model if Swendsen-Wang were not better still. More work will be required before we learn how useful these methods are, but they do seem to be worth investigating.

How Safe Is Markov Chain Monte Carlo? Racine-Poon "remains quite worried" about convergence of Markov chain Monte Carlo, and this seems appropriate. So long as there are many problems in spatial statistics, expert systems and statistical genetics for which no one knows how to construct rapidly mixing samplers, the worries will remain. Even ignoring these areas and sticking to what Raftery and Lewis call "standard statistical models," it is not clear that rapidly mixing samplers can be constructed for all such problems.

If one has a sampler that mixes too slowly, multiple starts and diagnostics cannot save the situation. It is necessary to change the sampling scheme so that it mixes more rapidly. Fortunately, the Metropolis-Hastings algorithm offers an enormous scope for experimentation. Experience shows that for many problems standard schemes such as one-variable-at-a-time Gibbs updating work well. Experience also shows that some very hard problems have been cracked using clever sampling schemes.

Rejoinder: Replication without Contrition

Andrew Gelman and Donald B. Rubin

We thank all the discussants and congratulate the editorial board for providing the readers of *Statistical Science* with multiple independent discussions of our article, which surely provide a better picture of the uncertainty about the distribution of positions on iterative simulation than one longer article by us, even though we might have eventually presented all possible theoretical positions had we been allowed to write ad infinitum. Even so, the readers would have obtained a more accurate impression of what users of iterative simulation actually do in practice had the discussants focused more on this pragmatic topic and less on theoretical advice concerning what others should do; after many public presentations and personal conversations, we know of no one who uses iterative simulation to obtain posterior distributions with real data and eschews multiple sequences, despite possible theoretical contrition at doing so. For a specific example, an anonymous reviewer of one of our research proposals wrote: "The convergence tests he has helped to develop for Gibbs sampling are certainly straightforward to implement. Moreover, the multiple starts upon which they are based appear to me to be essential in practical applications. Nonetheless, they are by no means widely accepted." To help disseminate our ideas, we summarize our recommendations in Table 1.

It is difficult to overstate the importance of replication in applied statistics. Whether dealing with experiments or surveys, the heritage beginning with Fisher (1925) and Neyman (1934) and followed by a host of other contributions and contributors is that, for statistical inference, a point estimate without a reliable assessment of uncertainty is of little scientific value relative to an estimate that includes such an assessment, and the most straightforward path to this objective is to use independent replication. This conclusion is also true in the context of iterative simulation where the estimand itself is a distribution rather than a point. Multiple sequences of an iterative simulation provide replication, whereas a single sequence is analogous to a systematic design. Although systematic designs can produce more precise estimates for equivalent costs and hence be useful especially in pilot investigations (e.g., for exploring efficient stratification schemes) or in very well-studied settings where sources of variability are easily controlled (e.g., some routine laboratory situations), in general scientific practice where variability is not fully understood and valid inferences are critical, systematic designs are far less attractive than those with independent replication. Of course, essentially all relevant statistical inferences are subject to some unassessed uncertainty (e.g., extrapolation into the future), and so "validity" of inference is relative, referring to the substantially larger class of problems successfully handled by replicated rather than systematic designs.

Somewhat surprisingly, many of the discussants' comments suggest an abandonment of this heritage, and some even appear to recommend reversing the accepted practice by using multiple sequences, with their independent replication and consequent superior inferential validity, for a pilot phase, and a systematic

TABLE 1
Summary of advice for inference from iterative simulation using multiple sequences

Steps	State of the art	Usually adequate	Works in simple problems
1. Create a starting distribution	Overdispersed approx. based on modes Importance resampling (Sections 2.1, 4.2 of our article)	Samples from each mode scattered about parameter space (Section 4.8)	Systematic samples at extreme points of parameter space (e.g., Gelman and Rubin, 1991, for Ising model)
2. Simulate multiple sequences	Iterative simulation algorithm tailored to problem at hand Possibly use auxiliary variables (i.e., data augmentation) Metropolis-Hastings Many replications using various algorithms	Metropolis or Gibbs on the natural space of random variables Several replications	Approximations (e.g., Gibbs on approximate conjugate model) Correction via rejection or importance sampling 2 or 3 replications for each mode
3. Monitor convergence	Use potential scale reduction for all estimands of interest, based on simple-sequence and multi-sequence variances (Tables 1-4)	Compare inferences from separate sequences and see how much they overlap	Examine time-series plots of multiple sequences overlaid (Figures 3, 4)
4. Inference about the target distribution	Multiple samples of simulated sequences that are very close to convergence (Sections 4.6, 4.7 last five columns of Table 2)	Distributional estimates that account for simulation variability based on simulations not far from convergence (Sections 3.4-3.6, first three columns of Table 2)	Distributional estimates based on simulations that are not close to convergence (Tables 3, 4)
5. What to do next if current inference is not sharp enough	Several options: • Run simulations longer, perhaps using the time series of potential scale reductions as design information • Restart using draws from a new estimate of the target distribution • Alter the iterative simulation algorithm(s) If potential scale reductions are already near 1, continue the science.		

