Bayes Factors and BIC

Comment on "A Critique of the Bayesian Information Criterion for Model Selection"

ADRIAN E. RAFTERY
University of Washington

I would like to thank David L. Weakliem (1999 [this issue]) for a thought-provoking discussion of the basis of the Bayesian information criterion (BIC). We may be in closer agreement than one might think from reading his article. When writing about Bayesian model selection for social researchers,[1] I focused on the BIC approximation on the grounds that it is easily implemented and often reasonable, and simplifies the exposition of an already technical topic. As Weakliem says, BIC corresponds to one of many possible priors, although I will argue that this prior is such as to make BIC appropriate for baseline reference use and reporting, albeit not necessarily always appropriate for drawing final conclusions. When writing about the same subject for statistical journals,[2] however, I have paid considerable attention to the choice of priors for Bayes factors. I thank Weakliem for bringing this subtle but important topic to the attention of sociologists.

In 1986, I proposed replacing *P* values by Bayes factors as the basis for hypothesis testing and model selection in social research, and I suggested BIC as a simple and convenient, albeit crude, approximation. Since then, a great deal has been learned about Bayes factors in general, and about BIC in particular. Weakliem seems to agree that the Bayes factor framework is a useful one for hypothesis testing and model selection; his concern is with how the Bayes factors are to be evaluated.

Weakliem makes two main points about the BIC approximation. The first is that BIC yields an approximation to Bayes factors that corresponds closely to a particular prior (the unit information prior) on

SOCIOLOGICAL METHODS & RESEARCH, Vol. 27 No. 3, February 1999 411-427

the model parameters. This may or may not be similar to the researcher's actual prior information or beliefs, and the Bayes factor it yields may or may not be close to the Bayes factor resulting from the researcher's prior.

This is clearly true, although I feel that the conclusions about BIC that Weakliem draws from it are somewhat overstated. The unit information prior is often a reasonable, well spread out reference prior for Bayes factors, and I discuss the rationale for it in more detail below. Most of the criticisms of the unit information prior on which BIC is based imply that it is too spread out. Now, usually, the less spread out the prior, the more the Bayes factor favors the alternative hypothesis when the models are nested; that is, the more evidence it implies for the "effect" being studied. Thus, in most cases, BIC is more conservative than the alternative priors proposed by Weakliem and other critics of BIC, in the sense that the alternative priors are more likely to find evidence for the effect being studied.

In practice, therefore, there seems to be some agreement that BIC is sufficiently conservative, perhaps too conservative. It follows that, in most cases, if BIC finds evidence for an effect, we should agree that the data support the effect (although not necessarily conversely). It follows that, if the Bayes factor based on the researcher's prior supports an effect, but BIC does not, the decisive additional evidence comes from the researcher's prior, and this should be made explicit in the research report. Of course, this does not mean that the researcher's prior should not be used, as it may well command widespread agreement; rather, the prior used should be clearly set out so that other researchers can decide whether they agree with it or not. This also seems to argue in favor of reporting BIC as a baseline reference analysis, even if the final conclusions are drawn using a different prior.

But how can one obtain accurate Bayes factors based on other priors? Here again, Weakliem and I are in broad agreement. Ideally, the researcher should carefully assess his or her prior and compute the Bayes factor based on it, at the end checking that the conclusions are not too sensitive to the precise prior specification via sensitivity analysis. There are two difficulties with this, which Weakliem seems to find daunting, but which I think we are well on the way to overcoming. The first is prior assessment, and here I think Weakliem has provided useful guidance with his two examples: assessing the prior distribution of

an odds ratio in his section on Bayes Factors and Prior Distributions, and the prior distribution of the asymmetry parameter in his section on Sample Size and the BIC. The kind of thought experiment that he describes is a good way to assess priors, and his laying that out is a major contribution of the article.

The second difficulty is that of computing the relevant integrals. Both the mathematical solution and the software for doing this now exist for many types of statistical models. For many models used in social research, Bayes factors can be computed almost exactly using the Laplace method (Raftery 1996)—essentially a more exact form of Weakliem's equation (2)—and this can be done readily using the GLIB software, which is freely and publicly available (see below for details).

