# ALTERNATIVE DESIGNS FOR THE CONSUMER EXPENDITURE SURVEY

**F. Jay Breidt**
**Iowa State University**

**Abstract.** The Consumer Expenditure Survey is a two-stage probability sample of U.S. households. Like other U.S. household surveys, it is redesigned in conjunction with the decennial census. Two key stratification variables (contract rent and housing value) formerly collected on the short form of the census will be unavailable for the year 2000 redesign. In the 2000 census, these variables will be available only from the long form, a roughly one-in-six sample. Loss of these stratification variables has led to consideration of three alternatives: stratification on small domain estimates from the long form, two-phase sampling for stratification, and a combination of these two approaches. Selection bias may arise in the long form due to nonresponse of sampled households or "conversion" of nonsampled households, which may occur if long forms are delivered to the wrong addresses or follow-ups go to the wrong households. Quasi-random sampling models for both selection problems are developed. Mean squared error approximations are derived using an asymptotic framework in which the number of small domains becomes large. Optimal mean squared error designs are considered.

**Keywords.** Double sampling, nonresponse, selection bias, small domain, two-phase sampling.

## 1. Introduction

The Consumer Expenditure Survey is a two-stage probability sample of U.S. households, designed to represent the total U.S. civilian noninstitutional population. Primary sampling units (PSU's) consist of counties, groups of counties, or independent cities. Sampling within PSU's involves stratification on household size, tenure (rent/own), and contract rent or housing value categories, available from the short form of the decennial census (Bureau of Labor Statistics, 1997). The rent and value variables are particularly important stratifiers, since housing expenditures are typically large and highly correlated with other kinds of expenditures.

In the 2000 census, the rent and value variables will not be available from the short form, but will be available from the long form. (The long form is roughly a one in six subsample of the short form.) Because loss of the rent and value variables is expected to lead to a loss of efficiency in estimates from the Consumer Expenditure Survey, alternatives to the current stratification scheme are under investigation. In this paper, three alternative designs are considered, and contrasted with the current design.

The first alternative is to subsample from those households which responded to the long form, stratifying directly on rent or value (or on additional stratifiers available from the long form). This is the classical two-phase sampling for stratification. A potential complication

is selection bias in the long form respondents. Two selection mechanisms are of concern. The first is nonresponse: some sampled units fail to respond. The second will be called "conversion": some nonsampled units are converted to respondents. This can occur, for example, if long forms are delivered to the wrong addresses, or follow-ups go to the wrong households.

Both nonresponse and conversion can lead to bias if their propensity is related to the study variables of interest. In Section 2.2, this problem of two-phase sampling for stratification with nonresponse and conversions is studied in the context of quasi-random sampling models for both selection mechanisms. Households are assumed to independently respond if selected, or convert if not selected, with household-specific probabilities.

The second alternative is to compute small domain estimates of rent or value, and use these domain-level estimates to stratify all households within the domain. The small domains might be defined as block groups within particular tenure class by household size classifications. Then, for example, the block-group averages of rent from the long form for rented two-person households would be used in place of the actual rent for these households. Details are given in Section 2.3. This strategy of stratifying on small domain estimates entails some loss of efficiency.

The third design alternative is to combine the first two, drawing an efficiently stratified sample from the long form respondents, but guarding against potential bias by drawing an additional sample from the remaining units. This additional sample can be stratified on small domain estimates of rent and value from the long form. The two parts of the sample can be weighted so that bias is not a factor, or can be weighted to trade some bias for small variance. Optimal mean squared error designs are considered in Sections 3.2 and 3.3. Numerical examples illustrating the results are investigated in Section 4.

## 2. Preliminaries

In this section, notation for stratification is introduced, and quasi-random sampling models are developed to describe nonresponse and conversion problems.

Let $U = \{1, \ldots, j, \ldots, N\}$ denote the universe of immediate interest, which for our purposes would be a particular tenure class by household size category within a PSU; for example, two-person rented households in Cook County. The universe is divided into $D$ disjoint domains, so that $U = \cup_{i=1}^{D} U_i$. Useful domains for this discussion would be block groups or similar geographic subdivisions. Let $M_i$ denote the number of households in domain $U_i$, and note that $N = \sum_{i=1}^{D} M_i$. For convenience, we model the $\{M_i\}$ as independent and identically distributed (iid) random variables.