single sequence without replication for the final inference, almost as if the initial exploration using multiple sequences were so informative that uncertainty was reduced to that of a laboratory setting (e.g., Tierney, and Raftery and Lewis). When actually performing iterative simulations with real data, however, it appears as if at least some of the same discussants actually do use multiple-sequence methods to avoid errant conclusions and obtain valid final inferences [e.g., the presentation at the 1992 ASA meeting of Lange, Carlin and Gelfand (1992) and the two examples in Wakefield et al. (1992)]. Other recent public examples of the use of multiple sequences to avoid possibly invalid conclusions obtained from a single "convergent" sequence include a variety of presentations in July 1992 at the University of Nottingham on Applied Bayesian Statistics (e.g., L. I. Pettit's "Inferences about ordered parameters – an astronomical problem" used two sequences that nicely converged to different answers after 10,000 iterations; C. E. Buck, C. D. Litton and D. A. Stephens' "Detecting changes in the shape of prehistoric corbelled tombs" successfully monitored slow convergence using twenty sequences of 150,000 iterations; C. Ritter's "The analysis of electron spectroscopy data for chemical analysis" used 100 sequences and a week of CPU time on a DEC station to find adequate convergence for many parameters but more uncertainty in the distribution for others). Further examples of multiple sequences working with single sequences exhibiting local convergence appear in McCullogh and Rossi (1992), Gatsonis et al. (1992), Gelman and Rubin (1992), Rubin and Stern (1992), and Wakefield (1992). For all these examples, lack of convergence is immediately apparent from multiple sequences while being difficult or impossible to detect from a single sequence.

Even our example, initially labeled "too simple to

provide a test of methods" by Geyer, illustrates the futility of the single-sequence approach. First, as shown by Table 4 and Figure 4 in our Section 4.8, starting from a minor mode leads to "locally convergent" behavior in the sense that standard single-sequence methods claim convergence even though the sequence has not visited most of the target distribution. Second, Geyer's own single-sequence simulation of our example was incorrect; as stated in his note added in proof: ". . . my simulation of Gelman and Rubin's example was wrong . . . the Gibbs sampler apparently converged when there was no stationary distribution for it to converge to. A run of a million iterations gave no hint of lack of convergence . . . a different starting point might have diagnosed the problem. . . ."

Applied statisticians with experience in designing or analyzing either experiments, observational studies or surveys understand the value of replication for avoiding inappropriate conclusions and for creating valid inferences. It is also clear to applied statisticians that the relative costs of obtaining replications in computer simulations are far less than the analogous costs in scientific investigations and that these relative costs will become even more disparate in the future, especially with the proliferation of parallel computing. We know of no statistician who calls for the total abandonment of replication when collecting real data despite the extra costs involved; so should it be for the collection of data from iterative simulations, where costs tend to be trivial relative to those involved in collecting the real data being modeled by the target distribution of the iterative simulation.

EXTENDED DISCUSSION

At this point, our rejoinder could end, because we feel that we have communicated, at least to applied statisticians, the fundamental benefits of multiple sequences for drawing valid inferences for target distributions. Most of the discussants, however, addressed more theoretical topics, which also warrant our attention to promote closure. We first emphasize some important general points that apply to more than one of the discussions and then respond to those individual discussants' points not adequately addressed previously.

The Relation between Statistical Theory and Practice

Fortunately, underlying much statistical practice, there exists mathematical theory. In particular, a specific mathematical theory based on Bayesian inference underlies our proposals, which comprise a principled approach to the analysis of data from iterative simulation. The theorems state that under certain conditions, our procedure is essentially optimal within a class of procedures and, under more general conditions, still leads to valid statistical inferences. Just as when implementing other mathematical results in statistical practice (e.g., Gauss-Markov and normal theory for the analysis of variance), the fact that the "required" mathematical conditions do not exactly hold in practice does not obviate the utility of the procedure. Deviations from our conditions (e.g., overdispersed starting distributions) do not relegate our methods to irrelevancy, any more than deviations from normality make the analysis of variance worthless. If strict adherence to such regularity conditions were required for statistical theory to be relevant to statistical practice, professional statisticians would be essentially irrelevant to empirical science.

"Iterative Simulation" versus "Markov Chain Monte Carlo"

We specifically chose the more general title "iterative simulation" so as to include all methods that iteratively converge to a target distribution [e.g., non-Markovian iterative importance sampling; Gelman (1992)]. In addition, Markovian methods such as the Metropolis algorithm can lose that property in practice because of various convergence-enhancing schemes such as reparameterization, adaptive sampling rules and importance reweighting. Finally, even if the vector of parameters being simulated forms a Markov chain, subvectors of individual components (or other lower-dimensional estimands) do not in general form Markov chains.

