Journal of Official Statistics, Vol. 26, No. 1, 2010, pp. 127–143

Evaluation and Selection of Models for Attrition Nonresponse Adjustment

Eric V. Slud[1,2] *and Leroy Bailey*[1]

This article considers a longitudinal survey like the U.S. Survey of Income and Program Participation (SIPP), with successive "waves" of data collection from sampled individuals, in which nonresponse attrition occurs and is treated by weighting adjustment, either through adjustment cells or a model like logistic regression in terms of auxiliary covariates. We measure the biases in estimated initial-wave (Wave 1) attribute totals between the survey-weighted estimator in the first wave and the weight-adjusted estimator for the same Wave 1 item total based on later-wave respondents. Three new metrics of quality are defined for models used to adjust a longitudinal survey for attrition. The metrics combine estimated between-wave adjustment biases based on subsets of the sample, relative to the estimated total, for various survey items. The maximum of the biases for estimated totals of a survey item is calculated from the weight-adjusted subtotal of the first j sample units, as j ranges from 1 to the size of the entire (Wave 1) sample, after a random re-ordering either of the whole sample or of the units within distinguished cells (which are then also randomly reordered); and the average over re-orderings of the maximal adjustment bias is divided by the estimated Wave 1 attribute total to give the metric value. Confidence bands for the metrics are estimated, and the metrics are applied to judge the quality of and to select among a collection of logistic-regression models for attrition nonresponse adjustment in SIPP 96.

Key words: adjustment cell; logistic regression; weighting; random re-ordering; raking; subdomain.

1. Introduction

To measure the quality of adjustment for attrition in a longitudinal survey, one would certainly try to evaluate the biases of adjustments using external data on the sample frame and the survey variables whenever such external data are available. But external data for evaluation seldom are available, so survey investigators must usually evaluate and choose among competing adjustment methods based on comparisons internal to the survey. Yet there is little published methodological work on how to measure the biases of adjustment from the internal evidence of a longitudinal survey.

[1] U.S. Census Bureau, SRD, 4600 Silver Hill Road, Washington, DC 20233-9100, U.S.A. Emails: Eric.V.Slud@census.gov, bailee47@aol.com
[2] University of Maryland, College Park, MD 20742, U.S.A. Email: evs@math.umd.edu
Acknowledgments: The authors are grateful to Sam Sae Ung of the U.S. Census Bureau for his comments, and for the suggestion to recalculate the metric values using weights which have been raked after model-based adjustment, to Pat Cantwell and Lynn Weidman for comments based on their thoughtful reading of the article, and to the JOS Associate Editor and referees for useful suggestions.
The views expressed in this article on statistical and methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.

There has been a great deal of work on calibrating, reconciling, and benchmarking time series of differing reporting periods and accuracies (Dagum and Cholette 2006). There has also been previous theoretical work on the large-sample behavior of model-based nonresponse weight adjustment methods. One recent example is Kim and Kim (2007); and these same authors, in an unpublished 2007 preprint, studied the problem of choosing between alternative parametric models for survey nonresponse using the same data on which the estimated parameters are applied to adjust the weights. But our literature search has yielded few papers explicitly assessing adjustment effectiveness using evidence only within the adjusted survey. A notable example is Eltinge and Yansaneh (1997), which discusses several diagnostics and sensitivity checks for the definition of weighting adjustment cells in a survey.

An important paper on internal longitudinal evaluation of nonresponse adjustment methods from a calibration perspective is Dufour et al. (2001). That paper specifically considers calibration adjustments, and assesses magnitudes of adjustment by a metric the authors define for tracking weight changes through several stages of a weight-adjusted longitudinal survey. By conducting a large simulation study within which they randomly subsample from a large longitudinal survey dataset (SLID, the Canadian Survey of Labor and Income Dynamics), Dufour et al. compare the weight-changes due to nonresponse weighting adjustments calculated by two model-based adjustment approaches (Logistic regression with stepwise variable selection and Response Homogeneity Groups – what we call below the adjustment-cell method – with cells defined using a CHAID-based Segmentation Model). Calibration (Deville and Särndal 1992) adjusts weights according to a model (of adjustment-cell or logistic-regression type) minimally subject to matching the estimated population totals in designated subsets exactly with the totals from an external study. Then the estimated adjustment biases (as in Bailey 2004 and as described below in connection with Slud and Bailey 2006) for population totals of other early-stage variables could be used to judge the overall success of the modelling approach used in adjustment. This could have been, but was not, done in Dufour et al. (2001), nor were effects of weighting adjustment on population subdomains examined.

Slud and Bailey (2006) studied the estimates and standard errors of differences, in the U.S. Survey of Income and Program Participation (SIPP), between Wave 1 totals of various 1996 cross-sectional survey items and the nonresponse-adjusted totals of the same Wave 1 items using response data from a later Wave (4 or 12) of the same survey. The nonresponse adjustments studied were either derived by an adjustment cell method or by parsimonious logistic regression models for the later-wave response indicators. Relative and standardized estimated biases were seen to vary considerably and somewhat erratically from one adjustment model to another. As in Dufour et al. (2001), many competing adjustment models could be defined, depending on which attribute variables would be used in constructing adjustment cells or as logistic regression predictors. Slud and Bailey (2006) noted that including Poverty as a predictor did have the effect, akin to raking, of making the sample-wide estimated Wave 1 total of Poverty particularly small. However, since that effect directly stems from the estimating sample-wide equation defining the logistic regression coefficients, they conjectured that this artificial effect would be removed by considering estimated Wave 1 bias within a number of different subdomains. They also raised the possibility of customizing the attrition adjustment model to remove between-wave adjustment biases as far as possible. This suggests creating a composite metric by combining the estimated between-wave adjustment biases for various survey items. However, to avoid the artificial effect described above of adjustments which may only be effective on the whole population, such a metric ought to incorporate the estimated biases on multiple subdomains of the population.

