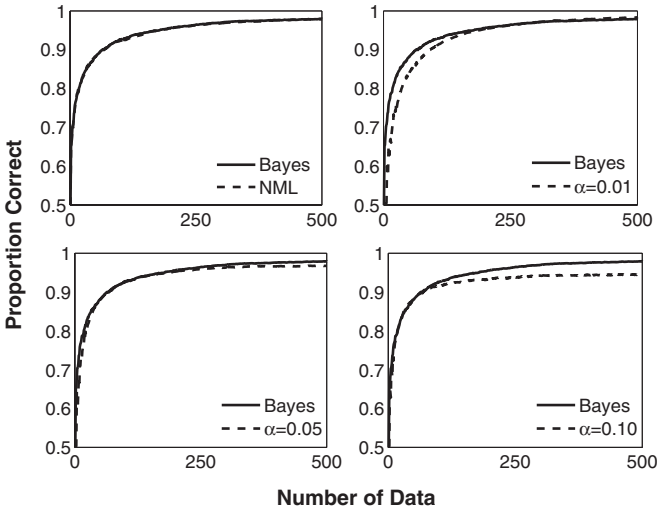

Fig. 8. The performance on the Bayesian, NML, and significance testing decision criteria as a function of the number of data available with data drawn from a Beta($\theta|0.5, 0.5$) distribution.

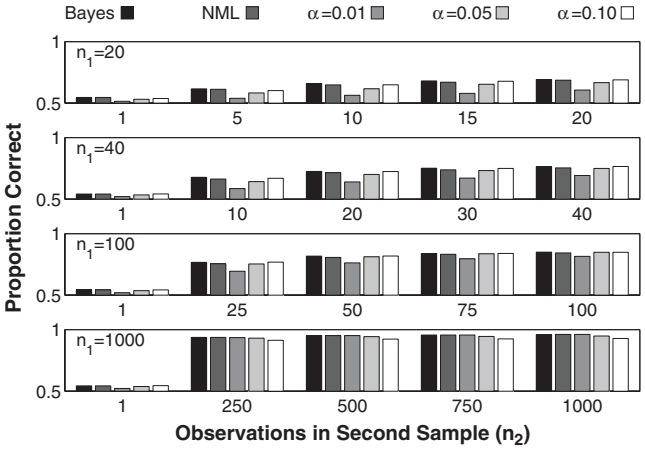

Fig. 9. The performance on the Bayesian, NML, and significance testing decision criteria for 20 rate problems with data drawn from a Beta($\theta|5, 2$) distribution.

superior to the Bayesian approach when data are generated with statistical properties that are different from those expressible by the models being compared. The preceding evaluations use as generating models the "same rate" and "different rates" models used to derive the Bayes criterion. For this reason, it is possible that, for real-world data, which are almost certainly not generated by such simple processes, the NML criterion could be more accurate.

To test this possibility, following the ideas discussed by Grünwald (1998, pp. 24–28), we repeated the two evaluations using a first-order Markov process to generate the data. This process specifies a probability γ_1 that the next observation will be a success given that the previous observation was a success, and a different probability γ_2 that the next observation will be a success given that the previous observation was not. In this way, Markov processes introduce dependencies between successive observations, and so violate the independency assumptions of the "same rate" and "different rates" models used to derive the Bayes criterion.

