Rejoinder

Charles J. Geyer

All of the discussants are to be thanked for their insightful comments. Two major points and a few minor points need some clarification.

What Use Is the Central Limit Theorem (CLT)? Both Polson and Raftery and Lewis criticize the use of variance calculations as a "convergence diagnostic." I agree. I had no intention of suggesting any such thing. My point was rather different. If one has a Markov chain Monte Carlo scheme that mixes rapidly enough so that estimates of means are approximately right, then the variance estimates will also be approximately right and hence will provide a useful estimate of the Monte Carlo error.

Theoretical calculations cannot replace empirical variance calculations, because the bounds from theoretical calculation are (so far at least) very conservative. They may overstate the error by many orders of magnitude. The only useful way to compare different sampling schemes or to estimate the actual Monte Carlo error seems to be to estimate the variance using time-series methods.

How Important Is Burn-In? Many of the discussants objected to my discussion of burn-in (warm-up, initialization) in Section 3.7. There was a point there, but perhaps it was not well made. Both Schmeiser and Polson make the point that "one long run" minimizes the influence of the starting point. It should be added that any point that might reasonably occur in the sample will do as a starting point. It is not necessary to be near the mode. It is only necessary that the starting point be not so far out in the tails as to be extremely improbable considering the sample size.

Thus it is typically not difficult to find a reasonable starting point. In Bayesian problems a crude approximation to the posterior mode will often do, and in frequentist problems the observed data (with missing data, if any, imputed in any reasonable fashion) will often do. Even in very hard problems a reasonable starting point can usually be found by simulated annealing, as suggested by Applegate, Kannan and Polson (1990). It is not necessary that the starting point be approximately from the stationary distribution. The ergodic theorem and the CLT hold *conditionally* on the starting point.

Even if one could get ideal burn-in, that is, starting from a realization from the stationary distribution, a sampler that mixes too slowly is useless for Monte Carlo. Thus it seems that the Gibbs stopper is aimed at the wrong problem. It tells where to start but not how long the run should be.

That having been said, I should concede that the wording in Section 3.7 of the paper is too strong. It does describe my experience with Markov chain Monte Carlo, but there may be problems in which "reasonable" starting points are hard to find.

Variance Estimation. Schmeiser points out the advantage of batch means that it avoids calculation of the autocovariances. While this is true, it is not always an advantage. The autocovariances provide a useful guide to selecting the spacing of samples (Section 3.6) and provide a better estimate of how much longer one needs to run when the run is too short. Both methods have their uses.

Raftery and Lewis assert that I recommend a truncated periodogram spectral estimator, but I did not. The new estimators in Section 3.3 are adaptive window estimators, so criticisms of fixed bandwidth estimators are not applicable. Moreover, only the initial positive sequence estimator uses sharp truncation. The other two use "windows" whose shape is adaptive. These methods were devised so that variance estimation could be more automatic (as Rosenthal requests). In any case the main point was to use the known properties (positivity, monotonicity, convexity) of the autocorrelation structure to improve the estimation. If the shape of the window were critical, the methods could be modified to use a better shape.

Operations Research Literature. Schmeiser points out that Markov chain simulation studies have a long history in the Operations Research literature and that most of the questions just now being addressed by statisticians have been well studied by operations researchers. I agree, and I hope that his comments and mine will lead to more interest among statisticians in the operations research literature on this subject.

Coupled Chains. Madras points out that my proposal for exploring a family of sampling schemes comes with no guarantees, which must be conceded. There do not seem to be any guarantees in this field. Even the theoretical bounds do not yet seem applicable to practical problems. However, what Madras takes as a counterexample is actually an example where the suggested method works. It is fairly easy to construct a sampler for the Ising model using coupled chains, and there is no problem seeing where the critical value is since the distribution changes so rapidly there. When the method of Metropolis-coupled chains (Geyer, 1991a) is used, the acceptance rate for swaps provides a built-in diagnostic of when the temperature gaps are too large. This would be a useful method for the Ising