REJOINDER

by Donald Rubin

I'd like to begin by thanking all the discussants for their interesting comments touching on a variety of issues. Since there appears to be only minor overlap in issues among the discussants, I'll respond discussant by discussant in an order that tries to create a logical flow.

Howard Wainer.--Doctor Wainer concurs with me in noting the important fact that in general there is no way to know precisely how a nonrespondent would have responded, and I think he also concurs with me in drawing the conclusion that any method for handling nonresponse therefore must implicitly or explicitly rely on a model for the nonresponse mechanism. Surprisingly, however, he appears to view the displaying of this sensitivity to models for nonresponse as a shortcoming of my proposals, and in his concluding sentence seems to look forward to a method that will make this sensitivity disappear. I feel that since the sensitivity to models is an essential feature of the problem of nonresponse, any appropriate method for handling the problem must display this sensitivity. I also feel that in many survey contexts, some reasonable (not absolute) checks on models are possible, for example, via administrative records or followup surveys. Presumably, by carefully using such data and the results of sequentially redesigned surveys, we will be able to eliminate clearly inappropriate models and narrow our attention to relatively restricted models, such as the "WRMS" of Dempster's comments.

David Hinkley.--Professor Hinkley points out the possible need to have relative weights for each of the plausible nonresponse models so that we can average their answers. Usually, I'd like to avoid averaging over reasonable nonresponse models unless (a) the weights--the posterior probabilities of the models--are largely determined by data rather than by prior probabilities assigned to the models (rarely the case with nonresponse), or (b) the answers under the models do not differ in their practical implications (in which case the averaging is rather irrelevant), or more generally (c) the combination of weights and answers is such that the average answer is insensitive to the prior probabilities on the models. When an average answer is sensitive to prior probabilities on reasonable models, then one answer is not a good summary of the relevant evidence in the data. Hinkley notes in his second paragraph that the need to display sensitivity to reasonable models is an important (and I feel rather neglected) theme in statistics. Finally, I am encouraged by the increasing interest in the development of nonignorable nonresponse models (nonrandom missingness models in the terminology of Rubin, 1976); incidently, another example is Stene's work, to appear in JASA.

Charles Patrick.--Doctor Patrick raises three points of disagreement. I don't understand the second and third points except if read to say that when faced with nonresponse, there exists sensitivity of answers to nonresponse models, and this sensitivity cannot be resolved by the data at hand; but then these are points of agreement! Whether the first point is one of disagreement depends upon the way we define "estimation" and "imputation" (I read these as short for "estimation under one model" and "multiple imputation under one model"). To me, multiple imputations simulate the predictive distribution of an unknown, and estimation may or may not refer to this distribution; if it does, it is logically equivalent to multiple imputation. However, when estimation refers to a best (point) estimate, a loss function is needed to define best, and then the estimate is a summary of some feature of the predictive distribution having an objective that is logically different from the objective of imputation. For example, suppose under a model a missing item is 1 with probability .6 and 0 with probability .4. Multiple imputation would simulate this distribution with random draws of 1's and 0's where p=.6. Using the second sense of estimation, however, we would estimate the missing value to be .6 if the loss is the squared error, 1 if the loss is the number of wrong guesses, and so forth. In this sense, we estimate only when we need one best guess; more generally our interest is in the entire predictive distribution which can be simulated by multiple imputations. Usually in surveys, the actual quantities (e.g., means) to be "point-estimated" are aggregates over individuals.

Estimating individual missing items by their best values and calculating the aggregate does not generally yield a best estimate of the aggregate, even when both "best's" refer to the same loss function. Consider estimating the variance for the entire sample under squared error; inserting the best estimate for missing items--the mean--produces an estimate of the sample variance that is always too small. Multiple imputations allow us to simulate the distribution of aggregated quantities so that appropriate point estimates can in principle be calculated.

Innis Sande.--Innis Sande has addressed the important issue of the practicality of my proposals. It would be unrealistic to be unconcerned about this. For fairly modest data sets, the plan is certainly feasible, at least for some classes of nonresponse models. Rubin (1977, JASA) illustrates many of my proposals with real data from 660 schools. In general, there are barriers to the multiple imputation approach; nonetheless I do not feel when designing a system to be used for many years to come that it is wise to be bound by current constraints on computing (especially considering the increasingly rapid development of

hardware breakthroughs) or by current constraints on the richness of statistical tools for good data analysis with large data sets. For current use, a modest version of my proposals may work well; for example, by sampling units and choosing important variables for multiple imputation we may learn much about the potential problems created by nonresponse in a large data set (of say 50,000 records).

I find myself uncomfortable with the suggestion that the client should get one data set with filled-in values because that's what he wants. I do not feel that there are many real surveys in which we know the consequence of pretending we know the missing values. Hence, I do not think it is practical (that is, implementable and sensible) to label an imputed data set as "clean" and ship it off to users who will act as if the imputed values have no uncertainty and perhaps draw very incorrect conclusions having serious real world implications. I am sympathetic with the desire of data producers to produce data sets that users can use, and that sympathy is the primary reason for turning to multiple imputations instead of insisting that each user of the data set be prepared to perform a full Bayesian analysis of the nonresponse problem. I think that it is currently reasonable and practical in many cases to expect a user to perform the same standard complete data analysis several times, and at least examine summaries of the variability of the answers within an imputation model (e.g., perhaps by means of the variance of the answers) and across imputation models (e.g., perhaps by minimums and maximums).

Arthur Dempster.--Professor Dempster and I agree on the need to be Bayesian in nonstandard situations, the need for repeated analysis using a variety of reasonable models, and the need to formalize classes of nonresponse models that can be useful in practice. However, we may disagree on the importance of the multiple-imputed data set scheme.

If the data producer, and data analyzer are in close contact, there may be no advantage in creating multiple imputations. But I think that there are many examples where the data producer must create data sets for many users. The typical data analyst may have available only limited statistical and computational tools (such as packaged computer programs), and thus may not have the resources for a full Bayesian attack on nonresponse. Since I don't see it as likely in the near future that the typical data analyst will be able to perform the appropriate kinds of nonresponse analyses, I think that the data producer, usually having greater resources, may have to create multiple-imputed data sets so that the typical user can simulate a proper Bayesian analysis. The sophisticated user may want to ignore the multiple imputations and perform his own analyses using specialized models tuned for his interests in the data set. However, it is possible that even the sophisticated user with specialized complete-data models may want to rely on the multiple imputations that are provided because the data producer may understand the nonresponse problem more intimately.

William Cochran.--Professor Cochran's comments serve as a perfect opening to briefly mention future plans. We are now engaged in the development of a large-scale statistical and computational system for multiple imputation in the proposed Survey of Income and Program Participation (SIPP). Experience suggests that there may be substantial nonresponse problems on some income items. Since no SIPP data are currently available, we have thus far restricted our attention to the income data in the Current Population Survey (CPS). Some preliminary results from this CPS effort should be available this Winter, with a final report about this time next year. Theoretical work will be driven by this application. Hyperparameter nonresponse models that allow the borrowing of strength in large data sets will be crucial. The development of such models and their application to real data constitute a challenging and exciting statistical task having important implications in many contexts.