VARIANCES FOR THE SURVEY OF INCOME AND PROGRAM PARTICIPATION 1984-1996 PANELS: A CHRONOLOGY AND EVALUATION OF DIRECT AND GENERALIZED VARIANCES

Jennifer Guarino, Vicki Huggins, Robert Fay, and Aref Dajani, U.S. Bureau of the Census
Jennifer Guarino, U.S. Bureau of the Census, Washington, DC 20233

Key words: Replicate variance estimation, generalized variance function, Primary Sampling Unit (PSU)

I. Introduction

The Survey of Income and Program Participation (SIPP) has undergone tremendous change over the last 14 years. The survey program began with three major goals:

- provide detailed monthly data on the income, program participation and program eligibility of the U.S. population
- provide a longitudinal cohort sample with detailed income and program data
- provide supplemental data on current, important topics such as child care.

The goals expanded in recent years to include the following:

- provide more accurate and more versatile poverty measures, both cross-sectionally and longitudinally
- provide data to help assess welfare reform.

With so many conflicting goals, it has been difficult for the SIPP program to meet any one with the highest level of confidence. In the early '90s, we polled data users to determine which of the conflicting goals were most important. SIPP data users strongly recommended that the Census Bureau focus resources on improving the reliability of longitudinal data with particular concern for the low income population. As a result, we redesigned the 1996+ SIPP panels to have larger longitudinal samples that remain in the field longer and contain an oversample of the low income population. To do this within a fixed budget, we discontinued the overlap of SIPP sample panels, and switched to an abutting 4-year panel design.

In addition to the sample design and reliability criteria changes through the years, we identified methods to improve the estimates of variances from the survey, for both direct variances of key statistics and for user-friendly generalized variance estimates that cover groups of characteristics.

The goal of this paper is to describe the changes that have taken place as well as their implications on the variance. This paper describes:

- Survey design changes in the SIPP since 1984
- Improvements in both direct and generalized variance estimation techniques
- An assessment of how our new methodologies have affected the reliability of key characteristics over time.

NOTE: For more details on this study, see Guarino et al (1998).

II. Survey Design Changes

The sample design parameters of a survey are key to ensuring adequate reliability of the survey estimates. Table 1 shows a chronology and description of the survey design changes.

Table 1. Changes in the SIPP sample design: 1984, 1990, 1996

Panel	PSUs per Stratum	Over-sample	Cluster Sizes in the Frames		
			Unit 70%	Area/ Permit 27%	GQ 3%
1984-1989	1	no	4	4	4
1990	2	yes	2	4	4
1991-1993	2	no	2	4	4
1996	2	yes	1	4	1

SIPP began as a 1 PSU per stratum design. SIPP borrowed leftover sample from the Current Population Survey (CPS) which has a 1-PSU per stratum design. As the goals for the SIPP became more defined, researchers including Shapiro (1981) decided that a 2-PSU per stratum sample design would be more advantageous for the SIPP since it allows for the calculation of Yates-Grundy estimators of the design variance, which are both unbiased and non-negative.

For the 1990 and 1996 panels, we started to

oversample households with certain characteristics. Specifically for the 1996 panel, we oversampled low income households based on their 1990 decennial characteristics. The goal of this oversampling was to improve the reliability of subgroups in poverty, such as blacks in poverty and hispanics in poverty, without substantially hurting the reliability of other characteristics. Based on analyses from the 1996 panel, we increased the sample size for total persons below 100% of poverty by 11%, and blacks in poverty by 24%.

In the earlier panels, we chose a systematic sample of clusters of sizes 2 to 4 within PSUs. By 1996, we selected clusters of size 1 within PSUs for those blocks with good addresses (roughly 70% of the sample households). Due to this drop in cluster size, we expected to see reductions in the design variance since we are virutally eliminating the between cluster component of the within-PSU variance. Since the total within-PSU variance component is generally around 80% of the total variance in the survey, the reductions in variance were substantial.

Finally, another change that will improve our estimates for the 1996 panel is an increase in sample size. The SIPP was originally designed to be an overlapping panel program with each sample panel including 20,000 interviewed households and a new panel starting each year. The largest panel contained about 23,000 interviewed households. Prior to the 1996 panel, SIPP switched to a 4-year abutting panel design yielding about 37,000 interviewed households for the 1996 panel, making it our largest sample ever.