This general strategy provides more accurate Bayes factors and in the process overcomes the difficulties with the BIC described by Weakliem. In particular, it removes the need for ad hoc proposals such as $MBIC_1$ and $MBIC_2$, neither of which approximates a Bayes factor for any prior, as far as we know, and which in this sense are less transparent than BIC itself. I have applied this strategy to Weakliem's examples using his own preferred priors, and the qualitative conclusions in each case are similar to those to be drawn from BIC.

Weakliem's final point is that in the social mobility example, he was able to find a better model that researchers who used BIC did not find. I congratulate him on this, but I do not think it discredits BIC. BIC finds Weakliem's preferred model to be better than any of the other models considered—so is it doing the wrong thing? Many researchers have analyzed the social mobility data set, some of them using BIC and some not, so the fact that previous researchers did not discover Weakliem's preferred model tells us little about how good BIC is; rather, it tells us that Weakliem is very good at data analysis!

Weakliem has misunderstood my advice about model building; what I advocated and illustrated in Raftery (1995, sec. 7) is actually quite close to what he did in his analysis of the social mobility example.[3] Both model checking and the search for better models should run their full course; the fact that Bayes factors prefer one model to another does not mean that one should stick with the current best model if substantial amounts of deviance remain to be explained or if the current model is not parsimonious or interpretable. Indeed, Bayes factors

(and BIC) can and should be used to guide an iterative model-building process.

THE UNIT INFORMATION PRIOR THAT UNDERLIES BIC

As Weakliem has reminded us, BIC provides a close approximation to the Bayes factor when the prior over the parameters is the unit information prior; that is, a multivariate normal prior with mean at the maximum likelihood estimate and variance equal to the expected information matrix for one observation.[4] This can be thought of as a prior distribution that contains the same amount of information as a single, typical observation.

I would like to review the rationale for this prior and suggest why it might be considered reasonable as a baseline reference prior for Bayes factors. The first question to consider is the difference between Bayesian estimation and hypothesis testing.[5] In Bayesian estimation, the prior often has little effect on the final estimates, and it is common to use a highly spread out prior (e.g., normal with a very large variance); precisely how spread out matters little. Sometimes, even infinitely spread out ("improper") priors are used and can lead to valid estimation results.

So why not do the same for Bayes factors? The reason is that Bayes factors are more sensitive to the prior than are Bayesian parameter estimates. As Weakliem illustrated in his Figure 1, with a normal prior and a single parameter, the Bayes factor for the null hypothesis (e.g., the independence hypothesis in Weakliem's 2×2 table examples) is roughly proportional to the prior standard deviation when the latter is large.[6] As a result, when the prior standard deviation is very large, the Bayes factor always favors the null hypothesis. This does not mean that one should not use Bayes factors; rather, it means that one should not use very spread out priors for this purpose, since very large prior standard deviations imply that the parameter is very large in absolute value, which is unrealistic. It also implies that we have to be more careful when choosing priors for Bayes factors than when choosing priors for Bayesian estimation.

We need a prior distribution that is sufficiently spread out—but not excessively spread out—to cover the parameter values thought

plausible. As Efron (1998) has explained recently, R. A. Fisher made many of the important statistical discoveries of the century by reducing inference problems to a very simple form for which there would be agreement on the answer; his preference was for inference about a normal mean when the standard deviation is known. It turns out that solutions for this model generalize to a very wide class of other statistical models. So let us think about our problem in this context also. Suppose data are from a normal distribution with unknown mean μ and known standard deviation, which without loss of generality we will take to be one. Then, consider testing the null hypothesis $\mu = 0$ against the alternative hypothesis $\mu \neq 0$, and suppose we seek to do this by computing the relevant Bayes factor. To do this, we need a prior distribution for μ.

In this situation, a unit information prior is normal with mean equal to the mean of the data and standard deviation equal to one. An example of data of this form, the likelihood, and the corresponding unit information prior is shown in Figure 1. The unit information prior is well spread out relative to the likelihood and is relatively flat within the part of parameter space where the likelihood is substantial, without being much larger outside it. It thus satisfies the conditions of Edwards, Lindman, and Savage (1963) for us to be in a stable estimation situation; that is, one where inference about μ is relatively insensitive to the prior. In this situation, we can say that the likelihood dominates the prior. The unit information prior usually allows us to be in this (often desirable) situation.