For each $j \in U$ there is an auxiliary variable, $X_j$, for the variable of interest, $Y_j$. The auxiliary variable could be contract rent or housing value in this application, while the study variable of interest would be some measure of consumer expenditures. The two variables

are assumed to be realized from an infinite superpopulation model in which

$$Y_j = \alpha + \beta X_j + \epsilon_j,$$

where $\mathrm{E}[\epsilon_j \mid X_j] = 0$ and $\mathrm{Var}(\epsilon_j \mid X_j) = v(X_j)$ for all $X_j$. Assume that the $\{\epsilon_j\}$ are independent of the domain sizes $\{M_i\}$.

It is desirable to stratify the population on $X$ at fixed break points $\{x_h\}$ into $H$ strata, $-\infty = x_0 < x \leq x_1, x_1 < x \leq x_2, \ldots, x_{H-1} < x < x_H = \infty$. Let $I_{hij} = 1$ if $x_{h-1} < X_j \leq x_h$ and $I_{hij} = 0$ otherwise. Define

$$N_h = \sum_{i=1}^{D} \sum_{j \subset U_i} I_{hij}.$$

*2.1. Two-Phase Sampling for Stratification*

If the $\{I_{hij}\}$ and hence the $\{N_h\}$ were known for all $h, i, j$, conventional stratification could be used. If the $\{I_{hij}\}$ were not known for all elements, but the $\{N_h\}$ were accurately known from some external source, then post-stratification of the realized sample could be employed. We consider the case in which neither the $\{I_{hij}\}$ nor the $\{N_h\}$ are known. In this situation, the usual remedy is to use two-phase sampling for stratification: the first phase, in which $X_j$ is measured, allows estimation of the $\{N_h\}$ and stratification of the sampled elements on $\{I_{hij}\}$; the second phase sample is then a stratified sample for which $Y_j$ is measured (e.g., Cochran, 1977, Chapter 12).

In our case, the first phase of sampling is the long form, which will be approximated as a simple random sample without replacement, with sampling rate $\lambda$. Let $A_{ij} = 1$ if element $j$ of domain $i$ is selected for the first phase sample and let $A_{ij} = 0$ otherwise.

*2.2. Selection Problems and Quasi-Random Sampling*

The first phase sample is subject to two kinds of selection problems. The first is nonresponse: some sampled elements fail to respond. The second will be called "conversion": some nonsampled elements are converted to respondents. This can occur in long form sampling in at least two ways: long forms can be delivered to the wrong addresses, or follow-ups can go to the wrong households. Such conversions can lead to bias problems if, for example, follow-up interviewers tend to replace non-vacant households with vacant households.

"Quasi-random" sampling mechanisms are used to describe both of these selection problems. Let $R_{ij} = 1$ if a response is obtained from element $j$ of domain $i$ and let $R_{ij} = 0$ otherwise. Given the first phase sample, we assume that the $\{R_{ij}\}$ are independently distributed as the Bernoulli mixtures

$$R_{ij} \mid A_{ij} \sim \mathrm{Bernoulli}(\pi_{ij}) A_{ij} + \mathrm{Bernoulli}(\kappa_{ij})(1 - A_{ij});$$

that is, if an element is selected in the first phase sample ($A_{ij} = 1$), it responds with probability $\pi_{ij}$; if an element is not selected in the first phase sample ($1 - A_{ij} = 1$), it is

converted to a respondent with probability $\kappa_{ij}$. To obtain an asymptotic mean squared error (mse) approximation, it is mathematically convenient to write

$$\pi_{ij} = \pi + \frac{\delta_{ij}^{\pi}}{\sqrt{D}}$$

and

$$\kappa_{ij} = \kappa + \frac{\delta_{ij}^{\kappa}}{\sqrt{D}}.$$

A stratified simple random subsample of overall size $m = \sum_{h=1}^{H} m_h$ is selected from the respondents in the first phase. There are $\sum_{i=1}^{D} \sum_{j \in U_i} I_{hij} R_{ij}$ respondents in stratum $h$. Define the $\{R_{ij}\}$ within stratum $h$ as the sample membership indicators of simple random sampling of size $m_h$ from $\sum_{i=1}^{D} \sum_{j \in U_i} I_{hij} R_{ij}$.