The primary goal of this research has been to devise metrics to aid in the comparison of different model-based methods of adjustment for nonresponse due to attrition, and thus to provide a basis for choosing among adjustment methods. Several earlier comparative investigations related to adjustment methods have been conducted, even within the SIPP survey structure, but they seem not to have resulted in clear advantage for any adjustment method over others. (See Rizzo et al. 1994 for example).

Our approach is to adjust weights using models for later-stage response, of adjustment-cell or logistic regression type, to calculate the later-stage survey estimates of first-stage totals of population and other survey variables. We then measure the biases of estimated late- versus early-stage weighted subtotals, for many different population subdomains. We define and study three related metrics for adjustment effectiveness, which combine the relative biases of specific survey variables over specified population subdomains. The ultimate objective is then to use the metrics to choose among adjustment models within SIPP 1996. In this research, as in Slud and Bailey (2006), adjustments are evaluated only for *attrition,* i.e., for between-wave losses of responders, but in practice Wave 1 adjusted for nonresponse is still only a surrogate for the full population. Our approach might also have been used to track correctness of attrition nonresponse between many pairs of waves. We have not done this, although a referee correctly points out the complication that in practice, the models which yield the best metric values between one pair of waves might be different from those which optimize metrics between a later pair of waves. We have also not attempted to track the magnitudes of weight adjustment changes across waves as Dufour et al. (2001) did, because our aim was to assess the effectiveness of adjustment models with respect to the best metrics of adjustment correctness available to us.

The article is organized as follows. Section 2 defines the metrics and presents bounds, theoretically valid as prediction bounds when the adjustment model is correct, which can be used in practice to flag inadequate weight adjustments. Section 3 applies the metrics to the comparison of a series of adjustment-cell or logistic-regression models which might have been used in attrition adjustment of the SIPP 1996 data, expanding on those studied in Slud and Bailey (2006). The adjusted SIPP 96 totals and metrics in Section 3 are based upon first-stage weights (incorporating nonresponse adjustments up to Wave 1) and model-based later-wave adjustments, but no population controls. However, raking or calibration to updated-census population totals in defined cells is done in practice whenever a weighting adjustment is applied to a large national longitudinal study like SIPP. So we present in Section 4 also the comparison among Wave 1 totals based on Wave 1 and later-wave adjusted weights which are then raked as was actually done in SIPP. Section 5 draws overall conclusions, both about the metrics studied and the consequences for model-based adjustment in SIPP.

2. Formal Development: Metrics and Bounds

We begin by formulating the survey design and nonresponse as a so-called *quasi-randomization model* (Oh and Scheuren 1983). Let S denote the sample of $n = |S|$ persons drawn from sampling frame $\mathcal{U}$, with known (or effectively adjusted) single inclusion probabilities $\{\pi_i\}_{i\in\mathcal{U}}$, and responding in Wave 1. For a series of cross-sectional survey measurements indexed by $k = 1, \ldots, K$, such as the $K = 11$ items studied by Bailey (2004) and Slud and Bailey (2006), denote by $y_i^{(k)}$ the Wave 1 item values and by $\mathbf{X}_i$ a vector of auxiliary variable values for all $i \in \mathcal{U}$. Let ρ_i denote individual response indicators (observed for all $i \in S$) in a specified later wave of the same survey, and let $p_i = \mathcal{P}(\rho_i = 1|S)$ denote the (unknown) conditional probabilities of individual response in that later wave. Let $\hat{p}_i = g(\mathbf{x}_i, \hat{\nu})$ denote estimators of these unknown probabilities derived (using a known function *g)* from a parametric model using auxiliary data $\mathbf{x}_i$, within which parameter-estimators $\hat{\nu}$ are obtained via estimating equations (Kim and Kim 2007). For any population attribute z_i, $i \in \mathcal{U}$, the frame-population total is denoted $t_z = \sum_{i\in\mathcal{U}} z_i$, and the corresponding Horvitz-Thompson estimator is $\hat{t}_z = \sum_{i\in S} z_i/\pi_i$.

For each survey item $y_i^{(k)}$, $i \in \mathcal{U}$, and the adjustment strategy embodied in the estimated response probabilities $\hat{p}_i$ define the estimated nonresponse bias for each population subdomain $\mathcal{D} \subset \mathcal{U}$ as

$$\hat{B}_k(\mathcal{D}) = \sum_{i\in\mathcal{D}\cap S} \left(\frac{\rho_i}{\hat{p}_i} - 1\right) y_i^{(k)}/\pi_i \tag{1}$$

In Slud and Bailey (2006) and earlier research of Bailey, the domain $\mathcal{D}$ was all of $\mathcal{U}$, and the quantity $\hat{B}_k(\mathcal{U})$ was interpreted as the difference between an adjusted estimator of $t_{y^{(k)}}$ using the data $(\rho_i, \rho_i y_i^{(k)}, x_i : i \in S)$ and the ordinary Horvitz-Thompson estimator $\hat{t}_{y^{(k)}}$, and was regarded as an estimator of attrition nonresponse bias due to the method of adjustment.