Inference and Diagnostics versus A Priori Analysis

The word "diagnostics," when applied to iterative simulation, typically implies assessing the distance, in some sense, between the estimated distribution at time t and the target distribution (or perhaps between the distribution of the simulations at time t and the target distribution), with the assessment based on the observed simulation results up to that time. We use the word "inference" rather than "diagnostics" throughout to emphasize the unity of estimating the target distribution and estimating the distance from convergence when conducting statistical analysis on the results of the iterative simulation.

Some of the discussants (notably Rosenthal, Polson and, to some extent, Geyer) seem uncomfortable with "diagnostics" and statistical analysis, and apparently would prefer that the assessment of convergence be done a priori and theoretically, before the simulations have begun. Eschewing statistical analysis essentially means drawing inferences about the result of an experiment by ignoring the observed data and instead using only the theory underlying the design of the experiment; there is certainly no theoretical advantage in throwing away information, either theoretical or empir-

ical, when making a decision or forming an estimate, even if the experiment is well designed. A strategy of shunning diagnostics because they might not work seems analogous to driving a car with a blindfold, after studying maps and traffic flows, because the strategy of looking at the road cannot ensure an accident-free journey.

Where's the Theorem?

The goal in Bayesian iterative simulation is inference about a target posterior *distribution*, and so judgments concerning the theoretical and practical efficacy of inferential procedures must be based on the quality of the inferences about this target distribution. Some of the discussants, notably Polson, offer theoretical results on the convergence of point estimates rather than on the convergence of distributional inferences from an iterative simulation as it proceeds from the starting to the target distribution. No one has offered nor even alluded to a theorem implying that single-sequence methods are optimal or even more efficient than multiple-sequence methods for distributional inference, although some have invoked the word "theory" to support this view (e.g., Raftery and Lewis state "Theory suggests that Markov chain Monte Carlo inference be based on a single long run," and Geyer seems to go even further, apparently citing "all the theory" for this claim). In our specific rejoinder to Geyer, we explicate the contrast between the attained objective of the single-sequence theorists and the real objective of iterative simulation in applied statistics.

Perhaps some of the theoretical confusion arises from the incorrect assumption that a multiple-sequence inference must use only the last iterate from each sequence, a practice analogous to performing cluster sampling, then basing an inference on just one unit from each cluster. Any inference created using the iterations from a single sequence can be replicated and improved using the iterations from each of two, or three, or m independent sequences; in fact, every survey practitioner would prefer having m clusters available rather than one. As in cluster sampling, it is straightforward to base inference about a distribution on samples that are identically distributed but not independent by using information about the between-within structure, and there are theorems confirming that there is more information in all the data than just in one cluster (one sequence) or in m observations, one from each cluster (the last draw in each sequence). Our inferences use both within- and between-sequence information.

The Role of a Starting Distribution in Helping to Detect Modeling and Programming Mistakes

In our example, the additional work expended to construct the starting distribution was more than repaid with confidence in our results, including confidence that there were no errors in our iterative simulation code itself. In addition, constructing the starting distribution took substantially less effort than programming the Gibbs sampler; for example, we obtained the ECM modes essentially by running the Gibbs program with all the random draws replaced by closed-form calculations of conditional means (E steps) or modes (CM steps).

In nearly all the applications of iterative simulation we know, establishing the statistical reasonableness of the model and debugging the simulation programs have been more difficult than running and obtaining inferences from the simulations. Certainly, in our own work, poor convergence of iterative simulations (i.e., large potential scale reductions for parameters of interest that do not decline after running the simulations for a long time) has often been caused by our programming or modeling errors. Even experts can make mistakes in programming, as well as, of course, in modeling (e.g., see Geyer's note added in proof), and it pays to design data collections and analyses, including of simulations, to detect such mistakes, not hide them. Having even a rough starting distribution for comparison can be very helpful.

Optimistic and Pessimistic Attitudes toward Iterative Simulation

Especially in his initial discussion, Geyer appears to be an optimist in the sense that he recommends using any "reasonable starting point," simulating "long enough," and then relying on asymptotically valid inferential techniques to estimate everything of interest. Also, there he appears to be an optimist because he is not worried about bugs in his computer code or underlying model. Raftery and Lewis appear nearly as optimistic; they recognize that some starting points can be bad but believe that trial and error is sufficient to find a single good starting point. Like Geyer, they appear convinced that their single sequence will not miss important parts of the target distribution, at least "in standard statistical models." The same sort of optimism supports the call for systematic instead of replicated sample surveys and experiments because they are good enough in "standard settings." Such optimism can easily lead to overconfident and false inferences when collecting real or simulation data, as we have already shown for the Ising model (Gelman and Rubin, 1992) and the random-effects mixture model (Section 4 of our article). As Geyer states in his note added in proof, the single sequence he obtained from his improper version of our model entirely missed all the mass in his posterior distribution, which is all concentrated about $\sigma_a^2 = 0$. Although Geyer discovered his error in modeling through theoretical insight, a practitioner could have discovered it with multiple

sequences with at least one starting near $\sigma_a^2 = 0$ or simply using a preliminary search for multiple modes, as we have recommended.