As a result of increasing the sample size of individual panels and oversampling in the 1996 panel, we expect the reliability for practically all estimates in the 1996 panel to be significantly improved as compared to the 1984 - 1993 panels.

III. Estimation of Direct Variances

We use a replication approach to estimate direct variances. Since 1985, the replication approach for internal replicates (that is, not for public release) consists of two basic components:

1. A replication-based representation of the Yates-Grundy estimator for the two-per-stratum selection of non-self-representing PSUs. A description of this method can be found in Fay (1989).

2. A modified version of half-sample replication, sometimes referred to as Fay's Method, to represent within-PSU variation. This method is used in all self-representing PSUs, but is also employed to represent a component of within variance in non-self-representing PSUs in conjunction with use of the Yates-Grundy estimator in a multistage situation. For $0 \leq k < 1$, the modification produces replicate estimates by weighting observations by the factor k or $2-k$ in place of the 0 or 2 used in the original version of half-sample replication. The corresponding variance estimator is

$$var(\theta_0) = \frac{1}{G(1-k)^2} \sum_{i=1}^{G} (\theta_i - \theta_0)^2$$

where G = number of replicates
1-k = perturbation factor
i = replicate i
θ_i = estimate of θ for replicate i
θ_0 = estimate for the full sample

The value of $k = .5$ has been used in all SIPP applications. For linear estimators, the modified version produces the same estimated variance as the original. With $k > 0$, all replicate totals will be positive for totals that are positive in the full sample, unlike half-sample replication. This feature has practical advantages in nonlinear applications.

The integration of these two components is described by Fay(1989) with respect to the design of the 1985 panel. In addition, the treatment for 1985 had the added complexity of representing between-stratum variance due to subsampling strata in the 1985 design. Strata were not subsampled in the 1996 design.

With $k = .5$, the method allows for a single set of replicate weights incorporating both theYates-Grundy and modified half-sample replicates simultaneously, without any negative replicate weights, for use in the variance estimator above. Use of the original form of half-sample replication, $k = 0$, would not have yielded a similar outcome.

Because the majority of the SIPP sample is drawn by systematic sampling from the census frame, alternate assignment of segments to half-samples within a stratum captures much of the effect of the systematic sampling in a demographic survey like the SIPP. Arguably, jackknife variance estimation is an alternative to half-sample replication. Comparing the two, half-sample replication generally gives better confidence interval coverage and performs better across a variety of populations, as noted by Rust (1986), drawing on numerous studies comparing the properties of different variance estimators. Primarily, however, we selected modified half-sample replication for its closer representation of systematic sampling.

We have used the term *panel assignment* to refer to

the assignment of replicate weights to SIPP segments. In practice, this has meant assigning each segment to a set of rows of a Hadamard matrix with associated coefficients. Within each major redesign of the SIPP, we have tried to assign panels consistently across time. For example, we intend to assign panels for the next SIPP panel consistently with our 1996 assignments, so as to reflect the covariances across time. In assigning PSUs to rows of the appropriate Hadamard matrix, consideration was given to obtaining regional and national variance estimates and estimates for important demographic groups.

For the 1996 panel, we considered but decided against a newer replication method for systematic sampling, *successive difference replication,* now in use in the Current Population Survey (CPS), American Community Survey (ACS), and under consideration for the long form in Census 2000. The primary reason to remain with half-sample replication for the SIPP was flexibility, since successive difference replication primarily employs differences between neighboring segments in its variance estimate. We needed a replication method designed to allow consistent use of sample hits across multiple panels of the SIPP, but the past history of the SIPP has shown considerable variation in the realized sample size over time.

We carried out direct estimation of the variance for about 440 characteristics for 1983 third Quarter data of the 1984 SIPP panel. Replicates for the 1984 panel were not individually reweighted for the noninterview adjustment and ratio estimation to population totals. The ratio estimation substantially reduces the sampling variance of characteristics highly correlated with the population controls used. Since we did not replicate this in the 1984 half-samples, the resulting estimates of variances overestimated the sample variances. It is more difficult to know the effect of replicating the noninterview adjustment in variance computations. The purpose of noninterview adjustment is to reduce the bias of survey estimates associated with nonresponse. This could lead to increases or decreases in the sampling variance depending on the level on nonresponse and the noninterview cell definitions. Empirical experience with the CPS suggests that these differences are more minor than the effect of ratio estimation, although the magnitude of the SIPP noninterview adjustment is generally higher than CPS.