The unit information prior covers the range of the observed data and seems to be reasonable in the following additional sense. Imagine an investigator who knows a little but not too much about the problem at hand. One might expect him or her to have at least an idea in advance of the general range within which the data are likely to lie. The mean will be toward the middle of this range, and so the investigator is likely, at the very least, not to put much prior probability outside the range of the data. The unit information prior approximately coincides with the observed distribution of the data, and so it seems unlikely to be too condensed; it seems that it will at least cover the prior distributions of more knowledgeable investigators. It may well be too spread out, however, especially if specific prior information is available, and

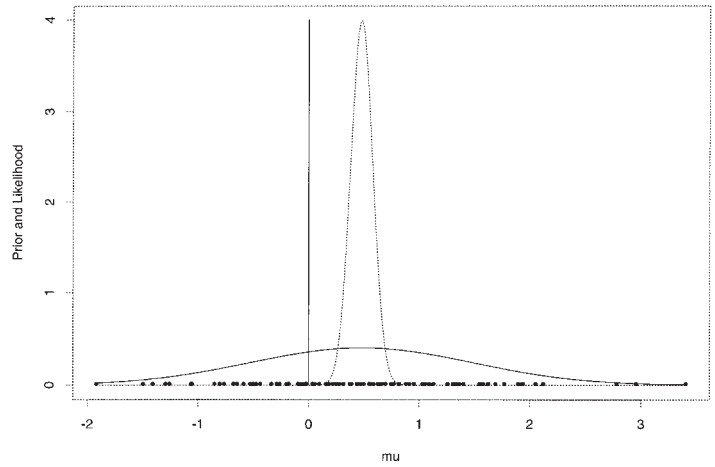

Figure 1: The Unit Information Prior for μ in the $N(\mu,1)$ Example ($n = 100$)

NOTE: The solid curve is the prior density, and the dotted curve is the likelihood. The dots show the data values, and the solid vertical line is at $\mu = 0$. This shows that the unit information prior is locally almost uniform in the region of parameter space favored by the likelihood, while remaining proper.

this is the brunt of Weakliem's article. These observations generalize to a wide range of statistical models.[7]

More spread out priors are conservative in the sense that they tend to favor the null hypothesis more (and hence find less evidence for an "effect" of interest in a study). As a result, it seems plausible that the unit information prior, and BIC, to which it corresponds, will tend to be conservative. Most criticisms of BIC to date have taken the view that it is too conservative; that is, implicitly, that the unit information prior is too spread out. This certainly seems to be Weakliem's view. Cox (1995) argued that the appropriate prior standard deviation will often decline as sample size increases, and hence for large samples the unit information prior is too spread out.[8] Volinsky (1997) has shown via a simulation study in the linear regression context that the performance of Bayes factors can be better than that of BIC if the prior used is less spread out than the unit information prior. Viallefont,

Raftery, and Richardson (1998) have shown that the unit information prior is too spread out substantively for epidemiological case control studies and that better performance results by using a less spread out, substantively motivated prior distribution for the treatment or other effect of interest.

The key thing to note about all these criticisms of BIC and suggestions for better priors is that they inject additional prior information into the problem, which then leads to less spread out priors and more evidence for the alternative hypothesis. Thus, BIC seems to be a conservative solution, which could be routinely reported as a baseline reference analysis, along with results from other priors. If BIC favors an "effect," we can feel that we are on solid ground in asserting that the data provide evidence for its existence. The converse is less clear: If BIC does not favor the effect, there might still be a justifiable prior that would support it. Weakliem has given us two good examples of how such priors can be assessed using imaginary experiments; I suspect that the priors he derives (which are less spread out than the unit information prior) would be widely acceptable to researchers. However, as we will see in these examples, using Weakliem's priors does not lead to qualitatively different conclusions from those obtained with BIC.

If BIC does not support an effect, but a Bayes factor using another prior does, the additional prior information has played a crucial role. This needs to be made very clear when the research is written so that readers can decide whether they agree with the prior used.