At the other, pessimistic, extreme are Rosenthal and Polson, who are, in Polson's words, "skeptical about the potential for any empirical diagnostics in the MCMC setting," apparently believing that inference about convergence should be based on theory alone. The obvious problem with this pessimistic position is that iterative simulation is designed for precisely those problems that are too hard to solve theoretically. After all, physicists have usefully studied the Ising model using Metropolis' algorithm [even with multiple sequences, as in Ehrman, Fosdick and Handscomb (1960)] for decades before the Swendsen and Wang (1987) algorithm appeared.

Most of the other discussants lie somewhere between; Racine-Poon may be typical of the applied Bayesian community in being willing to use iterative simulation when simpler methods fail but without completely trusting the results. Lewis (1992) and Taylor and Segal (1992) show similar concerns. Unlike theoretical optimists or pessimists, typical applied statisticians are concerned not only about whether a perfectly coded program will work for a properly formulated model but also about the effects that errors—in mathematics or coding—can have on complicated iterative simulation programs. Multiple independent overdispersed sequences address the concerns of practical users about detecting poor convergence—in Tierney's words, they "aid in detecting problems with the simulation."

INDIVIDUAL RESPONSES

Geyer

We appreciate the extensive comments offered by Geyer and especially value his note added in proof, which we hope will bring our positions to convergence. Though we agree with many of his specific comments, we disagree strongly with his general plan for the use of single-sequence methods and are puzzled by some of his adversarial comments in his initial discussion. For example, we agree with this comment near the end of his Section 1 that importance sampling and Markov chain methods are complementary; this is why Sections 2.1 and 4.2 of our article discuss the point. But then why does he label our proper dismissal of the aberrant sequence of Figure 4, using importance resampling, as an "inadequate justification"? Also, we cannot help but agree with his tautology that "there can be no valid inference from runs that are too short and that if runs are long enough, one run suffices," but what is added by such a statement? Not tautological but quite obviously true are the facts that (1) multiple sequences can reveal that a simulation is "too short" in problems where a single sequence can mislead, and (2) for many applications, m sequences, each of length n, can be "long enough," with one run of length n (or even mn) being "too short." Geyer's note added in proof provides an explicit example of this fact.

Despite Geyer's error, he assures us that his simulation is essentially correct for our distribution; henceforth, we will regard his final simulation as error-free for our model, which is a plausible assumption since his resulting inferences are in agreement with ours. The reason that his simulation, under a model with an improper posterior distribution, can accurately represent a simulation of our model with its proper posterior distribution, is that the two models differ only in their prior distributions for $(\sigma_{obs}^2, \sigma_a^2)$, and if the singularity around $\sigma_a^2 = 0$ is excluded, the prior distributions are nearly identical in the regions of parameter space where the likelihood is substantial, thereby implying nearly identical simulated posterior distributions if no draws are made from near $\sigma_a^2 = 0$. A brief review of his analysis of our example, presented in his Section 3.4, nevertheless dramatizes differences between our goals with iterative simulation.

Geyer has reanalyzed our model and data but does not provide a complete description of his procedure. In particular, where did he start his simulation run? Since his answers agreed with ours, he must not have started near $\sigma^2 = 0$ or near any of the minor modes for our model; had he started near the mode corresponding to the dotted graph on our Figure 4, his time-series methods would have declared convergence, despite being wildly incorrect. He further notes that "the samples seem so close to multivariate normality that Markov chain Monte Carlo does not seem necessary." This is nearly true for our model, but only in hindsight, and not if the single-sequence start were poorly chosen with the sequence locally convergent near a minor mode. Gibbs sampling was done for the original applied problem precisely because the adequacy of the normal approximation was not trusted for this complicated mixture model, especially with the small samples involved. To say that iterative simulation was not necessary because it didn't change our answer much is analogous to saying that getting the brakes tested on my car was a waste of money because they were defective only in the rain, and it did not rain yesterday.

To see the difference between our goals and Geyer's, note that the typical results of our method, displayed in our Table 2, provide interval estimates for the parameters (α, β, τ, etc.) in the target distribution, whereas his analogous Table 2 provides only point estimates (in this case, posterior means) of the parameters. As $n \to \infty$, our estimate approaches the target distribution, whereas his approaches a point. Although his emphasis is understandable in the context of his research on maximum likelihood estimation (Geyer, 1992), it is in-

complete for any applied statistics problem. An applied researcher wants to know, given this model and these data, that τ, the shift parameter for schizophrenics' delayed responses, is, for example, 95% likely to be in the range [0.74, 0.96], as provided by our Table 2, not simply that its posterior mean is in the range [0.845, 0.848], as provided by Geyer's Table 2.

