# DISCUSSION

**Roderick J.A. Little, University of Michigan**
**Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor MI 48109-2029**

## 1. Introduction

It is a pleasure to discuss this fine set of papers, which represent state-of-art methods applied to large, complex survey problems. The authors' pioneering efforts are extremely useful in advancing the field.

A major design change in the Decenniel Census is the proposal to subsample Nonresponse Follow-Up (NRFU) households who do not mail back Census forms. While aggregate data from such a design might be handled by standard methods for double sampling, much Census analysis involves micro-data files with information about individual households. In his excellent paper, Schafer (1995) argues persuasively that imputation is needed to create missing households that are not part of the NRFU subsample. Schafer proposes an ambitious model for multiply-imputing these households. The method factors the joint distribution of household characteristics into a sequence of conditional distributions, and applies Bayesian hierarchical logistic regression techniques. Factoring a joint distribution into a sequence of conditional distributions is an appealing strategy, particularly for household data where the joint distribution of characteristics has numerous structural zeros. Schafer's methods are principled in that they are based on a statistical model that can be criticized and elaborated. The approach is capable of handling item as well as unit nonresponse. Schafer shows great ingenuity in dealing with a variety of modeling issues, including a perceptive analysis of block heterogeneity and methods for incorporating correlation between neighboring blocks. The latter would be further enhanced by extension to two-dimensional spatial models. The stochastic simulation methods allow the fitting of a sophisticated model that would have been totally impractical until quite recently.

While much will be learnt from Schafer's analysis, it is not yet clear whether the gains over alternative methods, such as hot deck methods or the "top-down" approach of Zanutto and Zaslavsky (1994), justify the added complexity of Schafer's methods. Schafer dismisses hot-deck methods as "non-statistical", but they can be viewed as approximating draws from an implicit statistical model, if one that is not a very appealing since it includes all the high-order interactions between the classifiers (Lillard, Smith and Welch 1986; David, Little, Samuhel and Triest 1986). The hot deck can also be easily modified to create multiple imputes (Little 1988). While I agree with Schafer that his approach has clear advantages, careful comparisons with simpler methods like the hot-deck seem important to convince skeptics.

The alternative Zanutto and Zaslavsky (1994) approach has the advantage of concentrating on the outputs of primary interest, household counts, avoiding detailed modeling of the household structure. Schafer's approach is more comprehensive and hence better in principle, but there is a danger that computing limitations may lead to excessive simplification of the model structure. In particular, the preliminary analyses presented in the paper deal with block heterogeneity with random effects, and have very limited covariate information, mainly I suspect because of limitations in the dataset. The exchangeability of blocks and block groups implied by the random effects models is a strong assumption, and might lead to very unrealistic imputations for particular blocks. The data and model would be much improved by inclusion of good covariates characterizing blocks -- indeed I suggest that the formulation of useful covariates is a nontrivial and important topic of research. I fear that the inclusion of a rich set of covariates in Schafer's model may strain the computational limits, and approximate simplified models that retain the main features of the approach might be needed.

Heeringa (1995) presents an iterative simulation model for imputing multivariate asset amount data, where some of the data are available only as bracketed or "interval-censored" data. The paper presents a delightful combination of a clever data collection strategy -- asking for asset data in bracketed form to reduce the level of nonresponse -- with a clever analysis strategy -- multiply imputing the bracketed amounts to ease the subsequent analysis and allow propagation of uncertainty from the bracketing.

One attractive feature of the Gibbs' sampler applied to this problem is that it readily handles multivariate data, where more than one variable is reported in bracketed form. Rather than attempting to generate a multivariate draw from the joint distribution, Gibbs' allows a sequence of draws from the conditional distribution of one asset amount given parameters, partial information about that asset amount if available, and observed *or*

*drawn* values of the other amounts. The computational problem is thus reduced to a sequence of univariate draws, which is quite easy to handle. The Gibbs' sampler seems an ideal tool for this problem.

Heitjan's model for coarsened data provides a useful theoretical framework for Heeringa's problem. The definitions of Coarsened Completely at Random (CCAR), Coarsened at Random (CAR) and Not Coarsened at Random (NCAR) are useful analogues of corresponding notions of Missing Completely at Random, Missing at Random and Not Missing at Random in the missing-data literature. Consider the special case where the grouping indicator $G_i$ for subject *i* takes just two values, 0 when the actual value $X_i$ of a variable is observed, and 1 when the values is only known to lie in an interval $I = (X_{LB}, X_{UB})$. Then data are CCAR if $X_i$ is independent of $G_i$, so the distribution of $X$ looks the same for coarsened and non-coarsened data. Data are CAR $X_i$ is conditionally independent of $G_i$, given that $X_i \in I$. Heeringa's very interesting table showing the differences in the distribution of Stock and Mutual Funds for those who reported actual amounts and those who refused but reported coarsened amounts. This table provides information that the data are not CCAR, but does not provide direct information that the data are not CAR. However, it may be that a relatively parsimonious NCAR model would explain the observed differences in distribution and be more plausible than the CAR model, which implies that all the dependence of the coarsening mechanism on outcome is removed by conditioning on the coarsened form of the data. Such NCAR models provide an interesting topic for research.