PRECISE BAYES FACTORS: THE LAPLACE METHOD, REFERENCE SETS OF PRIORS, AND THE GLIB SOFTWARE

Weakliem and I seem to agree that the best approach is to carefully assess the best prior for the situation at hand and use it to compute the relevant Bayes factors. Ideally, the results should be backed up with a sensitivity analysis to show that the conclusions are not unduly sensitive to the precise prior specification.

Weakliem identifies two difficulties with this; namely, how to assess the priors and how to compute the intractable high-dimensional integrals involved. I believe that these obstacles have been largely overcome, at least for generalized linear models—which include

linear regression, logistic regression, and log-linear models—and for many other model classes. The details of the resulting solutions have been worked out, and software to implement them is freely available.

There are two parts to prior assessment. The first part is to identify a parsimonious but flexible class of (multivariate) priors for the multiparameter situation. For generalized linear models, a multivariate normal prior with all the parameters except the intercept centered at zero and all the covariances not involving the intercept equal to zero seems to be sufficient for many purposes. This involves only one hyperparameter to which the results are sensitive, a scale parameter denoted by ϕ that controls the prior standard deviations of the regression parameters.

The second part of prior assessment is to choose the value(s) of the hyperparameter(s). The best way to do this is to translate one's knowledge about the substantive situation into a prior distribution. Weakliem has given two good examples of ways in which this can be done. If this is not feasible, Raftery (1996) proposed a way of choosing a range of values of ϕ such that, in essence, the prior has a small effect on Bayes factors involving both nested and nonnested generalized linear models. The range is [1,5], with a "recommended" value of 1.65. This is referred to as a reference set of proper priors.

Although direct evaluation of the required integrals over parameter space, given by Weakliem's equation (1), is usually not feasible, the Laplace method gives a very accurate approximation for generalized linear models and at least some other model classes. The Laplace method essentially yields a more accurate version of Weakliem's equation (2), involving the posterior mode rather than the maximum likelihood estimator. The error is asymptotically of order $O(n^{-1})$, much smaller again than the $O(n^{-1/2})$ in Weakliem's equation (2). This method yields Bayes factors that are essentially exact, for all practical purposes.[9]

This entire methodology, specifying the prior family, choosing the range of hyperparameters, and evaluating the Bayes factors, can be carried out using the GLIB software, which is publicly available.[10]

It should be said that in most of the experience to date, the conclusions drawn from this more satisfying but also more complex methodology have not been qualitatively very different from those drawn

TABLE 1: Bayes Factors for Independence in the Anomia/Gender Example

Prior	*–2log (Bayes Factor)*	*Bayes Factor*	*Equivalent Prior Standard Deviation*	*Evidence for an Association*
BIC (unit information)	–3.0	4.6	4.07	None (favors independence)
GLIB: most spread out	–1.6	2.2	1.88	None (favors independence)
GLIB: least spread out	0.5	0.8	0.38	Weak
Weakliem	–0.8	1.5	1.35	None (favors independence)

NOTE: This is a 2×2 table with entries (412, 583, 584, 687), $n = 2{,}266$, $L^2 = 4.68$ ($P = .031$).

from the cruder but much simpler BIC. This turns out to be the case with Weakliem's examples as well.

EXAMPLES REVISITED

THE ANOMIA EXAMPLE

Bayes factors for independence in the anomia/gender example based on various priors are shown in Table 1 in both the raw form and in the form of –2log(Bayes factor), the latter being directly comparable with the BIC.[11] Conventional rules of thumb for interpreting the evidence for an association provided by these Bayes factors are shown in Table 2.[12]

The most striking thing is that the Bayes factors based on all the priors considered agree on the same basic qualitative conclusion: There is little or no evidence for an association between anomia and gender in these data. As we would expect, BIC is the most conservative (i.e., finds the least evidence for an association), but not by much; the difference between BIC and the result based on Weakliem's prior is only about two BIC points. This difference can be attributed mainly to the additional prior information that Weakliem has injected into the analysis (i.e., that in social survey data, odds ratios are usually between 1/20 and 20), which tends to increase the support for the hypothesis of an association slightly. The results from Weakliem's prior fit nicely between the upper and lower bounds given by GLIB. Thus,

TABLE 2: Grades of Evidence for an Association Corresponding to Values of the Bayes Factor in a 2 × 2 Table

–2log(Bayes Factor) or BIC	*Bayes Factor*	*Evidence for an Association*
< 0	> 1	None (favors independence)
0-2	0.37-1.00	Weak
2-6	0.05-0.37	Positive
6-10	0.01-0.05	Strong
> 10	< 0.01	Very strong

Weakliem may have overstated the importance of the differences between the different Bayes factor analyses in this example.