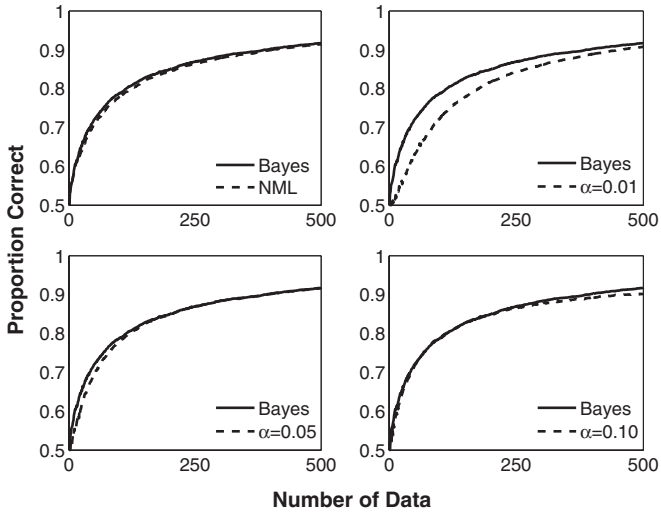

Fig. 10. The performance on the Bayesian, NML, and significance testing decision criteria as a function of the number of data available with data drawn from a Beta($\theta|5, 2$) distribution.

Any choice of γ_1 and γ_2, however, is naturally associated with a single rate θ according to the relationship $\theta = \gamma_1\theta + \gamma_2(1-\theta)$. This relationship allows different first-order Markov chains, specified by γ_1 and γ_2 to be assessed as either the same or different according to their associated θ rates. In this way, the Bayes and NML criteria can be assessed on rate problems where the data are generated by distributions that are different from the basic "same rate" and "different rates" models.

Figs. 11 and 12 show the results of repeating our first two evaluations using first-order Markov processes to generate the data. Fig. 11 shows the performance of the criteria for fixed sample sizes, Fig. 12 shows performance as the number of data available increase. The Bayes criterion is always either as accurate, or slightly more accurate, than the NML criterion in both analyses. Once again, these criteria outperform the significance testing approach, except perhaps for a moderate number of data and a stringent α level.

5. Discussion

5.1. Conclusions for the rate problem

Our comparisons and evaluations of significance testing, Bayesian and MDL approaches in solving the rate problem warrant two important conclusions.

5.1.1. Bayes and NML

First, the Bayes and NML criteria make extremely similar decisions for all counts and all sample sizes, and become indistinguishable for large sample sizes. The known asymptotic equivalence for exponential families of

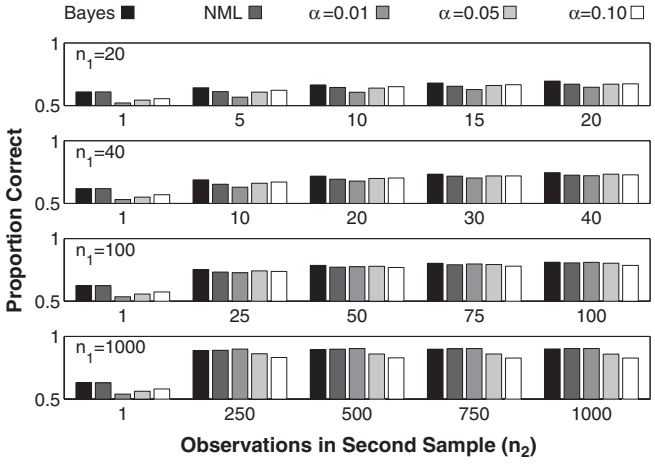

Fig. 11. The performance on the Bayesian, NML, and significance testing decision criteria for 20 rate problems with data drawn from a first-order Markov process.

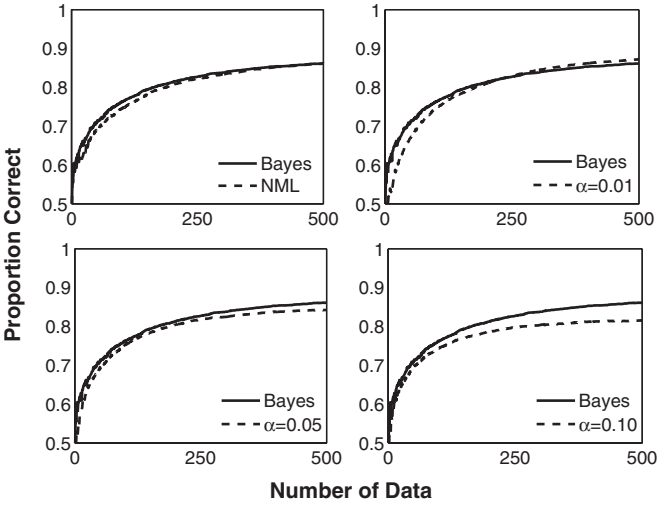

Fig. 12. The performance on the Bayesian, NML, and significance testing decision criteria as a function of the number of data available with data drawn from a first-order Markov process.