When new model-based adjustments are incorporated into the survey weighting for later waves, especially when model-terms related to specific survey items are introduced, our experience (Slud and Bailey 2006) suggests that the whole-population bias terms $\hat{\mathcal{B}}_k(\mathcal{U})$ may be reduced much more than biases on smaller domains $\mathcal{D}$. Our aim is to devise useful measures of maximum bias over many subdomains: good model-based adjustments ought to reduce biases over the interesting domains and likely cannot hope to reduce biases over all subdomains. To discourage models which correct simply for the biases of individual survey-item totals over the whole population, we compute the maximum of biases for the subdomains defined from consecutively indexed subsets either of the whole population, as in definition (2) below, or of each of a set of distinguished cells, as in (8). Since the sequential indexing of the sampled population is largely arbitrary, and is likely not related to any mechanism of selective nonresponse, we remove the effect of any particular indexing of sample (or of sample within the distinguished cells) by defining our metrics through averaging the maximum subsequence biases over all random reorderings of the sample.

2.1. Relative Subdomain Bias

We now propose a measure of the typical relative bias in estimating item totals over subdomains. The idea is to consider the largest value of absolute relative bias $B_k(\mathcal{D})/\hat{t}_{y^{(k)}}$ over a collection of different subsets $\mathcal{D} \subset \mathcal{U}$. We first randomly reorder the n elements of S, giving the new sequencing $\tau = (\tau(1), \tau(2), \ldots, \tau(n))$. The largest absolute bias in survey variable k over consecutively τ-indexed subdomains of S is

$$\max_{1 \le a \le b \le n} |\hat{B}_k(\{\tau(i) : a \le i \le b\})| \le 2 \cdot \max_{1 \le a \le n} |\hat{B}_k(\{\tau(1), \ldots, \tau(a)\})|$$

which follows immediately from the triangle inequality

$$|\hat{B}_k(\mathcal{D}_1)| \le |\hat{B}_k(\mathcal{D}_1 \cup \mathcal{D}_2)| + |\hat{B}_k(\mathcal{D}_2)|, \quad \mathcal{D}_1 = \{\tau(i) : a \le i \le b\},$$
$$\mathcal{D}_2 = \{\tau(i) : 1 \le i < a\}$$

To measure the overall relative bias in estimating item k totals over subdomains, we define

$$m_k = E_r\left(\max_{1 \le a \le n} |\hat{B}_k(\{\tau(1), \ldots, \tau(a)\})|\right)/\hat{t}_{y^{(k)}} \tag{2}$$

where the expectation is taken, for a fixed sample, over random permutations τ chosen equiprobably from the $n!$ permutations of the elements of S. The quantity m_k is smaller than the largest relative bias $|\hat{B}_k(\mathcal{D})|/\hat{t}_{y^{(k)}}$ over *all* subsets $C \subset \mathcal{U}$ – which is too large an estimate of error, and also too expensive to calculate – but does measure the expected magnitude of the worst relative bias from a subsequence in a typical (random) scanning order of the sampled population.

In settings where the relative estimated bias

$$\delta^{(k)} \equiv \hat{B}_k(\mathcal{U})/\hat{t}_{y^{(k)}} = \sum_{i \in S}\left(\frac{\rho_i}{\hat{p}_i} - 1\right)(y_i^{(k)}/\pi_i)/\hat{t}_{y^{(k)}} \tag{3}$$

is large, we will see below that m_k and its estimator $\hat{m}_k$ are not much different from $|\delta^{(k)}|$. However, if $\delta^{(k)}$ is small – which may be true for artificial reasons if the model used to define $\hat{p}_i$ prominently features the attributes $\{y_j^{(k)}, j \in S\}$ – then $\hat{m}_k$ may be much larger, since the model-fitting may not simultaneously adjust for weighted $y_i^{(k)}$ totals over arbitrary subsets of the sample. Definition (2) attempts to penalize models which directly adjust the population-wide total of an attribute.

The metric m_k depends only on the sample data. Although it is too complicated to evaluate exactly, it can be estimated well by evaluating its defining expectation over random permutations τ using a Monte Carlo simulation strategy. For each of a set $1, \ldots, R$ of indices r denoting Monte Carlo replicates, we define independent random permutations τ_r of the indices $i \in S$. For each $r, (\tau_r(j), 1 \le j \le n)$ is equiprobably chosen from the $n!$ possible reorderings of S, a choice which is easily implemented in a Monte Carlo simulation by defining a sample of independent Uniform(0,1) variates $\mathrm{V}_r = (\mathbf{V}_{ri}, i \in S)$ and letting $\tau_r(j)$ be the *antiranks,* i.e., the sequence of indices i of the V_{ri} observations written in increasing order. Then the estimator $\hat{m}_k$ of m_k defined in (2) is

$$\hat{m}_k = \frac{1}{R}\sum_{r=1}^{R} \max_{1\le j\le n} |\hat{B}_k(\{\tau_r(1), \ldots, \tau_r(j)\})| / \hat{t}_{y^{(k)}}$$

$$= \frac{1}{R\hat{t}_{y^{(k)}}}\sum_{r=1}^{R} \max_{0\le x\le 1}\left|\sum_{i\in S} I_{[V_{ri}\le x]}\left(\frac{\rho_i}{\hat{p}_i}-1\right)\frac{y_i^{(k)}}{\pi_i}\right| \tag{4}$$

A metric for nonresponse bias combining all survey variables indexed by $k = 1, \ldots, K$, could be

$$M \equiv K^{-1}\sum_{k=1}^{K} m_k = K^{-1}\sum_{k=1}^{K} \sup_{\mathcal{D}\subset S} |\hat{B}_k(\mathcal{D})| / \hat{t}_{y^{(k)}}, \quad \hat{M} = K^{-1}\sum_{k=1}^{K} \hat{m}_k \tag{5}$$

and a weighted-average definition would also make sense.