We did not compute direct variances for the 1985-1989 SIPP panels, primarily due to resource constraints.

In the later panels, we decided to reweight the replicate estimates to obtain the most accurate variance estimates possible. Weight components form an integral part of the estimation system, and so can not be ignored when it comes to estimating sampling variances. Lemeshow(1979) showed that using a single set of whole sample weights for each replicate estimate results in substantial bias in the variance estimation. For the 1990 panel, ratio estimation to population controls was included, and for 1996 noninterview and ratio estimation were replicated.

IV. Estimation of Generalized Variances

The 1984 panel marks the last time that SIPP generalized variance estimates were produced and published from raw data. The model used was $var(x) = ax^2 + bx$ or $v^2 = a + b/x$, where x is an estimate, var(x) is the variance estimate of x, and v^2 is the relvariance estimate of x. Since then we applied adjustments to model parameters to construct new generalized variance estimates for the panels thereafter.

Because of changes in the SIPP survey design, the 1996 Panel required completely new generalized variance estimates produced from raw data. We were also interested in determining whether the functional form used to calculate generalized variance estimates in 1984 was still appropriate as applied to the 1996 redesigned SIPP Panel.

For SIPP, direct variances were calculated for approximately 1,700 items. Once we obtained direct variance estimates, we had to do the following to calculate generalized variance estimates:

1. Generating candidate functional forms.
2. Transform data appropriately.
3. Defining domains of interest.
4. Eliminate outliers.
5. Determining an appropriate functional form.

1. Generate candidate functional forms.

Though the ordinary least squares model has been used traditionally to model variances of the major demographic surveys, we decided to investigate alternative functional forms. The functional forms we compared are as follows:

$v^2 = a + b/x$
$v^2 = a + b/x + c/x^2$
$v^2 = a + b/x + c/SQRT(x)$
$v^2 = a + b/x + c/x^2 + d/SQRT(x)$
$v^2 = 1 / (a + bx)$
$v^2 = 1 / (a + bx + cx^2)$
$v^2 = 1 / (a + bx + c*SQRT(x))$
$v^2 = 1 / (a + bx + cx^2 + d*SQRT(x))$
$log(v^2) = a + b*log(x)$

2. Transform data appropriately.

In order to look at the data to define domains of interest, we ran log-log plots of the direct relvariances and their corresponding estimates. We found that log-log plots were useful in detecting outliers. The log-log plots were used because the distribution of both the direct relvariances and their corresponding estimates clustered around zero. Taking the logs spread out the observations sufficiently to discriminate items into domains and locate outliers.

3. Define domains.

Using the approximately 1700 variance items, we grouped items with similar variance patterns and definition into domains. We created new domains when we could gather at least 100 items and collapsed domains with less than 100 items.

The final eleven domains are as follows:

1984 Domains

Household Measures

Household
Household, Black

Person Measures

Program Participation and Benefits, Poverty
Income and Labor Force
Other Person Measures
All Person Measures, Black

1996 New Domains

Household Measures

Household, Hispanic
Household, Metro and NonMetro

Person Measures

All Person Measures, Hispanic
All Person Measures, Metro and NonMetro
Poverty by Demographic Subgroup
(age by race by gender by marital status)

4. Eliminating outliers.

Ordinary least squares were run for each model by domain, using functions of the design-based final reweighted variance estimates as the dependent variable. We used output from the least squares runs to calculate the absolute relative deviations (ARD) for each observation, defined as the absolute difference between the predicted relvariance and the observed relvariance, divided by the observed relvariance. We then conservatively defined outliers to be those observations with either a studentized residual greater than 4 or an ARD greater than 50. Studentized residuals indicate outliers at the higher end of the curve of residual by predicted values and ARDs indicate outliers at the lower end of the curve of residual by predicted values.

5. Determining an appropriate functional form.

After removing outliers from further analysis, we then reran ordinary least squares for each model by domain and calculated 7 "goodness of fit" measures for each model by domain:

number of outliers deleted (across domains)
adjusted R^2 statistic
significance test statistics of model parameters
mean ARD
scatterplot of residuals by predicted values
normality test statistic for residuals
percent of items with predicted variances < zero.