These results contrast with those from a frequentist analysis, for which the *P* value is .031. Conventionally (at the most frequently used 5 percent level), this would be interpreted as meaning that there is evidence for an association, although there is some ambiguity here; frequentist statisticians would sometimes recommend that with such a large sample size (n = 2,266), a more demanding significance level such as 1 percent should be used. If this were done, the result would not be significant. There do not seem to be systematic guidelines for choosing frequentist significance levels.

THE BELIEF IN GOD/RACE EXAMPLE

The results for this example are summarized in Table 3.[13] The basic story is similar to that for the anomia example, in spite of the difference in the margins between the two examples. As before, the different Bayes factors agree that there is little or no evidence for an association, and the BIC is the most conservative, but again not by much. Again, the frequentist test is significant at the 5 percent level but not at the 1 percent level (P = .025).

One interesting point in both these examples is that although the data do not provide much evidence for the hypothesis of an association, they do not provide much evidence against it either; rather, the data are inconclusive (according to the Bayes factors). This is the case in both examples no matter which prior is used. Thus, in future research, one might reasonably continue to study these associations by collecting more data, albeit with only modest expectations of success. One feature of Bayes factors that frequentist significance testing does

TABLE 3: Bayes Factors for Independence in the Belief in God/Race Example

Prior	*–2log (Bayes Factor)*	*Bayes Factor*	*Equivalent Prior Standard Deviation*	*Evidence for an Association*
BIC (unit information)	–2.7	3.8	16.70	None
GLIB: most spread out	–3.0	4.4	19.11	None
GLIB: least spread out	–0.0	1.0	3.82	None
Weakliem	0.3	0.9	1.3	Weak

NOTE: This is a 2 × 2 table with entries (258, 1866, 9, 133), $n = 2,266$, $L^2 = 5.04$ ($P = .025$).

not possess is that they can distinguish between the two different situations where a null hypothesis is not rejected: when there are not enough data (and hence the data are inconclusive), and when the data actually support the null hypothesis. Both of these examples illustrate the first of these two situations.

THE SOCIAL MOBILITY EXAMPLE

In his reanalysis of the social mobility example, Weakliem emphasized the fact that he was able to find two models that fit better that any of those considered by researchers who analyzed these data previously using BIC (including myself).

But is this really an argument against BIC? Summary statistics for seven of the main models that have been considered for these data are shown in Table 4. According to BIC, Weakliem's proposed models are much better than any previously proposed, and his preferred model (number 7) has the best BIC score of all. So BIC and Weakliem seem to agree with one another about the bottom line in this example.

Yet, puzzlingly, in the final paragraph of his section on Sample Size and the BIC, Weakliem tries to argue that the failure of previous researchers to notice the models he so ingeniously discovered implies that BIC is a failure in this example. Many previous researchers have analyzed these data,[14] some using BIC and others not, and none had discovered the better models that Weakliem presented here. This tells us nothing about how good BIC is, and certainly does not imply that BIC fails in this example.

In Raftery (1995, sec. 7), I argued that the model search should continue in situations like this. By this I mean situations in which (1) a

TABLE 4: Fit of Models to Social Mobility Data

Number	*Model*	*Reference*	*Deviance*	*Degrees of Freedom*	*BIC*
1	Independence	GH, Table 5, model 1	42970	64	42227
2	Lipset-Zetterberg	GH, p. 22	18390	120	16997
3	Quasi-symmetry	GH, Table 5, model 2	150	16	–36
4	Saturated	—	0	0	0
5	Explanatory	GH, Table 5, model 4	490	46	–43
6	Uniform asymmetry	Weakliem	49	15	–125
7	Farm inheritance asymmetry	Weakliem	26	14	–137

NOTE: GH = Grusky and Hauser (1984).