Other modeling issues in Heeringa's proposed approach include the effects of misspecification of the lognormal model, particularly in the unbounded uppermost bracket. A sensitivity analysis to plausible alternative specifications may be needed here. Again it may be important to include covariates predictive of the missing items in the model. Finally, as in Schafer's analysis, it will be important to show that the modeling approach provides improvements over simpler more ad-hoc approaches, such as imputing interval midpoints or randomly drawing values within the intervals.

The paper by Judkins, Malec, Goksel, Hoffman, Shimizu and Monsour (1995) provides an interesting and relatively unusual application of Bayesian simulation methods in a design context. I like the idea of building "pseudo-populations" to address methodological issues, and was recently involved in such an effort in the context of assessing multiple imputation methods for the National Health and Examination Nutrition Survey (NHANES) (Ezzati-Rice, Johnson, Khare, Little, Rubin and Schafer 1995). Creating pseudo-populations is hard. If based on models fitted to observed data, model misspecification may bias method comparisons. If the pseudo-population is more empirically based, for example by bootstrapping an existing sample, then the resulting population is distorted by the restriction to sampled values. Information about the tails of distributions is limited, and some population structure may be impossible to capture; for example household composition cannot be recovered from a survey that sample's one individual per household. In our NCHS application, the pseudo-population was constructed by pooling a set of earlier samples. Simulation samples were then drawn by a weighted bootstrap that compensated for distortions arising from differential selection of the original data. Although the pooling somewhat ameliorated the limitations of bootstrapping, and the sample selection incorporated stratification, much of the clustering structure was lost because of limitations in the earlier datasets.

Given these problems, the need for a pseudo-population needs to be carefully motivated. For some purposes it may be avoidable. For example, if a sample design effect may be expressible as a function $DEFF = g(\sigma_1^2,\ldots,\sigma_q^2)$ of variance components $\sigma_1^2,\ldots,\sigma_q^2$. in such cases GIBS may be used to generate a draw from the posterior distribution of *DEFF*, by computing g at a draw from the joint posterior distribution of the variance components, estimated from the observed sample. I gather that here the *DEFF*'s of interest cannot be expressed in this simple form, thus motivating the creation of the pseudopopulation.

**Acknowledgments**

This research was supported by Bureau of the Census Contract No. YABC-2-66023.

**References**

David, M., Little, R.J.A., Samuhel, M.E. and Triest, R.K. (1986). "Alternative methods for CPS income imputation." *Journal of the American Statistical Association,* 81, 29-41.

Ezzati-Rice, T., Johnson, W., Khare, M., Little, R., Rubin, D. and Schafer, J. (1995). "A Simulation Study To Evaluate The Performance Of Model-Based Multiple Imputations In NCHS Health Examination Surveys." *Proceedings of the 1995 Annual Research Conference, U.S. Bureau of the Census*, 257-266.

Heeringa, S.G. (1995). "Application of Generalized Iterative Bayesian Simulation Methods to Estimation and Inference for Coarsened household Income and Asset Data." Invited paper, *Survey Research Methods Section, American Statistical Association 1995.*

Judkins,D., Malec, D., Goksel, H.A., Hoffman, K., Shimizu, I. and Monsour, M. (1995). using Mixed-Effects Modeling to Aid the Sample Design Process." Invited paper, *Survey Research Methods Section, American Statistical Association 1995.*

Lillard, L., Smith, J.P. and Welch, F. (1986). "what do we really know about wages: the importance of nonreporting and Census imputation. *Journal of Political Economy*, 94, 489-506.

Little, R.J.A. (1988). "Missing data in large surveys." *Journal of Business and Economic Statistics,* 6, 287-301 (with discussion).

Schafer, J. (1995). "Model-Based Imputation of Census Short-Form Items". Invited paper, *Survey Research Methods Section, American Statistical Association 1995.*

Zanutto, E. and Zaslavsky, A.M. (1994), Models for imputing nonsample households with sampled nonresponse followup". *Proceedings of the Survey Research Methods Section, American Statistical Association 1994*, 236-241.