The selection problems described above lead to the possibility of design bias if estimation is based solely on the subsample from the respondents in the first phase. One alternative is to draw two subsamples: one from the set of first-phase respondents, and one from all other elements. The first subsample can be efficiently stratified on the $\{X_j\}$ as described above, but this is not possible for the second subsample.

### *2.3. Stratification on Small Domain Estimates*

Instead, stratification for the remaining set is based on small domain estimates of the mean of $X_j$. The $\{X_j\}$ are assumed to be realized from a small domain model, with normal random effects for the domains. (The normality assumption makes certain computations more analytically tractable; similar computations can be obtained numerically for the non-normal case.) If $j \in U_i$, then

$$X_j = \mu_X + \gamma_i + \eta_j,$$

where $\{\gamma_i\}$ are iid $N(0, \sigma_\gamma^2)$, independent of $\{\eta_j\}$ iid $(0, \sigma_\eta^2)$ (not necessarily normal). Assume that the $\{\eta_j\}$ are independent of the domain sizes $\{M_i\}$. Without loss of generality for the variance computations of interest, take $\alpha = 0$ and $\mu_X = 0$.

Further assume that there exist estimates $\{\hat{\gamma}_i\}$, based on long form samples of size $\{\lambda M_i\}$ $(0 < \lambda < 1)$, of the domain means $\{\gamma_i\}$. Assume that

$$\hat{\gamma}_i \mid \gamma_i, M_i \sim N\left(\gamma_i, \sigma_\eta^2/(\lambda M_i)\right),$$

independent of $\{\eta_j\}$ and independent of $\{\epsilon_j\}$. Because the $\{\hat{\gamma}_i\}$ will be sample means or closely-related estimators, the assumption of normality should hold at least approximately provided domain sample sizes are moderately large.

The population is stratified on $\{\hat{\gamma}_i\}$ at the fixed break points $\{x_h\}$ into $H$ strata; that is, all elements $j$ in domain $i$ are in stratum $h$ if

$$x_{h-1} < \hat{\gamma}_i \leq x_h,$$

where $x_0 = -\infty$ and $x_H = \infty$. Let

$$J_{hi} = \begin{cases} 1, & \text{if } x_{h-1} < \hat{\gamma}_i \leq x_h, \\ 0, & \text{otherwise.} \end{cases}$$

A stratified simple random subsample of overall size $n = \sum_{h=1}^{H} n_h$ is selected from the set of all elements which were not respondents in the first phase. There are $\sum_{i=1}^{D} \sum_{j \in \Pi_i} J_{hi} \bar{R}_{ij}$ elements in stratum $h$, where

$$\bar{R}_{ij} = 1 - R_{ij}.$$

Define the $\{C_{ij}\}$ within stratum $h$ as the sample membership indicators of simple random sampling of size $n_h$ from $\sum_{i=1}^{D} \sum_{j \in \Pi_i} J_{hi} \bar{R}_{ij}$.

## 3. Main Results

### *3.1. Asymptotic Mean Squared Error*

Under the sampling framework described in Section 2, we seek an asymptotic approximation to the design mse of the estimator

$$\frac{1 - \psi\hat{\rho}}{1 - \hat{\rho}} \sum_{h=1}^{H} \frac{\sum_{i=1}^{D} \sum_{j \in \Pi_i} J_{hi} \bar{R}_{ij}}{N} \frac{\sum_{i=1}^{D} \sum_{j \in \Pi_i} Y_j J_{hi} C_{ij} \bar{R}_{ij}}{n_h}$$
$$+ \psi \sum_{h=1}^{H} \frac{\sum_{i=1}^{D} \sum_{j \in \Pi_i} J_{hij} R_{ij}}{N} \frac{\sum_{i=1}^{D} \sum_{j \in \Pi_i} Y_j J_{hij} B_{ij} R_{ij}}{m_h},$$