Bayes factor prefers a parsimonious model to a more complex one, (2) there is a substantial amount of deviance explained by the complex model but left unexplained by the parsimonious model, and (3) there is more than one degree of freedom in the comparison. When precisely should we keep going; that is, how substantial is "substantial" in the last sentence? One possible rule of thumb is to keep searching when the deviance difference exceeds about log(n), so that there is "room" for one additional parameter to yield a better Bayes factor or BIC. Another possibility, as Weakliem suggests, is to use a P value as a rough rule of thumb in this instance (I would not support using it as the basis for a final conclusion, but as a rough guide to whether to continue an iterative model search, it seems unexceptionable). Model search should also continue when the currently preferred model does not seem parsimonious or interpretable enough; then, the emphasis may be on explaining the same or nearly the same amount of deviance with fewer parameters. Grusky and Hauser (1984) and Hout (1988) give examples of this. In any event, the last paragraph of Weakliem's section on Sample Size and the BIC misrepresents my views on this issue.

Substantively, for social mobility research, it would seem interesting to combine Weakliem's ideas here with those of Grusky and Hauser (1984) by modeling some of the mobility parameters in Weakliem's preferred model as functions of country-specific independent variables. It would also seem worthwhile to apply some of the ideas discussed in this exchange to the more comprehensive, up-to-date, detailed, and comparable collection of social mobility tables developed by Ganzeboom, Luijkx, and Treiman (1989), and expanded since then.

Weakliem's other points about this example have to do with whether a properly calculated Bayes factor would really favor the quasi-independence model over the saturated model. The way to settle this is to actually do the calculation using, for example, the prior that Weakliem suggests (which seems reasonable). This can be done in a fairly straightforward way using GLIB. This would seem more satisfactory than ad hoc adjustments such as $MBIC_1$ and $MBIC_2$, which are not known to correspond to any particular prior and, hence, to any particular Bayes factor.

WHY BIC? WHY BAYES FACTORS?

Weakliem writes that there is no obvious reason to prefer BIC to other criteria such as L^2/df, Akaike's information criterion, or the index of dissimilarity. Where formal inference is concerned, I do not agree, although all these criteria can be useful for the informal assessment of models.

The main reason for using BIC is that it provides an approximation to a Bayes factor, and, as I will argue in a moment, Bayes factors have desirable properties for hypothesis testing and model selection.[15] The prior underlying BIC may often provide a reasonable representation of a situation where there is little prior information. Even if the prior underlying BIC is not the prior one would prefer to use, BIC still is likely to be conservative relative to Bayes factors based on informative priors, and so there is a case for using it as part of routine research reporting as a baseline reference quantity, perhaps in conjunction with other Bayes factors.

So why Bayes factors? Bayes factors provide the Bayesian solution to the question, "What evidence do the data provide for one model against another, competing model?" expressed as a ratio of posterior probabilities. This leads to several desirable properties. The first is that the hypothesis-testing procedure defined by choosing the model with the higher posterior probability minimizes the total error rate; that is, the sum of Type I and Type II error rates (Jeffreys 1961:396-97; Kass 1991). Note that frequentist statisticians sometimes recommend reducing the significance level in tests when the sample size is large; the Bayes factor does this automatically.

The second property is as follows. When there is model uncertainty (i.e., doubt about which is the best model to use), the Bayesian solution to inference about quantities of interest is Bayesian model averaging; that is, averaging the posterior densities of the quantity of interest across the different models, with weights proportional to their posterior probabilities (derived from the Bayes factors). Madigan and Raftery (1994) have shown theoretically that this leads to optimal predictive performance, and this has been verified on real data in a series of studies summarized by Raftery, Madigan, and Volinsky (1995).

NOTES

1. See Raftery (1986, 1993b, 1995).

2. See Madigan and Raftery (1994); Kass and Raftery (1995); Raftery (1996); Madigan, Gavrin, and Raftery (1995); Lewis and Raftery (1997); and Raftery, Madigan, and Hoeting (1997).