The quality of estimation of m_k in terms of $\hat{m}_k$, and relationships between these and $|\delta^{(k)}|$, are addressed in Section 2.3 below. We turn first to the modification of (2) and (4) to allow expected and estimated maximum absolute relative discrepancies with respect only to those random re-orderings which preserve distinguished cells of the population, such as the cells to which population totals would be raked or calibrated, or the population subdomains of particular interest to data users.

2.2. *Metrics for Subdomain Bias Over Distinguished Cells*

Most random permutations of the sample completely shatter any meaningful sample subdomains. Yet the idea behind raking or calibration is precisely that certain estimated subdomain totals – usually, the estimated population totals over the cells A_c of a specified geographic-demographic partition $\mathcal{U} = \cup_{c=1}^{C} A_c$ of the frame population – must be constrained equal to the benchmark subdomain totals (the controls) of a current (updated) census. For that reason, it makes sense to measure bias estimates $\hat{B}_k(A_c)$ over these cells, and to assume from now on that a partition $\mathcal{A}$ of $\mathcal{U}$ into cells A_c, $c = 1, \ldots, C$, has been fixed. The idea is to modify (2) so that the allowed permutations must maintain the membership of elements in each cell A_c.

One approach would be to aggregate cellwise biases into a *relative accumulated absolute bias*

$$m_k^{Cum} = \sum_{c=1}^{C} \omega_c^{(k)} |\hat{B}_k(A_c)| / \hat{t}_{y^{(k)}} \tag{6}$$

where $\omega_c^{(k)}(S)$ are a set of cell- and item-specific weights. This bias metric is very conservative, much larger than the relative bias numbers m_k, because it aggregates across cells the absolute biases over all cells, as though all domain totals in all cells could be simultaneously biased in the same direction. However, we will look at this metric in Table 6 below to see what it tells about choosing between adjustment models in SIPP 1996.

A less extreme modification of (2) is to restrict the permutations τ in the expectation – now denoting them by σ to reflect the new constraint – so that $\{\sigma(i) : i \in A_c \cap S\} = A_c \cap S$, i.e., the units are reordered within each sample block $A_c \cap S$ but not mingled across blocks. To explain this invariance, assume that the sample S is indexed in such a way that the $n_c \equiv |A_c \cap S|$ sampled elements in the c'th cell A_c appear as elements consecutively numbered from $\sum_{j=1}^{c-1} n_j + 1$ to $\sum_{j=1}^{c} n_j$ in the enumerated sample S. The invariance of the cells under σ means that

$$\begin{aligned} &\text{for all} \quad 1 \le c \le C \quad \text{and} \quad i \in A_c \cap S, \\ &\quad n_1 + n_2 + \ldots + n_{c-1} < \sigma(i) \le n_1 + n_2 + \ldots + n_c \end{aligned} \tag{7}$$

Now the allowed random permutations σ of the sample elements are chosen equiprobably from the $\prod_{c=1}^{C} n_c!$ permutations σ of $\{1, 2, \ldots, n\}$ which satisfy (7). Finally, our modfied metric is defined as the expectation over σ of the maximum absolute cumulative weighted sum of cellwise biases relative to $\hat{t}_{y^{(k)}}$ as follows:

$$\begin{aligned} m_k^* &\equiv E_\sigma\left(\max_{1 \le q \le n} |\hat{B}_k(\{\sigma(1), \ldots, \sigma(q)\})| \right) / \hat{t}_{y^{(k)}} \\ &= E_\sigma\left(\max_{1 \le c \le C, q \in A_c} \left| \sum_{l=1}^{c-1} \hat{B}_k(A_l) + \hat{B}_k(\{\tau(a) : a \in A_c, a \le q\}) \right| \right) \end{aligned} \tag{8}$$

An estimator for the modified quantity (8) can be implemented in terms of a collection of random batches $\mathbf{V}_r$ of n independent Uniform(0,1) random variates. The r'th reordering of the elements i within the reordered block $A_c \cap S$, is given by the indices in the range given by (7) of the variates $(V_{ir}, i \in A_c \cap S)$ ordered from smallest to largest. Then the estimator for (8) becomes

$$\hat{m}_k^* \equiv (R\hat{t}_{y^{(k)}})^{-1} \sum_{r=1}^{R} \max_{1 \le c \le C} \max_{q \in A_c} \left| \sum_{l-1}^{q-1} \hat{B}_k(A_l) + \hat{B}_k(\{i \in A_c \cap S, V_{ri} \le V_{rq}\}) \right| \tag{9}$$

or equivalently,

$$\hat{m}_k^* \equiv (R\hat{t}_{y^{(k)}})^{-1} \sum_{r=1}^{R} \max_{1 \le c \le C} \max_{q \in A_c} \left| \sum_{\{i \le n_1 + \ldots + n_{c-1}\} \cup \{i \in A_c \cap S : V_{ri} \le V_{rq}\}} \left(\frac{\rho_i}{\hat{p}_i} - 1 \right) \frac{y_i^{(k)}}{\pi_i} \right| \tag{10}$$

We next place confidence bounds on the differences between the quantities m_k, m_k^* and their estimates $\hat{m}_k, \hat{m}_k^*$ and on the differences between these quantities and $|\delta^{(k)}|$.

2.3. *Confidence Intervals and Bounds for* $\mathrm{m_k}$ *and* $\mathrm{m_k^*}$

All of the quantities m_k, m_k^* are functions of the sampled survey data, and the probability statements made at this stage concern only the chance element introduced by the random variates $\mathbf{V}_r$ used in defining (4) and (9), *conditionally given the sample.* At the end of the Section, we interpret the meaning of sample-based metric-estimators $\hat{m}_k$ for the survey population and adjustment model.