To summarize the competing analyses of our example: We created an overdispersed starting distribution, sampled ten starting points by importance resampling (SIR) and ran ten sequences in parallel for 200 steps, thereby automatically obtaining distributional inferences for all parameters of interest, along with a good deal of confidence that further simulations would not appreciably change our inference. Geyer chose one starting point in an unknown fashion, simulated 10,000 steps, examined some covariance functions (which required looking at graphs) and obtained point estimates and six different variance estimates for each posterior mean examined; he used one variance estimate and discarded the other five (it is not clear why they were computed) and still obtained inference only about the posterior means, not the distribution of the parameters themselves. Geyer concluded in his initial discussion that his inferences were fine on the basis of his Figure 1, not realizing that had his starting point been different, he could have obtained a vastly different answer, and not recognizing that his answer, consisting of posterior means, is statistically deficient for the scientific problem being addressed by the iterative simulation. Moreover, our comparisons with starting distributions lead us to believe in the correct coding of our simulation algorithm—did Geyer's misplaced confidence in his first simulation emanate from some feature of his single-sequence simulation or from the comparison with our multiple-sequence results? Furthermore, the same question may be asked of his current confidence in his simulation of the correct model.

Finally, in this practical example, multiple sequences are far more "efficient" than Geyer's single sequence: our 10×200 iterations yielded an immediate summary of all our substantive inferences (Table 2) with no additional thought—we just plugged the simulation results into our little S program—whereas Geyer initially simulated five or fifty times as many iterations (depending on whether parallel computing is available), had to examine correlation graphs and relied on asymptotic variance estimates. Moreover, he subsequently simulated a million iterations without a "hint of lack of convergence," even though "there was no stationary distribution for it to converge to." The evidence here supports the contention that, in practice, multiple-sequence methods yield useful inferences in less time and with less effort than single-sequence methods.

Racine-Poon

We believe that most of the practical concerns raised in this discussion are addressed in Table 1 and in our earlier comments here. In particular, the effort required to create an approximately overdispersed starting distribution (1) is typically less than required for successfully implementing rejection or importance sampling; (2) can often be reduced by adapting the routines used for iterative simulation, replacing simulation by maximization and conditional expectation to yield ECM; and (3) is useful for debugging. We share Racine-Poon's skepticism about Geyer's variance estimates (which fail in the Ising model and in our example) but are confident that following our prescription in her problems will lead to valid inferences about her target estimands, with high potential scale reductions when the simulations are far from convergence. In addition, much effort that would otherwise have been expended looking at sample sequences, correlation plots, spectra and so on is avoided by simply producing tables analogous to our Table 2, with a row for each estimand of interest, and stopping when inferences are sufficiently precise for practical purposes, as indicated by the potential scale reductions.

Cui et al.

This discussion, along with the work of Kong, Liu and Wong (1991), Gelman et al. (1992), Liu and Liu (1992) and Roberts (1992), shows that importance ratios have the potential for aiding the monitoring of iterative simulation. Under Markov chain simulations, the basic idea is that, if a sequence is started by sampling from the stationary distribution, it should look just as "likely" backward as forward; a related idea appears in Besag and Clifford (1989). Before stationarity, under appropriately strong conditions, a sort of second law of thermodynamics should hold, with the forward path of any chain looking more likely than the reverse path, just as a dropped dish often breaks into pieces, but the pieces of a broken dish are unlikely to fall into place.

Of course, as Cui et al. point out themselves, their statistic can be computed for several parallel sequences and monitored automatically using our methods, thereby eliminating the need for looking at graphs. Furthermore, the theoretical analogy to statistical mechanics suggests that their method would be improved by generalizing it to monitor the distribution of several parallel sequences at once. We would not trust their convergence diagnostic when based on only a single sequence, since it would have no hope of detecting that a simulation is locally convergent near a minor mode, as in our Figure 4 or as in Geyer's original, improper model.

The witch's hat example is brought up by several other discussants, as well as Cui et al., and is, as Racine-Poon points out, a good example of the pitfalls of single-sequence estimation. The maximum of the distribution can easily be found by stepwise ascent, and one can then start multiple sequences with some starting at the mode and several scattered about the sparser parts of the distribution. Slow convergence of an iterative simulation will immediately be revealed by high potential scale reductions.

Gelfand

The discussion extends the "Rao-Blackwellization" idea of Gelfand and Smith (1990) with more sophisticated ways of estimating the target distribution, given some samples (presumably obtained by an iterative simulation that is close to convergence) and some knowledge of the target density function. Not surprisingly, we suggest that the method be applied to independent sequences simulated from overdispersed starting points, thus combining the advantages of multiple sequences with the additional information used in Gelfand's monitoring statistics.