where

$$\hat{\rho} = \frac{\sum_{i=1}^{D} \sum_{j \in \Pi_i} R_{ij}}{N}$$

and $0 \leq \psi \leq \hat{\rho}^{-1}$. The asymptotic mse approximation is obtained by approximating the squared design bias and design variance of the estimator above, normalizing by multiplying by $D$, and computing the probability limit (under the superpopulation model) of the normalized mse as $D \to \infty$. The choice of the weights $(1 - \psi\hat{\rho})(1 - \hat{\rho})^{-1}$ and $\psi$ insures that the design bias does not dominate asymptotically. Different choices of $\psi$ lead to different degrees of asymptotic variance and bias:

- $\psi = 0$ implies $(1 - \psi\hat{\rho})(1 - \hat{\rho})^{-1} = (1 - \hat{\rho})^{-1}$, so all weight is on the inefficiently stratified sample. The estimator is biased.
- $\psi = 1$ implies $(1 - \psi\hat{\rho})(1 - \hat{\rho})^{-1} = 1$, so the two components are equally weighted. The estimator is unbiased.
- $\psi = \hat{\rho}^{-1}$ implies $(1 - \psi\hat{\rho})(1 - \hat{\rho})^{-1} = 0$, so all weight is on the efficiently stratified sample. The estimator is biased.

Detailed derivations of the limiting, normalized mse are omitted. The result, after considerable simplification, is

$$
\begin{aligned}
D(\text{design mse}) \xrightarrow{a} & \left[\frac{1-\psi}{1-\rho}\left\{\lambda \mathrm{Cov}_M(Y_j,\delta^{\dagger}_{ij})+(1-\lambda)\mathrm{Cov}_M(Y_j,\delta^{*}_{ij})\right\}\right]^2 \\
& +\frac{(1-\psi)^2}{(1-\rho)^2 \mathrm{E}[M_i]}\mathrm{Var}_M(Y_j)\lambda(1-\lambda)(\pi-\kappa)^2 \\
& +\frac{(1-\psi)^2}{(1-\rho)^2 \mathrm{E}[M_i]}\mathrm{Var}_M(Y_j)\left\{\lambda\pi(1-\pi)+(1-\lambda)\kappa(1-\kappa)\right\} \\
& +\frac{(1-\psi\rho)^2}{1-\rho}\sum_{h=1}^{H}\frac{\omega^{*}_h}{\mathrm{E}[M_i]}\mathrm{Var}_{M,\Pi^{*}_h}(Y_j)\frac{1-\varphi^{*}_h}{\varphi^{*}_h} \\
& +\psi^2\rho\sum_{h=1}^{H}\frac{\omega_h}{\mathrm{E}[M_i]}\mathrm{Var}_{M,\Pi_h}(Y_j)\frac{1-\varphi_h}{\varphi_h}, \qquad (1)
\end{aligned}
$$

where

$$
\rho=\lambda\pi+(1-\lambda)\kappa,
$$

$$
\mathrm{Cov}_M(Y_j,\delta_{ij})=\frac{\mathrm{E}[M_i Y_j \delta_{ij}]}{\mathrm{E}[M_i]}-\frac{\mathrm{E}[M_i Y_j]}{\mathrm{E}[M_i]}\frac{\mathrm{E}[M_i \delta_{ij}]}{\mathrm{E}[M_i]},
$$

$$
\mathrm{Var}_M(Y_j)=\mathrm{Cov}_M(Y_j,Y_j),
$$

$$
\omega^{*}_h=\frac{\mathrm{E}[M_i I_{hi}]}{\mathrm{E}[M_i]},
$$

$$
\omega_h=\frac{\mathrm{E}[M_i I_{hij}]}{\mathrm{E}[M_i]},
$$

$$
\mathrm{Var}_{M,\Pi^{*}_h}(Y_j)=\frac{\mathrm{E}[M_i Y_j^2 I_{hi}]}{\mathrm{E}[M_i I_{hi}]}-\frac{\mathrm{E}^2[M_i Y_j I_{hi}]}{\mathrm{E}^2[M_i I_{hi}]},
$$

