DISCUSSION

Joseph Waksberg, Westat, Inc.
1650 Research Blvd., Rockville, MD 20850

1. Oversampling Low Income Population

The authors have used an imaginative approach on an important and frequently occurring problem in U.S. government surveys. It's unfortunate that they had to restrict their analysis to the housing unit sample frame. This makes the results inapplicable to surveys conducted outside the census Bureau by researcher who do not have access to the Census lists of addresses, and more importantly to the Census data for the individual households. I hope that at some future time funds will become available to permit additional research for stratification carried out at the block or block group level.

The reductions in variance arising from oversampling specific addresses appear to be much greater than what could be accomplished by oversampling blocks containing high concentration of low income persons. For surveys carried out through area sampling, geography is not a very efficient stratification device for poverty. Part of the reason is that although the poor who live in central cities of metropolitan areas are generally concentrated, most of the poverty population lives outside central cities. The vast majority of poor outside central cities are geographically dispersed. Recent CPS reports indicate that less than 40 percent of persons below the poverty level live in areas defined by the Census Bureau as poverty areas. This 40 percent was greatly influenced by the fact that two-thirds of the blacks and 55 percent of Hispanics below poverty were in poverty areas. Of the 56% of all poor who were non-blacks and non-Hispanic only 22 percent lived in poverty areas.

The poverty areas in the CPS reports are defined to consist of complete census tracts or MCD's. Using smaller areas does provide somewhat better discrimination but the basic problem still exists. Some years ago we prepared a tabulation from the 1970 census showing the distribution of poverty among block groups and ED's. Only 28 percent of persons below the poverty level were living in BG's or ED's that had over 30 percent of their residents who were below the poverty level, with another 45 percent in areas with 10 to 30 percent in poverty. As is the case in the more recent CPS reports, most of the black and Hispanic poor were in low income areas but only a minority of the white poor.

In fact it is necessary to get down to ED's and BG's that have 10 to 20 percent of the population in poverty to cover as much as 50 percent of the poverty population. This implies that most of the oversampled persons in these areas turn out to be above poverty. Oversampling on the basis of geography thus turns out to be a very expensive way of increasing the sample of low-income persons and is not an effective sampling method with area samples. It is interesting to find out that the Census Bureau can use its microfile of Census data to make oversampling viable.

I want to move to other issues discussed in the paper. First, I am puzzled by the large reductions in variance reported for general poverty statistics, e.g., number of persons below 150% of poverty. The auxiliary variable for oversampling concentrated on black and Hispanic characteristics, and this is reflected in the sample size increases shown in Table 3. The bottom two lines in Table 3 indicate the sample size increase for persons in poverty was entirely for blacks and Hispanics. Since blacks and Hispanics account for less than half of the persons in poverty, one would not expect the oversampling to be that effective. Could the results be an artifact of the limited number of PSU's used for the study.

The authors report only a slight effect of elapsed time on the efficiency of oversampling, and I'm not sure whether to believe it. A recent report from SIPP on transitions in and out of poverty indicates that about 25 percent of persons who are in poverty in one year move out of poverty in the following year. They are replaced by approximately the same number who move into poverty. This is a little lower than the 30% about 20 years earlier, reported in my 1973 paper that the authors refer to, but it is still quite high. The transitions reported in SIPP thus do not appear to reflect the situation in an unusual set of years.

In the course of a few years these transitions obviously increase. For this amount of movement not to affect the efficiency of oversampling, it would be necessary for most of the people entering poverty to move to the housing units previously occupied by persons who left poverty. I find it hard to believe that this occurs on such a large scale and suggest the analysis be reviewed. I wonder if these results are due to measuring the effect of time changes by looking at the deffs rather than at CV's. With part of the oversampled group leaving poverty, I would expect the sample size of persons in poverty in the second year to be much lower than the first year. Has this been reflected in the analysis.