Madras

We completely agree with this discussion, especially the first four sentences. We are pleased to learn about dimerization—yet another application for iterative simulation. In this situation, we suggest simulating multiple samples from the exact distribution, so that valid inference may be obtained from iterative simulations of any length, even lengths that would be "too short" for single-sequence methods. Deciding how many iid variates to simulate at the beginning is really a problem of experimental design—determining if the additional precision afforded by a new simulation is worth the cost.

Schmeiser

Most of Schmeiser's comments are already addressed, especially by our response to Geyer. We wish to point out, however, that Schmeiser may fundamentally agree with us about the utility of multiple sequences, because in problems of Bayesian simulation, it is typically easy (i.e., requiring much less effort than the iterative simulation itself) to obtain overdispersed starting points, which are effectively, in Schmeiser's language, "stratified or antithetic initial states."

Rosenthal

We agree with Rosenthal that iterative simulation "can be used with greater confidence if it is more automated and requires less 'poking around'"; in fact, this concern was a primary motivation for our effort. More specifically, our Table 2, which summarizes the inference for our example, was automatically created by our little S program; no "poking around" was required, except for the creation of the starting distribution, which, as we discussed previously, was certainly worth the effort. Rosenthal may "share Geyer's concern about the difficulty of obtaining useful starting distributions in the first place," but even a single sequence must be started somewhere, and finding a realistic approximate starting distribution is made no easier by restricting oneself to Dirac delta functions.

Polson

Rosenthal and Polson both refer to a vigorous and expanding theory of Markov chain simulation that has the promise of leading us to more effective simulation algorithms and expanding the class of distributions that we can simulate in practice. We agree that this is an exciting area for research that promises to create many useful suggestions for practical application.

Although important for the design of simulation algorithms, Polson's arguments, like theoretical convergence results in general, have an asymptotic emphasis that can be misleading if applied in practice to the problem of inference. For example, in his second paragraph, the phrase "approximate . . . to any specified level of accuracy" is too strong a condition in many practical examples (e.g., distributions with minor modes). In the example in our Section 4, multiple sequences and "diagnostics" (or a lucky starting value) greatly facilitate valid inference about our estimands, and, although Polson's condition (1) is not strictly satisfied (because of the minor modes with negligible, but nonzero, probability), we are confident that we estimated the target distribution to the specified degree of accuracy. For our example, the stated asymptotic result would only apply if we were interested in probability masses in the target distribution smaller than one in a billion, based on the relative masses of the major and minor modes.

Raftery and Lewis

Raftery and Lewis propose to run a single sequence until their specific criterion (possibly based on a shorter pilot simulation) says to stop, then use the iterates after a specified "burn-in" time for distributional inference. Unfortunately, this approach fails in many real examples, including the one in our Section 4 and too many of the other examples in the recent literature previously referenced. For example, notwithstanding the claim in their discussion, their computer program falsely diagnoses convergence when applied a single Gibbs sampler sequence for the Ising model. We applied their computer program separately to the two

Gibbs sampler sequences, each of length 2,000, from Gelman and Rubin (1992) and obtained the recommendation that either 1,148 or 995 iterations is sufficient, following a "burn-in" period of either seven or five iterations, to have a 99% chance of estimating the 95% quantile of the distribution to within an accuracy of 2%. A glance at Figures 1 and 3 from Gelman and Rubin (1992) shows immediately that the two sequences, which have widely dispersed starting points, are still far apart after 2,000 iterations, and there is no possibility that they can both yield accurate estimates of the 95th percentile point of the target distribution. It is also obvious from the two sequences considered together that at least one of the sequences requires a "burn-in" period alone of more than 2,000 iterations. Furthermore, restarting and using single-sequence methods cannot be as informative as comparing multiple sequences; at best, restarting leaves the analyst with some sequences that individually look fine, but, as in our Figure 4, a between-series comparison is still necessary to tell whether the sequences are in agreement.

Although Raftery and Lewis may have obtained acceptable results for their own problems (possibly aided by multiple sequences as in their own Figures 1 and 2), we distrust their general prescription, because it gives wrong answers in many problems—a high price to pay for the largely hypothetical convenience of simulating only one sequence. Perhaps their approach may prove more useful when combined with multiple sequences, as they suggest themselves in Raftery and Lewis (1992).

Tierney

At the end of our Section 2.4, we point out that more accurate within-sequence estimates than ours can be constructed using time-series information, and we are pleased that Tierney also notes this possibility; also see Gelman et al. (1992). However, Tierney's claim that "the resulting variance estimates [from multiple sequences] are very inefficient unless the number of chains is quite large" is incorrect. In our example, relatively efficient variance estimates are obtained using ten independent sequences, with far more than nine "effective degrees of freedom." The variance estimate $\hat{\sigma}^2$ uses both B and W, not just B alone, and the precision of $\hat{\sigma}^2$ is what the Satterthwaite (1946) approximation for the effective degrees of freedom (from our Equation 4) addresses. If each of m sequences has, say, 100 effective degrees of freedom, and the sequences are close to convergence, then the estimate $\hat{\sigma}^2$ based on m sequences together will have nearly $100m + (m - 1)$ degrees of freedom. If the m sequences are far from convergence, then $\hat{\sigma}^2$ will have little more than m degrees of freedom, but, in this case, $W << \hat{\sigma}^2$, and the single-sequence inferences are all falsely precise anyway.