$$
\mathrm{Var}_{M,\Pi_h}(Y_j)=\frac{\mathrm{E}[M_i Y_j^2 I_{hij}]}{\mathrm{E}[M_i I_{hij}]}-\frac{\mathrm{E}^2[M_i Y_j I_{hij}]}{\mathrm{E}^2[M_i I_{hij}]},
$$

$$
\varphi^{*}_h=\frac{1}{(1-\rho)\mathrm{E}[M_i I_{hi}]}\lim_{D\to\infty}\frac{m_h}{D},
$$

and

$$
\varphi_h=\frac{1}{\rho\mathrm{E}[M_i I_{hij}]}\lim_{D\to\infty}\frac{m_h}{D}.
$$

It is assumed that $\varphi^{*}_h>0$ and $\varphi_h>0$ for all $h$. If the domain sizes are constant, $M_i\equiv\mu_M$, then $\mathrm{Cov}_M(Y_j,\delta_{ij})$ and $\mathrm{Var}_M(Y_j)$ reduce to the usual unconditional covariance and variance, while $\mathrm{Var}_{M,\Pi^{*}_h}(Y_j)$ and $\mathrm{Var}_{M,\Pi_h}(Y_j)$ reduce to within-stratum variances.

The result (1) can be rewritten as

$$
\begin{aligned}
\text{design mse} \;\simeq\; & \left[\frac{1-\psi}{1-\rho}\left\{\lambda \mathrm{Cov}_M(Y_j,\pi_{ij})+(1-\lambda)\mathrm{Cov}_M(Y_j,\kappa_{ij})\right\}\right]^2 \\
& +\frac{(1-\psi)^2}{N(1-\rho)^2}\mathrm{Var}_M(Y_j)\lambda(1-\lambda)(\pi-\kappa)^2 \\
& +\frac{(1-\psi)^2}{N(1-\rho)^2}\mathrm{Var}_M(Y_j)\left\{\lambda\pi(1-\pi)+(1-\lambda)\kappa(1-\kappa)\right\} \\
& +\frac{(1-\psi\rho)^2}{N(1-\rho)}\sum_{h=1}^{H}\omega_h^*\mathrm{Var}_{M,\Pi_h^*}(Y_j)\frac{1-\varphi_h^*}{\varphi_h^*} \\
& +\frac{\psi^2\rho}{N}\sum_{h=1}^{H}\omega_h\mathrm{Var}_{M,\Pi_h}(Y_j)\frac{1-\varphi_h}{\varphi_h}, \qquad (2)
\end{aligned}
$$

because $DE[M_i]N^{-1} \xrightarrow{p} 1$ as $D \to \infty$.

The first term in (2) is the squared bias, due to nonresponse and conversion. The second term is the variance contribution from simple random sampling in the first phase. The third term is the variance contribution from the quasi-random Bernoulli mixture sampling, and the final two terms are the variance contributions from the two kinds of stratified sampling in the last phase.

*3.2. Optimal Allocation*

We next consider the allocation of the sampling resources to the two subsamples. Assume that the overall sampling rate is fixed: $\varphi^\dagger = \rho\varphi + (1-\rho)\varphi^*$. Define

$$
V_{II}^* = \sum_{h=1}^{H}\frac{\omega_h^*}{(1-\rho)N}\mathrm{Var}_{M,\Pi_h^*}(Y_j)
$$

and

$$
V_{II} = \sum_{h=1}^{H}\frac{\omega_h}{\rho N}\mathrm{Var}_{M,\Pi_h}(Y_j).
$$

Assume proportional allocation within each subsample, so that $\varphi_h^* \equiv \varphi^*$ and $\varphi_h \equiv \varphi$. Then, by a standard constrained optimization using Lagrange multipliers, it can be shown that the optimal subsample sampling rates are

$$
\varphi^* = \frac{(1-\psi\rho)V_{II}^{*1/2}}{(1-\rho)^{1/2}\left\{\psi\rho^{3/2}V_{II}^{1/2}+(1-\rho)^{1/2}(1-\psi\rho)V_{II}^{*1/2}\right\}}\varphi^\dagger
$$

and

$$
\varphi = \frac{\psi\rho^{1/2}V_{II}^{1/2}}{\psi\rho^{3/2}V_{II}^{1/2}+(1-\rho)^{1/2}(1-\psi\rho)V_{II}^{*1/2}}\varphi^\dagger.
$$