My final comment on the paper concerns the procedure for determining oversampling rates. These rates are determined separately for each PSU in order to keep the interviewers' workloads constant among PSU's. The authors state that this does not have much effect on the optimum allocation. I am surprised at that. The constraint to keep workloads constant has a peculiar effect on the sample. At the extremes, the greater the number of low-income households in a PSU, the smaller the oversampling rate. For example, if the entire population of a PSU is low-income, no oversampling is possible, whereas oversampling at quite high levels will be applied in PSU's with very low poverty rates. This is almost exactly the opposite of what one would like to do. I suggest examining the possibility of bending the constraint. There must be many cases in which interviewers can absorb an increased sample or in which a 10 or 20 percent reduction in sample size would still provide an efficient workload.

2. Household Clustering

Pat Cantwell's paper also involves using resources for sampling that are uniquely available to the Census Bureau. The hypothesis that deliberately introducing heterogeneity into the clusters would reduce variances seems plausible, and I was disappointed that the gains were too trivial to bother with.

The decision to retain the current plan of basically using compact clusters unless important gains are possible with an alternative clustering method makes sense. Pat mentioned that compact clusters are less expensive than more dispersed ones because the travel costs are lower. If the noncompact clusters are kept within the same block or block group then the decreased travel is fairly small, but there is another reason for preferring compact clusters. I was largely responsible for introducing address samples into the Census household surveys. The main reason for using compact clusters was that this usually resulted in all housing units in small structures being in the sample simultaneously. Apartment numbers frequently do not exist in these small buildings, and the descriptions of the individual addresses may be ambiguous or nonexistent. Having the entire building in the sample eliminates the problem of trying to identify specific sample households and thus avoids potential biases. The sample is much more highly controlled. The variance reduction of an alternate scheme for clustering should be quite strong to outweigh this advantage of compact clusters.

I am not familiar with the current detailed procedures for creating clusters in the sample surveys when address lists compose the sampling frame. If it is not being done now, I would encourage the Bureau staff to examine the feasibility of establishing clusters in such a way that almost all one to four unit buildings remain intact. This might occasionally require combining addresses that are physically separated but it would provide tighter control on the sample.

3. Within PSU Sort and Stratification

My earlier comment that oversampling geographic areas for low-income statistics is not effective for variance reduction does not apply to sorting and stratification. The problems of oversampling geographic areas arise from the fact that it results in oversampling persons who are not low-income but who happen to live in the same blocks as those with low income. This wastes considerable resources. Simply sorting and stratifying the blocks without oversampling does not incur any of these problems. The procedure is equivalent to stratification with proportionate allocation which almost always produces reductions in variances, generally fairly modest.

Greater gains from stratification were reported in urban than in rural PSU's presumably because poverty is not as geographically concentrated in rural as in urban PSU's. The overall variance reduction in the U.S. will therefore be less than shown in the paper's tables and chart, which reflect the urban situation only. However, since there is only trivial cost in stratifying and sorting, any gains are worthwhile.

The variables used in stratification and sorting were auxiliary variables rather than income itself, because income data are not available at the block level. Block groups are larger geographic units and as a result cruder indicators of household characteristics. However, since income distributions are available for block groups, it may be that income stratification at the block group level is more effective than auxiliary variables for blocks. It would be interesting to explore this possibility.

My final comment relates to the issue of whether boundaries for the stratification classes should be uniform across all PSU's or be uniquely established for each PSU. The authors take it for granted that a separate determinations for each PSU would produce lower variances. It's not self evident to me. Where there are multiple levels of stratification, the value of the final level or levels is affected by what has gone on before. Gains from stratification at the last level are frequently reduced by earlier stratification. An ineffective but detailed stratification at the first or second levels could interfere with the potential value of the last level. It seems sensible to me to use stratification and sorting of the type described, but I would explore more fully the difference between choosing overall stratum boundaries and making a separate decision in each PSU.