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