*3.3. Optimal Weighting*

It remains to choose the optimal weighting constant, $\psi \in [0, 1/\rho]$. Assuming optimal allocation as given above, the optimal $\psi$ is given by

$$\psi_{\text{opt}} = \frac{V_7 - \Delta(1-\rho)^{1/2} V_{77}^{*-1/2}/\varphi^\dagger - \rho V_{77}^*}{V_7 + \Delta^2/\varphi^\dagger - \rho^2(V_{77}^* + V_{77})},$$

where

$$\begin{aligned} V_7 &= \left[\frac{1}{1-\rho}\left\{\lambda \text{Cov}_M(Y_j, \pi_{ij}) + (1-\lambda)\text{Cov}_M(Y_j, \kappa_{ij})\right\}\right]^2 \\ &+ \frac{1}{N(1-\rho)^2}\text{Var}_M(Y_j)\lambda(1-\lambda)(\pi-\kappa)^2 \\ &+ \frac{1}{N(1-\rho)^2}\text{Var}_M(Y_j)\left\{\lambda\pi(1-\pi) + (1-\lambda)\kappa(1-\kappa)\right\} \end{aligned}$$

and

$$\Delta = \rho^{3/2} V_{77}^{*-1/2} - \rho(1-\rho)^{1/2} V_{77}^{*-1/2}.$$

## 4. Numerical Example

Figure 1 shows asymptotic mean squared error as a function of $\psi$ for various degrees of selection bias. Stratum break points $\{x_h\}$ were chosen to match break points for coded value of owned home in the 1990 CE design: $(x_1, x_2, x_3, x_4) = (9, 16, 25, 35)$. The mean and variance for $X_j$ were computed as the grouped mean and variance from a frequency distribution of housing value (using an upper endpoint of 85 and treating $X_j$ as uniformly distributed on each of the resulting five intervals): $\mu_X = 19.055$ and $\sigma_X^2 = 316.6536$. The size of each of the $D = 200$ domains was set equal to a constant, $M_i \equiv 50$. Within each domain, the $X_j$ were distributed as dependent normal random variables with correlation $10/13$, and the $Y_j$ were modeled as

$$Y_j = \frac{1}{60}X_j + \epsilon_j,$$

with $\{\epsilon_j\}$ iid $N(0, (0.3952)^2)$. This implies that the correlation between $X_j$ and $Y_j$ is 0.6.

The first phase sample was simple random sampling with rate $\lambda = 1/6$, to reflect the long form sampling rate. Parameters of the selection mechanism were set at 0, 0.4, or 0.8 for the correlations $\text{Corr}(Y_j, \pi_{ij})$ and $\text{Corr}(Y_j, \kappa_{ij})$; 0.9 for the response rate, $\pi$; and 0.02 for the conversion rate, $\kappa$. The final phase of sampling consisted of two stratified subsamples, with overall sampling rate $\varphi^\dagger = 0.02$. Sampling resources were optimally allocated across the two subsamples, and proportionally allocated within each subsample.