The analysis for the traditional model $v^2 = a + b/x$ was run using iterative weighted least squares as well as unweighted ordinary least squares. For any functional form, iterative weighted least squares is a standard technique used to stabilize regression model parameters. Iteratively weighting by the squared inverse of predicted relvariances and removing outliers at each step until those predicted relvariances converge ensure that items with low relvariances are given more weight than those with high relvariances. This is because many consider items with high relvariance to have less reliable variance estimates than those with low relvariance. Evaluating the results, we used iterative reweighting to ensure that upper tail estimates such as poverty and program participation are estimated accurately by the model.

Running the analysis across functional forms and comparing goodness-of-fit measures, the following two models performed well for the 1990 and 1996 panels with the advantage of each listed by the ✚ below:

$v^2 = a + b/x$ (with and without population controls)

✚ no change in functional form from 1984
✚ property of variances of proportions is preserved:
Var(p) = Var (1-p)
✚ with population controls, variances are never negative

$\log(v^2) = a + b*\log(x)$

✚ better goodness-of-fit measures
✚ variances are never negative

For the 1996 Panel, the specific model we selected was the iterative weighted least squares model $v^2 = a + b/x$ using population controls because of a specific favorable property of variances of proportions found in

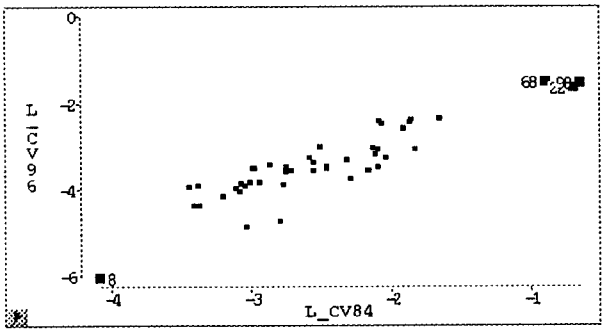
simple random sampling, variance of a proportion is equal to the variance of one minus that proportion. When comparing iterative weighted and ordinary least squares, we observed that the generalized variance estimates were more conservative for iterative weighted variance estimates, that they were less likely to understate the direct variance estimates.

V. Comparison of 1996 to 1984 and 1990

This section summarizes the results of graphical comparison of SIPP estimates and direct variance estimates for the poverty and program participation domains from the 1984, 1990, and 1996 panels. The main focus of this section is comparing the 1996 panel estimates with their corresponding estimates from the 1990 and 1984 panels.

Graph 1 below shows the comparison of the natural log of the coefficients of variation(CV) between 1984 and 1996, for items in the poverty and program participation domain. Three variance items are clustered away from all other items at the high end of the log scale of CV. In the original scale, this translates into CVs in the range of 40 to 55 percent for the three items in the 1984 panel while the next highest CV was 20 percent. This graphs shows that SIPP poverty and program participation CVs are consistantly smaller in 1996 than 1984, regardless of the relative size of the CVs.

Graph 1. Scatteplot of ln(CV '96) and ln(CV '84)



To further summarize reliability results, we look at some CVs, variances, and design effects for key estimates: persons ages 16+ in low income households for all persons, blacks, Hispanics, and for households receiving food stamps. Table 2 compares CVs for certain poverty items over the 1984, 1990, and 1996 panels. Note that CVs for 1996 are less than half of their value in 1984.

Table 2. Comparison of CVs for Key Estimates

Item	1984 CV	1990 CV	1996 CV
Persons 16+ in Low Income HH	.033	.021	.013
Blk Persons 16+ in Low Income HH	.086	.040	.031
Hisp Persons 16+ in Low Income HH	.078	.046	.031
Blk Persons 16+ in HH with Food Stmps	.116	.047	.034
Hisp Persons 16+ in HH with Food Stmps	.132	.079	.049

Similarly, Graph 2 below shows that there has been major reductions in the variance of poverty and program participation variables from 1984 to 1990 to 1996. The first three items in the graph correspond to the items in rows 1, 2, and 3 respectively in Table 2 above. The last item listed in the graph is the variance for the estimate of all persons 16+ in low income households receiving food stamps.

Graph 2.