ACKNOWLEDGMENT

We thank Tom Belin for useful comments.

ADDITIONAL REFERENCES

Aldous, D. (1987). On the Markov chain simulation method for uniform combinatorial distributions and simulation annealing. *Probability and Engineering Information Science* 1 33–46.

Belisle, C. J. P., Romeijn, H. E. and Smith, R. L. (1992). Hit-and-run algorithms for generating multivariate distributions. *Math. Oper. Res.* To appear.

Besag, J. E. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika* 76 633–642.

Buck, C. E., Litton, C. D. and Stephens, D. A. (1992). Inferences about ordered parameters—an astronomical problem. *The Statistician.* To appear.

Damerdji, H. (1991). Strong consistency and other properties of the spectral variance estimator. *Management Sci.* 37 1424–1440.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* Springer, New York.

Diaconis, P. (1988). *Group Representations in Probability and Statistics.* IMS, Hayward, Calif.

Diaconis, P. and Hanlon, P. (1992). Eigen analysis for some examples of the Metropolis algorithm. Technical report, Dept. Mathematics, Harvard Univ.

Ehrman, J. R., Fosdick, L. D. and Handscomb, D. C. (1960). Computation of order parameters in an Ising lattice by the Monte Carlo method. *J. Math. Physics* 1 547–558.

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh.

Fishman, G. S. and Moore, L. R. III (1986). An exhaustive analysis of multiplicative congruential random number generators with modulus $2^{31} - 1$. *SIAM J. Sci. Statist. Comput.* 7 24–45.

Gatsonis, C., Liu, C., Morris, C. and Normand, S. L. (1992). Hierarchical Bayes models for medical services utilization data: Bayesian methods for hierarchical logistic and normal models. Technical report, Dept. Health Care Policy, Harvard Univ.

Gelman, A., Liu, C., Liu, J. and Rubin, D. B. (1992). New methods of inference from iterative simulation using multiple sequences, Technical report, Dept. Statistics, Univ. California, Berkeley.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57 1317–1339.

Glasserman, P. (1991). *Gradient Estimation via Perturbation Analysis.* Kluwer, Boston.

Glynn, P. (1987) Limit theorems for the method of replication. *Commun. Statist. Stochastic Models* 3 343–355.

Glynn, P. and Heidelberger, P. (1992). Analysis of initial transient deletion for replicated steady-state simulations. *Oper. Res. Lett.* 11. To appear.

Goldsman, D., Meketon, M. and Schruben, L. (1990). Properties of standardized time series weighted area variance estimators. *Management Sci.* 36 393–397.

Goldsman, D. and Schruben, L. (1990). New confidence interval estimators using standardized time series. *Management Sci.* 36 393–397.

Jerrum, M. and Sinclair, A. (1990). Polynomial-time approximation algorithms for the Ising model. Technical report, Univ. Edinburgh.

KELTON, W. D. (1989). Random initialization methods in simulation. *IIE Transactions* 21 355-367.

KIRKPATRICK, S., GELATT, C. D., JR. and VECCHI, M. P. (1983). Optimization by simulated annealing. *Science* 220 671-680.

KONG, A., LIU, J. and WONG, W. H. (1991). Sequential imputations and Bayesian missing data problems. Technical Report 321, Dept. Statistics, Univ. Chicago.

LAL, M. (1969). "Monte Carlo" computer simulations of chain molecules. I. *Molecular Phys.* 17 57-64.

LANGE, N., CARLIN, B. P. and GELFAND, A. E. (1992). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbrs (with discussion). *J. Amer. Statist. Assoc.* 87 615-632.

LEE, T.-M. and GELFAND, A. E. (1992). On convergence diagnosis and acceleration of Markov chain Monte Carlo algorithms. Technical report, Dept. Statistics, Univ. Connecticut.

LEWIS, S. M. (1992). Contribution to the discussion of three papers on Gibbs sampling and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B.* To appear.

LIU, C. and LIU, J. (1992). Comment on "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods," by A. F. M. Smith and G. O. Roberts. *J. Roy. Statist. Soc. Ser. B* 55. To appear.

LIU, J., WONG, W. H. and KONG, A. (1991a). Correlation structure and covergence rate of the Gibbs sampler (I): Applications to the comparison of estimators and augmentation schemes. Technical report, Dept. Statistics, Univ. Chicago.

LOUIS, T. A. (1992). Comment on "Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers," by N. Lange, B. P. Carlin and A. E. Gelfand. *J. Amer. Statist. Assoc.* 87 626-628.