We begin with the simplest and clearest confidence statement. Since $\hat{m}_k$ is calculated as the empirical average over quantities calculated from a series of R random permutations of the sample, its sampling variability due to those permutations can be assessed by empirical standard errors

$$se(\hat{m}_k) = \frac{1}{|\hat{t}_{y^{(k)}}|}\left[\frac{1}{R(R-1)}\sum_{c=1}^{R}\left(\max_{0<x\leq 1}\left|\sum_{i\in S} I_{[V_{ci}\leq x]}\left(\frac{\rho_i}{\hat{p}_i}-1\right)\frac{y_i^{(k)}}{\pi_i}\right| - \hat{t}_{y^{(k)}}\hat{m}_k\right)^2\right]^{1/2}$$

Thus, with approximate 99% confidence when R is large,

$$|m_k - \hat{m}_k| \leq 2.576 \cdot se(\hat{m}_k) \tag{11}$$

the right-hand side being asymptotically, for large R, proportional to $1/\sqrt{R}$. Similar confidence statements with respect to the randomness of the permutations σ_r can be given bounding $m_k^* - \hat{m}_k^*$.

The difference between the metric value m_k and the overall relative bias $|\delta^{(k)}|$ is due to the fluctuations with varying $x \in [0,1]$ of the quantities

$$Z_k(x) \equiv Z_{kr}(x) = \sum_{i\in S} I_{[V_{ri}\leq x]}\left(\frac{\rho_i}{\hat{p}_i}-1\right)(y_i^{(k)}/\pi_i) \tag{12}$$

being maximized in (4), for fixed r, where V_{ri} are independent identically distributed Uniform(0,1) variates, and where $Z_k(1)/\hat{t}_{y^{(k)}}$ is by definition equal to $\delta^{(k)}$ given in (3). If the quantities $Z_k(x)$ were replaced by their expectations (i.e., if $I_{[V_{ri}\leq x]}$ were replaced by x), then expression (4) would become $|\delta^{(k)}|$. Thus, the discrepancy $\hat{m}_k - |\delta^{(k)}|$ can be bounded by the maximum absolute value of the random *weighted empirical* process indexed by a continuous argument $x \in [0,1]$,

$$\beta_n^{(k)}(x) = \frac{1}{\sqrt{n}}\sum_{i\in S}(I_{[V_{ri}\leq x]} - x)\left(\frac{\rho_i}{\hat{p}_i}-1\right)\frac{y_i^{(k)}}{\pi_i} = \frac{1}{\sqrt{n}}(Z_k(x) - x\delta^{(k)}\hat{t}_{y^{(k)}}) \tag{13}$$

conditionally given all sample data $\{i, \rho_i, x_i, (\rho_i y_i^{(k)}, 1 \leq k \leq K) : i \in S\}$, with only the variates $V_{ri}, i \in S$, regarded as random. The process $\beta_n^{(k)}(\cdot)$ has mean 0, and according to a slight extension of the Donsker Theorem (Pollard 1980), has approximate distribution for large n the same as

$$\sqrt{\gamma^{(k)}}W^\circ(x) \equiv \left[n^{-1}\sum_{i\in S}\left(\frac{\rho_i}{\hat{p}_i}-1\right)^2(y_i^{(k)}/\pi_i)^2\right]^{1/2} W^\circ(x) \tag{14}$$

as a random continuous function of $x \in [0,1]$, where $W^\circ(x)$ denotes a *tied-down Wiener process* or Gaussian process with mean 0 and

$$Cov(W^\circ(v), W^\circ(u)) = \min(v,u) - v \cdot u$$

The scaling constants in (14) governing the amplitude of fluctuations of $\beta_k(\cdot)$,

$$\gamma^{(k)} = n^{-1} \sum_{i \in S} \left(\frac{\rho_i}{\hat{p}_i} - 1 \right)^2 (y_i^{(k)}/\pi_i)^2 \tag{15}$$

are readily computed from the sample data, and under general assumptions are bounded for large n. By definition of m_k and the remark that $Z_k(1) = \delta^{(k)}/\hat{t}_{y^{(k)}}$ in (12) and (3),

$$\begin{aligned} |m_k - |\delta^{(k)}|| &= m_k - |\delta^{(k)}| \\ &\leq \frac{1}{\hat{t}_{y^{(k)}}} E_V \left(\max_{0<x\leq 1} \left| \sum_{i \in S} (I_{[V_{i \leq x}]} - x) \left(\frac{\rho_i}{\hat{p}_i} - 1 \right) (y_i^{(k)}/\pi_i) \right| \right) \\ &= (\sqrt{n}/\hat{t}_{y^{(k)}}) E_V \left(\max_{0<x\leq 1} |\beta_k(x)| \right) \approx 1.2286 \sqrt{n} \left(\sqrt{\gamma^{(k)}}/\hat{t}_{y^{(k)}} \right) \end{aligned} \tag{16}$$

since 1.2286 is the expectation of $\sup_{x \in [0,1]} |W^\circ(x)|$ which arises in calculating percentage points of the one-sample Kolmogorov-Smirnoff statistic, and is readily calculated using the density of this random variable given by Kolmogorov and reproduced by Feller (1948).

A similar argument, using the representation (10), proves that $m_k^* - |\delta^{(k)}|$ is positive and bounded above by the same quantity on the right-hand side of (16).

