DISCUSSION

Margaret Gurney, Bureau of the Census

My comments have to do chiefly with the reliability of estimates of the variance made from the sample, whether the method used is the collapsed stratum method, the replication method, or the "direct" method described by Dr. Tepping.

The paper by Kish and Frankel uses $\sqrt{1/2L}$ as a measure of the coefficient of variation of the standard error of an estimate from a design with $L = 47$. This result is obtained if we assume that β is approximately 3, in the well-known formula for the rel-variance of the estimate of the variance [1].

The assumption that β is about 3 may be far from true, as is indicated in Tables 1 and 2 of the Tepping paper: much of the variance, and substantially more of the variance of the variance (or the CV of the standard error) may come from a few (sometimes only 1) of the strata. In examining the between stratum variances for several of the characteristics in his tables we found values of β of 33, 68, and 157, for individual strata. Combining 2 strata (as is done in the collapsed stratum technique, which is used in all of these papers) may result in an average value of β which is nearly as large as the larger of the β's for the two strata, if the pairing of strata for collapsing has been inefficient for the particular estimate being considered.

If, as implied in the Tepping Tables 1 and 2, a few strata may dominate not only the variance, but also the variance of the estimated variance, the average β for the whole sample design may be much larger than 3. If, for example, the average β is of the order of 33, the formula for the CV of the standard error with the CPS design and 120 paired strata would lead to a CV of more than 25 percent; with fewer strata, such as 47 pairs, the CV would be larger, about 40 percent.

Admittedly, a β of 33 for the whole sample design seems quite large, but there are many important agricultural crops (for example rice, sugar cane, citrus fruits) which are grown in only a few parts of the country, and for these β may be appreciably greater than 3. Similarly, there are many important industries which are localized, and for which a national sample might produce β's which are larger than 3.

It is important, therefore, to know what is going into the variance. With the replication method this is difficult, since we do not see the individual original strata. If most of the variance comes from one collapsed stratum, one-half of the balanced replication estimates may be much larger than the other half; if most of it comes from two collapsed strata one-fourth of the replicates may be inordinately large. But the individual contributions to the estimate of the variance are not displayed.

This discussion of the distribution of the replication estimates leads to mention of Addendum II on page 34 of the Simmons-Baird paper, where it is suggested that much could be done with the individual estimates which are produced by the replication method (28 estimates in the HEW survey of the paper). The idea of displaying these individual estimates as a routine part of production of the data is a good one -- it is not new, having been stressed on numerous occasions by the chairman of our meeting. It brings the analyst one step closer to the original data: it can be used to some extent as a quality control device, in that finding an estimate which deviates considerably from the other estimates may indicate that something has happened in the processing of that replicate. The distribution of the estimates gives some feeling for the variability of the replication estimates, and may indicate that most of the variability comes from one or two strata. It could perhaps take the place of variance calculations, for less important or infrequently collected statistics.

[1] Hansen, M.H., W.N. Hurwitz, and W.G. Madow. (1953). Sample Survey Methods and Theory, Vol. I, p. 427. New York: John Wiley & Sons, Inc.