3. See Kass and Raftery (1995, sec. 7.3) for another illustration of this.

4. This was shown for nested models under certain conditions by Kass and Wasserman (1995); the result is reviewed in Kass and Raftery (1995). Raftery (1995, sec. 4.1) provided another version of the result that shows that it also applies to integrated likelihoods in general, but under more restrictive conditions. It follows that BIC can also be a good approximation for the comparison of nonnested models.

5. A brief introduction to Bayesian estimation is given in Raftery (1995, sec. 3.1); several good references for further study are recommended there. To these I would add Gelman et al. (1995).

6. A quite general result along these lines can be derived using equation (14) of Kass and Raftery (1995).

7. Essentially, those for which the maximum likelihood estimator is asymptotically normally distributed with variance matrix equal to the inverse of the expected Fisher information matrix.

8. I feel that this may well be true in some settings, but whether it is or not in any particular case is an empirical question that could in principle be assessed empirically. Cox's (1995) rationale is that studies looking for big effects will tend to have small samples, and studies that expect to find small effects will tend to use large samples. Of course, this tendency may be mitigated by a preference for large samples when the effect being studied is important so as to try to put the conclusion beyond controversy; an example of this is provided by the large studies that were carried out on smoking and lung cancer (a large effect). For sociology, the point is even more moot. A great deal of sociology consists of secondary analyses of existing (often large) data sets that were collected for other purposes, and so it is hard to see how in such situations an association between size of effect and sample size could arise.

9. The use of the Laplace method for Bayes factors was first proposed by Jeffreys (1961). The methodology for generalized linear models was described by Raftery (1996), building on Raftery (1988, 1993a). More pedagogical expositions are provided by Kass and Raftery (1995) and Raftery and Richardson (1996).

10. GLIB is an S-PLUS function, which is available from the Statlib archive at http://lib.stat.cmu.edu/S/glib, or from the Bayesian Model Averaging Homepage at http://www.research.att.com/~volinsky/software/glib. The most recent updates will be posted to the Bayesian Model Averaging Homepage at http://www.research.att.com/~volinsky/bma.html, which is maintained by Chris T. Volinsky (volinsky@research.att.com).

11. The result for Weakliem's prior is taken from his article. His analysis has an unsatisfactory aspect because computing a Bayes factor involves specifying a prior for α, β_1, and β_2 in his equation (4), as well as a prior for Θ, and integrating over all four parameters. However, he assigns a point mass prior distribution to α, β_1, and β_2 at their maximum likelihood estimates for ease of computation. This prior is clearly unrealistic. In fact, it is easy to both specify reasonable priors for these parameters and integrate over them using GLIB. However, in this example, how the prior for α, β_1, and β_2 is specified does not make much difference, as Weakliem surmised, so I have just reported his result for ease of comparison.

12. These rules of thumb were adapted by Raftery (1995) from Jeffreys (1961, appendix B). Jeffreys argued that grades of evidence are most appropriately assessed on the logarithmic scale.

13. For this example, Weakliem did not report a Bayes factor based on his preferred prior standard deviation of 1.35 for the log odds ratio, Θ. I have therefore computed it using GLIB with the default prior family, setting the prior scale parameter ϕ in such a way that the prior standard deviation of Θ is 1.35. This is $\phi = 0.35$, somewhat outside the "reference range" of 1-5 derived in Raftery (1996). The prior standard deviation of 1.35 represents real prior information, however, so its being outside a reference range designed for the situation where there is little prior information is not a cause for concern.

14. See Grusky and Hauser (1984), Xie (1992), and the many references therein.

15. Weakliem seems to agree on this point, but it is still worth developing, as it is a crucial one.

REFERENCES

Cox, David R. 1995. "The Relation Between Theory and Application in Statistics [with discussion]." *Test* 4:207-61.

Edwards, Walt, H. Lindman, and Leonard J. Savage. 1963. "Bayesian Statistical Inference for Psychological Research." *Psychological Review* 70:193-242.

Efron, Bradley. 1998. "R. A. Fisher in the 21st Century [with discussion]." *Statistical Science* 13:95-122.

Ganzeboom, Harry B. G., Ruud Luijkx, and Donald J. Treiman. 1989. "Intergenerational Class Mobility in Comparative Perspective." *Research in Social Stratification and Mobility* 8:3-79.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. London: Chapman and Hall.