For specific items k, we find when n is large that for moderate numbers R of random permutations, the difference $\hat{m}_k - m_k$ (or $\hat{m}_k^* - m_k^*$) is generally very small compared to m_k (respectively m_k^*). Then, by calculating the right-hand side of (16), we find roughly how small the value $\hat{m}_k$ must be in order that the sample data be compatible with a zero relative bias $\delta^{(k)}$. The objective of this kind of analysis is first of all to flag as "inadequately adjusted" those items for which model-based attrition nonresponse adjustment has resulted in estimated metric values $\hat{m}_k$ greater than the sum of the right-hand sides of (11) and of (16). Since we will find generally that the values of $\hat{m}_k^*$ and $\hat{m}_k$ are roughly the same, we will use the same threshold for metric values $\hat{m}_k^*$.

We can now interpret the bounds (16) and (11). First, we have seen that it is easy to choose R large enough so that the right-hand side of (11) is much smaller than that of (16), which implies that we can essentially disregard the difference between m_k and $\hat{m}_k$ (or between m_k^* and $\hat{m}_k^*$). Second, (16) tells us that when one of three quantities m_k or m_k^* or $|\delta^{(k)}|$ is much larger than the bound in (16), then all three will be, indicating that this item k has been badly adjusted (for the Wave and model under consideration). Third, when the quantities m_k, m_k^*, $|\delta^{(k)}|$ are of the same or smaller size compared with the bound in (16), the quantity m_k^* gives the quality of adjustment under a meaningful metric which takes account of various population subdomains including all of the distinguished cells A_c.

Next, we compare the estimated metric values $\hat{m}_k$ and $\hat{m}_k^*$, individually or aggregated as in (5), across different adjustment models in order to choose a "best" model in a specific survey application.

3. Adjustment Metric Values in SIPP 96

For the case of the U.S. SIPP 1996 survey, with $K = 11$ cross-sectional items, response probabilities $\hat{p}_i$ were estimated by the specific adjustment-cell and logistic-regression models mentioned above, all as described in detail by Slud and Bailey (2006). Briefly, the cross-sectional survey items $y_i^{(k)}$ studied are: indicators that the individual lives in a Household which receives (i) Food Stamps (Foodst), or (ii) Aid to Families with Dependent Children (AFDC); or indicators that the individual receives (iii) Medicaid (Mdcd), or (iv) Social Security (SocSec); and indicators that the individual (v) has health insurance (Heins), (vi) is in poverty (Pov), (vii) is employed (Emp), (viii) is unemployed (UnEmp), (ix) is not in the labor force (NILF), (x) is married (MAR), or (xi) is divorced (DIV).

In this data example, nonresponse is adjusted in one of two ways: either using a SIPP adjustment-cell model based on 149 standard cells (Tupek 2002) defined in terms of variables including age, sex, race, Hispanic origin, educational level, and labor force status; or using one of a series of logistic regression models A–III summarized in Table 1. (Of these models Model A and B were the ones used in Slud and Bailey 2006). The models C–E were selected to have progressively better fit, using an indicator of Wave 4 response as response-variable within the 94,444 SIPP Wave 1 sample records with positive base-weights. The variables used in these regression models include race, Hispanic origin, Renter versus Owner of housing unit, indicator that individual is the Household Reference Person, indicator of College education, a 4-category variable of Family type used in other raking-adjustment cells defined for use in SIPP by Tupek (2002), U.S. Census region, an indicator of ownership of Assets, plus some or all of the 11 SIPP survey items listed above.

*Table 1. Logistic regression models used to adjust Wave 4 or Wave 12 nonresponse in SIPP 96. Df is the number of independent coefficients in each model, including Intercept, and Dev the deviance for the 94,444-record SIPP 96 sample data in Wave 4. AIC is equal to Dev +2*Df*

Model	Df	Variables	Dev	AIC
A	8	Wnotsp Renter College RefPer Black Renter*College Black*College	76,558	76,574
B	9	same as A, plus Pov	76,545	76,563
C	14	same as B, plus Foodst Mdcd Heins UnEmp Div	76,299	76,327
D	14	same as B, minus Black*College plus Mdcd Heins UnEmp Pov*Heins Mdcd*Heins Heins*College	76,242	76,270
E	18	same as D, plus hisp + Famtyp	76,017	76,053
F	18	same as C, plus Afdc SocSec Emp Mar	76,280	76,316
O	10	College Assets hisp Region Famtyp	66,285	66,305
I	20	same as O, plus Renter Refper Pov Mdcd Heins UnEmp Foodst SocSec Mar Renter*College	65,678	65,718
II	22	same as I, plus Pov*Reg3 Pov*SocSec	65,652	65,696
III	31	same as II, plus AFDC NILF Div Wnotsp Black Black*College Pov*Heins UnEmp*Assets Heins*College	65,638	65,700

As can be seen by the dramatic decrease in deviance in the models O-III by comparison with the previous models, the binary variable Assets turns out to be by far the single strongest variable: 97% of individuals [in Wave 1 responding households] with Assets $= 1$ responded in Wave 4, while only 76% of individual with Assets $= 2$ did so. Forward and backward selection among models involving Assets resulted in the progressively better models I, II, and III. (In these models as defined in the table, Reg3 denotes the indicator of the third of four U.S. Census regions). For purposes of comparison, the "model" which treats nonresponse probabilities as constant but otherwise unconstrained within each of the 149 SIPP adjustment cells is calculated to have deviance 76,117 based on 149 degrees of freedom.

The method of model selection followed in the remaining data analysis of this section, as described and justified in the previous section, is to search for models and items with few metric values $\hat{m}_k, \hat{m}_k^*$, which are large compared to the bounds obtained by adding the right-hand sides of (11) and (16). This contrasts with the approach of Slud and Bailey (2006) who studied models A and B and, in the present notation, compared estimated population-wide adjustment biases $\delta^{(k)}$ with their design-based standard errors as found by a Balanced Repeated Replication method.