NML codelengths and log-Bayesian marginal likelihoods with Jeffreys' priors (Barron, Rissanen, & Yu, 1998; Grünwald, 2005; Rissanen, 1996) means the result for large samples is not surprising, given the similarities between the Jeffreys' prior and uniform prior for this problem. It is interesting, however, to observe how quickly the two measures converge for the sample sizes typical in psychology, and using data generated ways that differ in one aspect or another from those assumed by the Bayes criterion.

In practice, one potential advantage of the Bayes criterion is that it is relatively less demanding to compute. The denominator of the NML criterion involves a large number of terms for problems dealing with large sample sizes. This has necessitated the development of elegant and sophisticated recursive methods for the efficient calculation of the NML criterion (Kontkanen, Buntine, Myllymäaki, Rissanen, & Tirri, 2003) that are not trivial to implement. The Bayes criterion, on the other hand, is conceptually and computationally easy to calculate for any plausible sample size. Given the result that the Bayes and NML criteria behave very similarly theoretically, this practical advantage encourages the use of the Bayesian approach.

5.1.2. *Bayes and significance testing*

The second conclusion is that the Bayesian criterion perform at least as well as the significance testing criteria. An elegant way of summarizing all of the results we have presented,[6] is by noting the natural symmetry between those tests favoring the Bayesian perspective and those favoring the significance testing perspective. In those comparisons focusing on decision accuracy, with data generated from known (if mis-specified) prior distributions, the Bayes criterion outperforms significance testing, unless significance testing is allowed to vary its α level as a function of sample size. In those comparisons focused on fixing Type I errors, however, the Bayes criterion needs an analogous flexibility in determining its evidence threshold to match the decision-making performance of significance testing.

Based on this analysis, our conclusion is that the comparisons presented here demonstrate that the Bayes criterion is at least as well performed as significance testing. Perhaps this is not surprising, since Jaynes (2003, p. 550; see also Lee & Wagenmakers, 2005) argued that significance testing criteria are usable and useful when dealing with the problems where all the relevant information comes (or can accurately be conceived as coming) from independent runs of simple random experiments. This means that significance testing works well when (1) the variables of interest vary according to simple distributions, in which a small number of parameters adequately describe their distributional form; (2) there is no important prior information available about the variables of interest; and (3) there are many data. Each of these conditions is satisfied by the rate problems we have considered.

Nevertheless, it is easy to imagine variations on the rate problem that would violate the last two conditions, by presenting strongly constraining prior information about the rates of one or both of the data sources, or by having access to only a small number of data. Interestingly, both of these possibilities loom larger in applied rather than laboratory settings. In the real-world much is usually already known about a problem before data are collected or observed, and it is often the case that additional data are expensive, dangerous, or otherwise difficult to obtain.

5.2. *Model selection and data analysis in psychology*

More generally, we think our results highlight the potential of using Bayesian and MDL methods for

[6]We are extremely grateful to a reviewer for providing this insight.