Grusky, David B. and Robert M. Hauser. 1984. "Comparative Social Mobility Revisited: Models of Convergence and Divergence in 16 Countries." *American Sociological Review* 49:19-38.

Hout, Michael. 1988. "More Universalism, Less Structural Mobility: The American Occupational Structure in the 1980s." *American Journal of Sociology* 93:1358-1400.

Jeffreys, Harold. 1961. *Theory of Probability*, 3rd ed. Oxford: Oxford University Press.

Kass, Robert E. 1991. "About *Theory of Probability*." *Chance* 4:13.

Kass, Robert E. and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90:773-95.

Kass, Robert E. and Larry Wasserman. 1995. "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association* 90:928-34.

Lewis, Steven M. and Adrian E. Raftery. 1997. "Estimating Bayes Factors via Posterior Simulation With the Laplace-Metropolis Estimator." *Journal of the American Statistical Association* 92:648-55.

Madigan, David, Jonathan Gavrin, and Adrian E. Raftery. 1995. "Enhancing the Predictive Performance of Bayesian Graphical Models." *Communications in Statistics—Theory and Methods* 24:2271-92.

Madigan, David and Adrian E. Raftery. 1994. "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window." *Journal of the American Statistical Association* 89:1535-46.

Raftery, Adrian E. 1986. "Choosing Models for Cross-Classifications." *American Sociological Review* 51:145-46.

———. 1988. "Approximate Bayes Factors for Generalized Linear Models." Technical Report No. 121, Department of Statistics, University of Washington.

———. 1993a. "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models." Technical Report No. 255, Department of Statistics, University of Washington. Available at http://www.stat.washington.edu/tech.reports/tr255.ps

———. 1993b. "Bayesian Model Selection in Structural Equation Models." Pp. 163-80 in *Testing Structural Equation Models*, edited by K. A. Bollen and J. S. Long. Newbury Park, CA: Sage.

———. 1995. "Bayesian Model Selection in Social Research [with discussion]." Pp. 111-95 in *Sociological Methodology 1995*, edited by Peter V. Marsden. Cambridge, MA: Blackwell.

———. 1996. "Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models." *Biometrika* 83:251-66.

Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. 1997. "Model Selection and Accounting for Model Uncertainty in Linear Regression Models." *Journal of the American Statistical Association* 92:179-91.

Raftery, Adrian E., David Madigan, and Chris T. Volinsky. 1995. "Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance [with discussion]." Pp. 323-49 in *Bayesian Statistics 5*, edited by J. M. Bernardo et al. Oxford: Oxford University Press.

Raftery, Adrian E. and Sylvia Richardson. 1996. "Model Selection for Generalized Linear Models via GLIB, With Application to Epidemiology." Pp. 321-54 in *Bayesian Biostatistics*, edited by D. A. Berry and D. K. Stangl. New York: Dekker.

Viallefont, Valerie, Adrian E. Raftery, and Sylvia Richardson. 1998. "Bayesian Model Selection in Logistic Regression, and an Epidemiological Context." Technical Report No. 343, Department of Statistics, University of Washington, Seattle.

Volinsky, Chris T. 1997. "Bayesian Model Averaging for Censored Survival Models." Unpublished Ph.D. dissertation, Department of Statistics, University of Washington.

Weakliem, David L. 1999. "A Critique of the Bayesian Information Criterion for Model Selection." *Sociological Methods & Research* 27:359-97.

Xie, Yu. 1992. "The Log-Multiplicative Layer Effect for Comparing Mobility Tables." *American Sociological Review* 57:380-95.

Adrian E. Raftery is professor of statistics and sociology at the University of Washington in Seattle. He is currently coordinating and applications editor of the Journal of the

American Statistical Association, *and he was editor of* Sociological Methodology *from 1996 to 1998. He is a fellow of the American Statistical Association and winner of its Award for Outstanding Statistical Application. In 1998, he was the corecipient of the Population Association of America's Clifford C. Clogg Award for population statistics. He is currently working on Bayesian model selection and Bayesian model averaging in social research, on inference for mechanistic models in whale population dynamics and environmental monitoring, and on family structure and social mobility (with Timothy J. Biblarz).*