Calculations of $\hat{m}_k$ have been made with $R = 100$ random-permutation Monte Carlo replications, with the results for Model B presented in Table 2. (Because $n = 94,444$ is so large, the between-replication differences are small and this choice of R is ample). The final columns of Table 2 respectively display the bounds $b_{4,k}$, $b_{12,k}$ on the right-hand sides of (16) (which turn out to be virtually identical for the adjustment-cell and logistic-regression adjustment methods) for adjustments of Waves 4 and 12 nonresponse. It also turns out that for all items and combinations 4C, 4L, 12C, and 12L, the bounds on the right-hand side of (11) range from 1-5% of the corresponding bounds (16). The analogous table with logistic models A and D and F, also calculated with $R = 100$ iterations, is displayed as Table 3. However, the columns of bounds $b_{4,k}$, $b_{12,k}$ are included in the latter

Table 2. Quantities $\hat{m}_k$ in (4) estimated from SIPP96 data, for later-wave nonresponse adjustment either to Wave 4 or 12, and by either the Adjustment-Cell (C) or Logistic-Regression (L) method (Model B) and based on $R = 100$ replications. The last two columns are the bounds in (16), with $a = .01$

Item	$\hat{m}^{4C}$	$\hat{m}^{4L}$	$\hat{m}^{12C}$	$\hat{m}^{12L}$	$b_{4,k}$	$b_{12,k}$
Foodst	.0052	.0186	.0442	.0130	.0056	.0123
AFDC	.0067	.0248	.1040	.0350	.0078	.0173
Mdcd	.0066	.0279	.0163	.0426	.0053	.0119
SocSec	.0191	.0116	.1118	.1038	.0041	.0086
Heins	.0085	.0065	.0197	.0133	.0019	.0040
Pov	.0187	.0033	.0372	.0091	.0047	.0097
Emp	.0016	.0017	.0082	.0122	.0020	.0041
UnEmp	.0534	.0594	.1176	.1280	.0131	.0250
NILF	.0032	.0034	.0333	.0462	.0033	.0069
MAR	.0111	.0018	.0508	.0226	.0025	.0051
DIV	.0124	.0201	.0235	.0390	.0067	.0133

Table 3. Quantities $\hat{m}_k$ estimated from SIPP96 data based on $R = 100$ random permutations, for wave 4 or 12 nonresponse adjustment by logistic regression model A (first two columns) or model D (next two columns). The last two columns are the bounds $b_{4,k}$, $b_{12,k}$ from (16) using Model D

Item	$\hat{m}^{4,A}$	$\hat{m}^{12,A}$	$\hat{m}^{4,D}$	$\hat{m}^{12,D}$	$\hat{m}^{4,F}$	$\hat{m}^{12,F}$	$b^D_{4,k}$	$b^D_{12,k}$
Foodst	.0120	.0086	.0076	.0110	.0039	.0093	.0056	.0123
AFDC	.0175	.0446	.0067	.0624	.0053	.0134	.0077	.0170
Mdcd	.0219	.0346	.0035	.0078	.0037	.0084	.0052	.0114
SocSec	.0117	.1040	.0125	.1066	.0027	.0073	.0041	.0086
Heins	.0076	.0148	.0013	.0027	.0012	.0028	.0019	.0039
Pov	.0123	.0127	.0032	.0074	.0032	.0085	.0047	.0098
Emp	.0021	.0116	.0015	.0161	.0014	.0034	.0020	.0041
UnEmp	.0626	.1322	.0095	.0207	.0098	.0184	.0139	.0288
NILF	.0026	.0447	.0029	.0456	.0023	.0063	.0033	.0069
MAR	.0023	.0236	.0018	.0213	.0017	.0037	.0025	.0051
DIV	.0201	.0390	.0168	.0334	.0048	.0098	.0068	.0139

table only for model D, because the bounds for the other models are virtually identical with these, and again the bounds from (11) are only a few percent of the bounds (16).

Inspection of Tables 2 and 3 reveals that the metric $\hat{m}_k$ with very few exceptions in Wave 12 clearly exceeds the corresponding bounds b_k for the adjustment-cell model and for all of the logistic regression models. One notable exception is Pov, where as seen by Slud and Bailey (2006), Model B includes Pov as a predictor and does adjust effectively both in Waves 4 and 12. Similarly, we see that Model D, which includes variables Pov, Mdcd, Heins, and UnEmp as predictors, does a particularly good job of adjusting the totals of these same variables as measured by the metric $\hat{m}_k$. Indeed, the most striking preliminary conclusion from examining the tables of metric values under these various logistic regression models is that including a variable as a predictor generally results in

Table 4. Estimated metric values $\hat{m}^*_k$ defined in (9) for Wave 4 adjustment (except for Wave 12 in last row) with $R = 100$, based on the Adjustment cell and logistic regression models, using SIPP 96 data with demographic (raking) cells as partition elements A_c. Final two rows of the Table give metric $\hat{m}^*_k$ values averaged over Items k