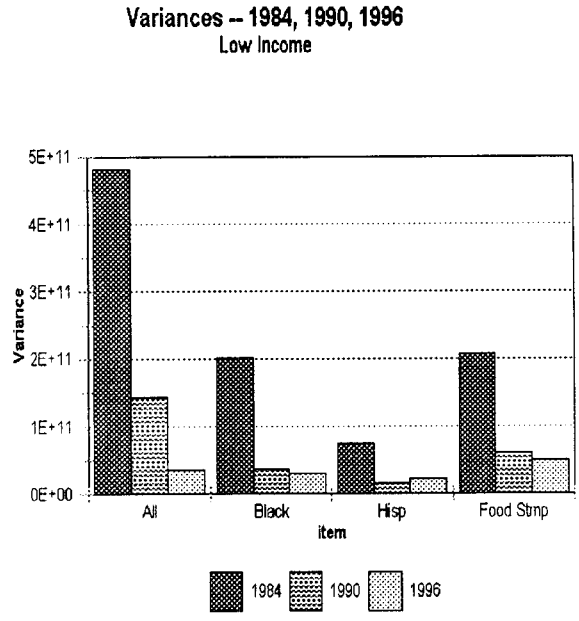


Table 3 contains design effects for the poverty and

program participation domain for 1984, 1990, and 1996. Note the substantial reduction in the design effect from 1984 and 1990 to 1996. Much of this drop can be attributed to the decrease in cluster size over the panels. The introduction of oversampling in 1990 and 1996 may also play an important role here. From these results, we see that our changes over the years have produced results which are consistent with our goals of reducing sampling variance for statistics of key interest.

Table 3. Comparison of Design Effects: 1984, 1990, 1996

Domain	DEFF 1984	DEFF 1990	DEFF 1996
Poverty and Program Participation	4.32	4.10	1.82

Table 4 below contains summary statistics for the comparison of CVs between the 1984 and 1996 panels for the poverty and program participation domain. Note that both the mean and median CV for this domain in 1984 are notably higher than that of 1996 (roughly 5.6% higher for the mean, 4.3% higher for the median). Likewise, the maximum value or highest outlier of the 1984 CVs in this domain is twice that of 1996. Furthermore, the average difference between CVs in 1984 and 1996 is about 5.5%.

Table 4. Comparisons of CVs in Poverty/Program Participation Domain: 1984 to 1996

Summary Statistic	1984 CVs	1996 CVs
Mean	10.3%	4.7%
Median	7.4%	3.1%
Max	42.7%	22.1%
Average Difference	5.5%	

VI. Concluding Remarks

Our evaluation shows that changes made in the SIPP survey design and methodology for estimating variances have improved the reliability of many key estimates in SIPP. The model testing performed validates previous assumptions about the best functional form for modeling SIPP variances. Likewise, including all steps of our weighing process in the reweighting of the replicates helps improve precision, as it reduces bias associated with non-response and undercoverage. With these changes and the results of our evaluation in part V, we are confident that SIPP estimates are more accurate and reliable than ever.

VII. Acknowledgments

The authors extend special thanks to: Tracy Mattingly and Elaine Hock of the Census Bureau for their help in the design, implementation, and documentation of the 1996 direct variance system, Tom Krenzke of WESTAT and Chris Moriarity of the National Center for Health Statistics for helpful discussions which are reflected in various ways in this paper and for posing several of the questions that led to the work in generalized variance estimation reported here, and Yuki Ellis who allowed us to use parts of her evaluation as well as some of her graphs in Section V.

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

REFERENCES

Fay, Robert E. (1989), "Theory and Application of Replicate Weighting for Variance Calculations," *1989 Proceedings of the Section on Survey Research Methods*, American Statistical Association, Alexandria, VA, pp. 212-217.

Guarino, J., Huggins, V., Fay, R., and Dajani, A. (1998), "Variances For The Survey of Income and Program Participation (SIPP) 1984 to 1996 panels: A Chronology and Evaluation of Direct and Generalized Variances", Census Bureau Memorandum Dated September 17, 1998.

Lemeshow, S. (1979), "The Use of Unique Statistical Weights for Estimating Variances with the Balanced Half-Sample Technique" *Journal of Statistical Planning and Inference* 3, pp. 315-323.

Rust, K. (1986), "Efficient Replicated Variance Estimation," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, American Statistical Association, Alexandria, VA, pp. 81-87.

Shapiro, Gary (1981), "Research on One vs. Two PSUs Per Stratum: Things to Consider in the Next Stage of the Study", Census Bureau Memorandum Dated November 25, 1981.