In the absence of selection bias problems (i.e., $\text{Corr}(Y_j, \pi_{ij}) = 0$ and $\text{Corr}(Y_j, \kappa_{ij}) = 0$),

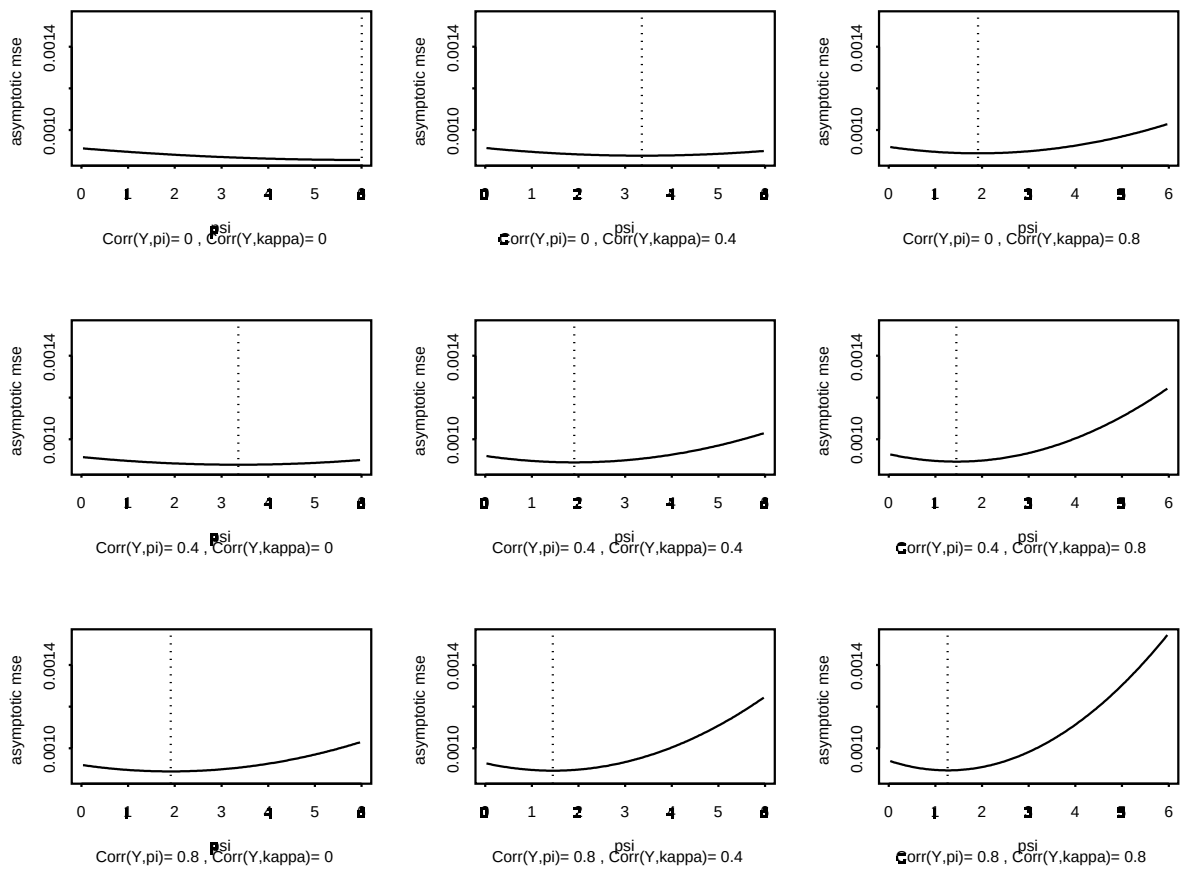


Figure 1: *Asymptotic mean squared error for the combined estimator with weighting parameter $\psi$. Vertical reference line is at the optimal value of $\psi$.*

the optimal procedure is to put all weight on the efficiently stratified subsample from the first phase ($\psi = \rho^{-1}$). As bias becomes more of a problem, the optimal procedure shifts toward the unbiased weighting ($\psi = 1$). The unbiased weighting procedure is never far from optimal in this example.

**Acknowledgments.** This work was conducted while the author was a senior research fellow at the U.S. Census Bureau and the Bureau of Labor Statistics, supported by the American Statistical Association and the National Science Foundation.

## References

Bureau of Labor Statistics. (1997). *BLS Handbook of Methods*, Bulletin 2490. U.S. Department of Labor, Washington, D.C.

Cochran, W.G. (1977). *Sampling Techniques*. 3rd ed. Wiley, New York.