MADRAS, N. and SOKAL, A. D. (1988). The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk. *J. Statist. Phys.* 50 109-186.

MATTHEWS, P. (1991). A slowly mixing Markov chain with implications for Gibbs sampling. Technical report, Dept. Mathematics and Statistics, Univ. Maryland.

MEKETON, M. S. and SCHMEISER, B. W. (1984). Overlapping batch means: Something for nothing? In *1984 Winter Simulation Conference Proceedings* (S. Sheppard, U. Pooch, and D. Pegden, eds.) 227-230

MÜLLER, P. (1992). A generic approach to posterior integration and Gibbs sampling. *J. Amer. Statist. Assoc.* To appear.

NELSON, B. L. (1989). Batch size effects on the efficiency of control variates in simulation. *European J. Oper. Res.* 43 184-196.

NEYMAN, J. (1934). On the different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc.* 97 558-606.

PETTIT, L. I. (1992). Inferences about ordered parameters—an astronomical problem. Presented at the Nottingham meeting on practical Bayesian statistics.

RAFTERY, A. E., LEWIS, S. M. and AGHAJANIAN, A. (1992). Event history modeling of World Fertility Survey data, with application to the fertility decline in Iran. Unpublished manuscript, Center for Studies in Demography and Ecology, Univ. Washington.

RITTER, C. (1992). The analysis of electron spectroscopy data for chemical analysis. Presented at the Nottingham meeting on practical Bayesian statistics.

RITTER, C. and TANNER, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy Gibbs sampler. *J. Amer. Statist. Assoc.* To appear.

ROBERTS, G. O. (1992). Convergence diagnostics for the Gibbs sampler. In *Bayesian Statistics 4* (J. M. Bernardo et al., eds.) 775-782. Oxford Univ. Press.

ROSENTHAL, J. S. (1991a). Rates of convergence for data augmentation on finite sample spaces. Technical report, Dept. Mathematics, Harvard Univ.

RUBIN, D.B. and STERN, H. S. (1992). Testing in latent class models using a posterior predictive check distribution. Technical report, Dept. Statistics, Harvard Univ.

SCHERVISH, M. J. and CARLIN, B. P. (1990). On the convergence of successive substitution sampling. Technical report, Dept. Statistics, Carnegie Mellon Univ.

SCHMEISER, B. (1990). Simulation experiments. In *Handbooks in Operations Research and Management Science, Vol. 2: Stochastic Models* (D. P. Heyman and M. J. Sobel, eds.) 295-330. North-Holland, Amsterdam.

SCHMEISER, B., AVRAMIDIS, T. and HASHEM, S. (1990). Overlapping batch statistics. In *1990 Winter Simulation Conference Proceedings* (O. Balci, R. P. Sadowski and R. E. Nance eds.) 395-398.

SCHRUBEN, L. (1981). Control of initialization bias in multivariate simulation response. *Comm. ACM* 24 246-252.

SCHRUBEN, L. (1982). Detecting initialization bias in simulation output. *Oper. Res.* 30 569-590.

SHEINER, L. B. and BEAL, S. L. (1980). Evaluation of methods of estimating pharmacokinetic parameters I. Machalis-Menten model: Routine clinical pharmacokinetics data. *Journal of Pharmacokinetics and Biopharmacology* 8 533-571.

SINCLAIR, A. and JERRUM, M. (1989). Approximate counting, uniform generation and rapidly mixing Markov chains. *Inform. and Comput.* 82 93-133.

SMITH, R. L. (1984). Efficient Monte Carlo procedures for generating random feasible points uniformly over bounded regions. *Oper. Res.* 32 1296-1308.

SPIEGELHALTER, D. J. (1988). Fast algorithms for probablistic reasoning, with applications in genetics and expert systems. In *Proceedings of the Conference on Influence Diagrams*, Berkeley, Calif.

TAYLOR, J. M. G. and SEGAL, M. R. (1992). Comment on "Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers," by N. Lange, B. P. Carlin and A. E. Gelfand. *J. Amer. Statist. Assoc.* 87 628-631.

WAKEFIELD, J. (1992). The prediction of concentration levels in population pharmacokinetic studies. Presented at the Nottingham meeting on practical Bayesian statistics.

WAKEFIELD, J., SMITH, A. F. M., RACINE-POON, A. and GELFAND, A. E. (1992). Bayesian analysis of linear and nonlinear population models using the Gibbs sampler. Technical report, Dept. Mathematics, Imperial College, London.

WAKEFIELD, J. (1992). The Bayesian analysis of pharmacokinetic models. Ph.D. dissertation, Univ. Nottingham.

WEI, G. C. G. and TANNER, M. A. (1990). Posterior computations with censored regression data. *J. Amer. Statist. Assoc.* 85 829-839.

WHITT, W. (1990). The efficiency of one long run versus independent realizations in steady state simulation. *Management Sci.* 37 645-666.