analyzing psychological data. For a combination of historical and perhaps sociological reasons, most statistical inferences in psychology are made using significance testing methods, despite compelling evidence that they can be inefficient, inapplicable or even pathological (e.g., Berger & Wolpert, 1984; Edwards, Lindman, & Savage, 1963; Jaynes, 2003; Lindley, 1972). It is especially surprising that Bayesian methods are not more widely used, given many of the questions experimental psychology typically asks of its data are naturally interpreted as model selection questions. For example, rather than use *t*-tests, Lee, Loughlin, and Lundberg (2002) drew inferences about whether sets of scores had the same means using Bayes Factors; Lee and Cummins (2004) and Lee and Corlett (2003) used Bayes Factors to decide whether accuracy, confidence and response time distributions were significantly different; Karabatsos (2005) used Bayes Factors to compare deterministic axiomatic accounts of choice and judgment; and Vickers, Lee, Dry, and Hughes (2003) used Bayes Factors to compare competing meaningful interpretations of data from a factorial experimental design that would usually be subjected to ANOVA methods. To the extent that our findings for the rate problem generalize—and Bayesian and MDL methods prove to make good statistical inferences—experimental psychology would benefit from adopting modern model selection methods for data analysis.

Acknowledgments

We wish to thank Yong Su, Eric-Jan Wagenmakers, Lourens Waldorp, and two anonymous referees for very helpful comments.

References

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, *44*(6), 2743–2760.

Berger, J. O., & Wolpert, R. L. (1984). *The likelihood principle*. Hayward, CA: Institute of Mathematical Statistics.

DeGroot, M. H., Fienberg, S. E., & Kaldane, J. B. (1986). *Statistics and the law*. New York: Wiley.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (Second ed.). New York: Wiley.

Francis, T., Jr., Korns, R., Voight, R., Boisen, M., Hemphill, F., & Napier, J. (1955). An evaluation of the 1954 poliomyelitis vaccine trials: Summary report. *American Journal of Public Health*, *45*(supplement), 1–50.

Grünwald, P. D. (1998). *The minimum description length principle and reasoning under uncertainty*. Amsterdam: Institute for Logic, Language and Computation, University of Amsterdam.

Grünwald, P. D. (2005). Minimum description length tutorial. In P. D. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (pp. 23–80). Cambridge, MA: MIT Press.

Jaynes, E. T. (2003). In G. L. Bretthorst (Ed.), *Probability theory: The logic of science*. New York: Cambridge University Press.

Karabatsos, G. (2005). The exchangeable multinomial model as an approach to testing deterministic axioms of choice and measurement. *Journal of Mathematical Psychology*, *49*, 51–69.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *50*(430), 773–795.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*(435), 1343–1370.

Kontkanen, P., Buntine, W., Myllymäaki, P., Rissanen, J., & Tirri, H. (2003). Efficient computation of stochastic complexity. In C. M. Bishop, & B. J. Frey (Eds.), *Proceedings of the ninth international workshop on artificial intelligence and statistics* (pp. 181–188). NY: Society for Artificial Intelligence and Statistics.

Lee, M. D. (2004). An efficient method for the minimum description length evaluation of cognitive models. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the cognitive science society* (pp. 807–812). Mahwah, NJ: Erlbaum.

Lee, M. D., & Corlett, E. Y. (2003). Sequential sampling models of human text classification. *Cognitive Science*, *27*(2), 159–193.

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and "rational" models. *Psychonomic Bulletin & Review*, *11*(2), 343–352.

Lee, M. D., Loughlin, N., & Lundberg, I. B. (2002). Applying one reason decision making: The prioritization of literature searches. *Australian Journal of Psychology*, *54*(3), 137–143.

Lee, M. D., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*(3), 662–668.

Lindley, D. V. (1972). *Bayesian statistics: A review*. Philadelphia, PA: SIAM.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin & Review*, *11*(6), 961–974.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 445–471.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society*, *Series B*, *49*(3), 223–239, 252–265.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*(1), 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*(5), 1712–1717.

Sellin, T. (1980). *The penalty of death*. Beverly Hills, CA: Sage.

Vickers, D., Lee, M. D., Dry, M., & Hughes, P. (2003). The roles of the convex hull and number of intersections upon performance on visually presented traveling salesperson problems. *Memory & Cognition*, *31*(7), 1094–1104.

Wynder, E. L. (1954). Tobacco as a cause of lung cancer with special reference to the infrequency of lung cancer among non-smokers. *The Pennsylvania Medical Journal*, *57*(11), 1073–1083.