DISCUSSION

David W. Chapman, U.S. Bureau of the Census
Washington, D.C. 20233

My discussion covers the papers by Eugene Burns and by Donaldson and Borgmann.

1. "Multiple Imputation in a Complex Survey," by Eugene Burns.

It is encouraging that survey statisticians are concerned with incorporating the effects of imputation in survey variance estimates. I hope that this concern will continue at EIA and at other organizations.

Although the method described by Burns is basically sound, it is technically not "multiple imputation" as defined by Rubin (1987). With Rubin's approach, the imputation process is repeated several times, creating a complete data set for each replication of the imputation process. The variation among estimates from the various data sets provides a measure of the imputation component of the total variance.

The procedure described by Burns applies to the special case of the use of a replication or pseudo-replication method to estimate survey variances. His procedure is simply to make imputations separately for each replication, rather than using the whole-sample imputations for the replicate data. Although this is not a new idea, it is often not done in practice due to the additional calculations required. With the increased power of computers, this approach of making separate imputations for each replicate has become more feasible.

In this special case of replication variance estimation, the approach described by Burns is ideal. In fact, it is preferable to Rubin's multiple imputation approach as a method of incorporating the effects of imputation in the variance estimates because it does not require creating multiple data sets.

Regarding the comparisons made by Burns between his approach and the procedure which does not generate separate imputations for each replicate, the differences in variance estimates were surprisingly large. Though I have not seen similar comparisons for other surveys, it is hard to believe that variance estimates could be as much as 45% higher when the imputations are made separately by replicate. I suggest that the calculations be checked carefully. If the calculations are correct, his results are quite significant.

When estimating the impact of imputation on variances for any survey, the effect of the response rate needs to be considered. In general, the higher the response rate, the lower the impact will be of imputation on variances. Future research in this area should try to address the effect of the response rate level on the imputation component of the variance.

Finally, I was not sure what the value of the simulation was. The main objective of the study was to compare the single-imputation variance estimates with those derived from the approach which imputes separately by replicate. Simulated nonrespondents are not needed to make this comparison. A possible value of simulation would be that the relationship between response rate and the effect of imputation on variances could be investigated by varying the response rate.

2. "A Multivariate Analysis of Farm Costs and Returns Survey Data," by W. W. Donaldson and R. E. Bargmann.

This research appears to be a good first step to developing an imputation procedure. However, this did not seem to be the main purpose of the study. The stated purpose (p. 2) "was to determine if there exists a subset of version 1 questionnaire item responses that can be imputed from the remaining questionnaire item responses with minimal information loss." They did not include all the survey items in their regression prediction analysis. They included items that satisfied three criteria (p. 2).

The only important criterion for selecting items for imputation analysis is whether or not they have nonresponse rates that are of concern. All items that suffer from considerable unreporting need to be addressed in an imputation procedure. The three criteria listed on p. 2 should not be used.

In some cases, special imputation procedures will be needed. For example, for items with a large number of zero responses, the first step in the imputation process may have to be a zero-nonzero (often a recipiency-nonrecipiency) imputation. For those receiving a nonzero imputation, the specific value would then be imputed next. For items that are not on an ordinal scale, regression imputation may not be the best choice. Perhaps a hot deck procedure, where missing values for a nonrespondent are supplied by those of a respondent having similar characteristics (i.e., a matched donor), would be most appropriate.

Regarding the development of their regression equations, they were developed among "homogeneous groups." This might work well but should be compared to alternative approaches. The authors assumed that normalization of item values was good. This should be investigated by comparing the results with those not using normalization. The purpose of smoothing data sets using the Gamma distribution, prior to comparing observed and imputed raw scores, needs clarification.

The authors should consider imputation procedures, other than regression prediction, to apply to their data. A residual should perhaps be added to regression imputes, especially if the preservation of variances among responses is important. Hot deck imputation, of which there are at least two types, should certainly be considered. A good discussion of imputation

methods is given by Kalton and Kasprzyk (1982).

Finally, in any study of imputation procedures the evaluation of alternative methods is difficult since the missing values (by definition) are not available. The authors used a random split of the sample to allow for evaluation. This is acceptable as part of the development of a good imputation procedure. But the key for evaluation is to investigate how well the imputations work for nonrespondents, not for a random half-sample.

One approach is to treat the set of respondents as the entire sample and generate "pseudo nonrespondents" from among the respondents by selecting appropriate numbers of respondents from various subgroups. This is not easy to do in a useful way but there are only a few evaluation methods available. In some cases, many, or perhaps all, of the missing values can be obtained from administrative records. If item definitions are consistent between the survey and administrative source, this is the ideal situation for evaluation.

REFERENCES

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons, Inc.

Kalton, G. and Kasprzyk (1982). "Imputing for Missing Survey Responses." Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 22-31.