# DISCUSSION

**Roderick J. A. Little**
**Department of Biomathematics, UCLA School of Medicine, Los Angeles CA 90024**

*1. Introduction*

My reaction to this strong session was to recall Mr. Squeer's comment on the food at "Do–The–Boys Hall" in Dickens' novel, Nicholas Nickelby:

"Here's Richness!"

Who says survey statistics are dull? The problems tackled here – complex non–rectangular data sets, highly skewed non–normal data, complex correlation structures from cluster sampling, differing degrees of aggregation and repeated measures, and multiple analysis objectives – are challenging and exciting, far from routine text–book statistics. The authors (and their sponsors) are to be congratulated for trying to do the right thing with hard problems, in real–world settings.

*2. Kennickell's Paper*

Arthur Kennickell's paper is extremely ambitious and path–breaking. Elsewhere (Little 1986), I have opined that imputations should be

a) based on models
b) appropriately conditioned
c) draws, not means
d) multiple, a la Rubin (1987)

This is easier said than done, and Kennickell attempts it in a dauntingly complex setting.

Some general features of his approach are A) a reliance on parametric, normal linear models, with some modifications for categorical data; B) some attention to edit constraints, by constraining the imputes within ranges; C) use of state–of–the–art technology such as multiple imputation (Rubin 1987), EM algorithms (Dempster, Laird and Rubin 1977, Little and Rubin 1987) and Gibbs' sampling/data augmentation (e.g. Tanner and Wong 1987); and D) a high degree of effort and complexity, possibly only practical in a repeated survey rather than a single study.

Continuous variables are treated by transforming (usually to the log scale), fitting normal linear models, adding normal noise and back–transforming. Care is needed in handling non–constant variance in this process. In David et al. (1986), we found that the variance of the logged outcome was inversely related to the mean, resulting in too much noise being added to the large means. On exponentiation this yielded some unacceptably large imputes. To alleviate the problem we imputed the conditional mean plus residuals stratified by the predicted mean. Kennickell's draws may solve the problem by applying constraints to avoid large values, but an extension of David et al.'s (1986) method to Kennickell's more general situation would be useful.

Kennickell applies a linear model to binary outcomes, which (as he acknowledges) is problematic for small or large proportions. An alternative, which avoids the iterative computations of logistic regression but has better statistical properties than Kennickell's method, is to apply discriminant analysis and impute according to the posterior classification probabilities, which always lie between zero and one. This form of discriminant analysis occurs in Little and Schluchter's (1985) maximum likelihood algorithm for mixed continuous and categorical variables with missing values, which might be useful in Kennickell's setting.

A more difficult issue is the treatment of variables involving presence of an attribute (such as an income source), and an amount if that source is present. Little and Su (1987) showed how the maximum likelihood algorithm of Little and Schluchter (1985) can be tricked to handle multivariate

missing data with variables of this kind.

Full parametric modeling of the kind attempted by Kennickell is a major undertaking in a large survey. It may be simpler to apply multiple–imputation versions of hot–deck methods that impute based on matches to respondent cases. More work on comparing the two approaches, along the lines of David et al. (1986) and Taylor, Lazzeroni and Schenker (1990), appears desirable.

*3. Lent's Paper*

The paper by Janice Lent is an impressive and detailed study of techniques that have been developed over a number of years. Three methods for assessing variability in complex surveys can be distinguished:

A) *direct methods*, where design–based variances are estimated by direct calculations from the sample, using Taylor Series expansions and sample reuse methods such as jackknifing or replication;

B) *models for design effects*, where models are used to estimate the inflation in variance due to the complex sample design, using data on design effects from prior surveys or other statistics from the same survey; and

C) *model–based variance estimation*, where the variance is based on super–population models for the survey variables; see for example Skinner, Holt and Smith (1989).

Lent compares examples of A (namely, the generalized replication method) and B (namely, the current GVF method) using labor force data from the Current Population Survey. In general, Lent's paper suggests that comparisons with the direct method estimates have provided a useful basis for refining the existing GVF method.

It would be interesting to see comparisons of the two approaches considered with a model–based approach C. Building models for all the survey variables is often a major task, but seems feasible for the small number of variables that are the focus of this inquiry. (I was a bit surprised by the emphasis on employment and unemployment *totals* rather than *rates*, but perhaps this displays my ignorance about the subject matter).

Part of the reason why design–based methods are favored is simplicity, in that they base variance estimates on variability between primary sampling units (PSUs), without requiring detail about design structure below the PSU level. In contrast, models generally involve the individual survey units, and hence require a more detailed specification. A possibility that might be worth pursuing is to develop simple models for PSU summary statistics. As a starting point, consider inference for a population mean $\bar{X}$ of a variable X, and in PSU k with selection probability $\pi_k$, let $\bar{x}_k$ be the sample mean for the $n_k$ sampled households. A basic model has the form

$$\bar{x}_k \sim N(\mu_k, \sigma_k^2),$$

where N denotes the normal distribution and $\mu_k = \mu(n_k, \pi_k)$, $\sigma_k = \sigma(n_k, \pi_k)$. The objective is then to find sensible parsimonious forms for the mean and variance functions $\mu_k$ and $\sigma_k$, the latter being particularly important for purposes of variance estimation.

*4. Hostetter and O'Conor's Paper*

The Hostetter and O'Conor paper is different in nature from the other papers, being more concerned with organizational issues. It provides an interesting case study in how to manage the redesign of a large government survey with multiple conflicting objectives. I leave it to others more qualified to debate the management issues. Many of the approaches adopted seem like "common sense", but achieving "a common sense" with many players is clearly not as obvious as it might seem, and requires management of a high order.

The statistical issues involve development of efficient designs for multivariate outcomes that are often highly skewed, where any sensible

design involves selection rates that vary greatly between strata. It would seem that existing theory is somewhat limited, and many potential Ph.D. theses lurk. To take just one example, the authors describe a method for discovering problems in the existing design by examining the selection rates of the largest units in the sample, and ensuring that small selection rates for these large units are increased in the new design. This idea seems useful but has a somewhat *ad hoc* flavor. More systematic approaches to such problems would seem worth future development.

## ACKNOWLEDGEMENT

This research was supported by USPHS grant MH 37188 from the National Institute of Mental Health

## 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.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society B, 39, 1–38.

Little, R.J.A. (1986), "Missing Data in Large Surveys," Journal of Business and Economic Statistics, 6, 287–301 (with discussion).

Little, R.J.A. and Rubin, D.B. (1987), Statistical Analysis with Missing Data, New York: John Wiley and Sons.

Little, R.J.A., and M.D. Schluchter (1985). Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values. Biometrika, 72, 497–512.

Little, R.J.A. and Su, H–L (1987), "Missing–Data Adjustments for Partially–Scaled Variables," Proceedings of the Survey Research Methods Section, American Statistical Association 1987, 644–649.

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

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989, eds.), Analysis of Complex Surveys, New York: John Wiley.

Tanner, M. and Wong, W. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), Journal of the American Statistical Association, 82, 528–550.

Taylor, J.M.G., Lazzeroni, L.C. and Schenker, N. (1990), "Robustness of Multiple Imputation Techniques with Application to AIDS," Proceedings of the Survey Research Methods Section, American Statistical Association 1990.