Item	ModA	ModC	ModD	ModF	ModI	ModIII	Adj.Cell
Fdst	.0126	.0057	.0084	.0057	.0050	.0051	.0060
AFDC	.0179	.0065	.0075	.0064	.0063	.0059	.0072
Mdcd	.0220	.0044	.0042	.0043	.0046	.0043	.0069
SocS	.0118	.0115	.0127	.0038	.0048	.0045	.0192
Hins	.0078	.0021	.0022	.0020	.0027	.0026	.0086
Pov	.0132	.0050	.0052	.0051	.0049	.0049	.0191
Emp	.0030	.0026	.0028	.0026	.0024	.0022	.0023
UnEmp	.0627	.0097	.0092	.0098	.0103	.0102	.0535
NILF	.0032	.0032	.0035	.0030	.0033	.0031	.0039
MAR	.0028	.0025	.0025	.0025	.0038	.0035	.0112
DIV	.0205	.0057	.0170	.0056	.0055	.0054	.0129
POP	.0023	.0022	.0023	.0022	.0020	.0019	.0020
Wav4.Avg	.0150	.0051	.0065	.0044	.0046	.0045	.0127
Wav12.Av	.0411	.0273	.0304	.0110	.0163	.0113	.0492

very good adjustment as measured either by metric $\hat{m}_k$ or $\hat{m}_k^*$. This is true even under Model F, where we can see from Table 1 that the last batch of variables entered between Models D and F did not seem very important as measured by an increase in maximized loglikelihood, or equivalently a decrease in Deviance.

Recall that we devised the metrics $\hat{m}_k, \hat{m}_k^*$, in part to penalize model-based adjustment which, like raking, removes bias directly in terms of population totals. Recall also that $\hat{m}_k^*$ differed only by finding maximum absolute discrepancies over consecutive sequences of reordered indices which keep distinguished cells consecutively indexed. (In our computations, the distinguished cells used in the metrics were not the 149 nonresponse adjustment cells, but rather a system of 101 cells defined by Sex, Age-intervals, and Race, which are used by SIPP in raking to demographic population totals). In fact, the metric values $\hat{m}_k^*$ turn out to be only slightly larger than $\hat{m}_k$, and they follow a very similar pattern across the different models. Consider Table 5 charting the progression of averaged $\hat{m}_k$ metrics (over $k = 1, \ldots, 11$ and Population Count) as the adjustment model varies over the Adjustment Cell model and the ten Logistic models in Table 1, computed as in (5), with equal weights $w_k = 1/12$. The logistic regression models are all, except for Model O, clearly better than the cell-based model in adjusting at Wave 12, but at Wave 4, Models A and B actually seem a little worse and O seems much worse than the cell-based adjustment method. Since Models A–E are listed in order of decreasing Deviance or AIC, and Models O-III are much better than the first six logistic models from this viewpoint, there is no strict relationship between decreasing AIC and decreasing $\hat{M}$. Model F looks to be the clearly best adjustment model at both waves in Table 5, although Models I–III are strong competitors and would be preferred from examination of deviances. The metric $\hat{M}$ seems to reward Model F for including many of the SIPP items as predictors, and the much lower-deviance models I–III which incorporate some of the same predictor variables used to form the raking cells used as distinguished evaluation cells A_c are not rewarded for their predictive accuracy by the metrics m_k and m_k^*.

Table 5. Metric values $\hat{m}_k$ calculated on SIPP 96 data for Adjustment-cell model and for logistic regression models A-F and averaged over $k = 1, \ldots, 12$, where item 12 is Population Count ($y_i^{(12)} \equiv 1$)

Model	Wave 4	Wave 12
Adj.Cell	0.01228	0.04741
LReg, A	0.01451	0.03942
LReg, B	0.01504	0.03893
LReg, C	0.00426	0.02475
LReg, D	0.00571	0.02812
LReg, E	0.00481	0.02654
LReg, F	0.00342	0.00782
LReg, O	0.03078	0.05269
LReg, I	0.00393	0.01371
LReg, II	0.00393	0.01363
LReg, III	0.00392	0.00880

Table 6. Relative accumulated absoluted bias values $\hat{m}_k^{Cum}$ as in (6), averaged over $k = 1, \ldots, 12$, for each of 6 logistic regression models and the Adjustment Cell model, for each of Waves 4 and 12, using SIPP 96 data with distinguished demographic (raking) cells as partition elements A_c

	ModA	ModC	ModD	ModF	ModI	ModIII	Adj.Cell
Wave 4	.0483	.0444	.0448	.0439	.0462	.0460	.0448
Wave 12	.1226	.1198	.1203	.1061	.1119	.1080	.1307

Although we would not have chosen Model F from likelihood considerations, this model may be an excellent choice from the vantage point of nonresponse adjustment. The SIPP dataset is large enough ($n = 94,444$) that all of the SIPP survey items except AFDC and Emp have highly significant coefficients. Moreover, the parametric adjustment models F, II, and III are accomplishing something that the adjustment cell model cannot: they are generating response probabilities with good behavior over raking cells considered as subdomains. To see this more clearly, consider the less forgiving metric m_k^* defined in (8): for each item k, the absolute estimated biases $|\hat{B}_k(\sigma(1), \ldots, \sigma(q))|$ are accumulated over consecutive subsets of the sample dataset which have been randomly reordered in such a way as to leave distinguished demographic cells intact, and m_k^* is the ratio of this absolute total to $\hat{t}_{y^{(k)}}$. The result on the SIPP 96 data is given in Table 4. The logistic regression models, especially F and III, do better, item by item, with far fewer parameters than the 101 distinguished-cell response fractions

$$\sum_{i \in A_c \cap S} (r_i / \pi_i) / \sum_{i \in A_c \cap S} (1/\pi_i)$$

(estimates of the cellwise response probabilities) by which the adjustment cell model divides the survey weights in cell A_c. Model F is still the overall best choice, with III a close second, and this conclusion becomes stronger when we find it confirmed in Table 6 also by the relative total absolute bias which simply sums cellwise absolute biases without any reordering of sample items.

4. Models and Metrics Using Raked SIPP Weights

So far, metrics have been calculated and models evaluated based on direct adjustments using substituted estimates of response probabilities, with weights changing according to

Table 7. Estimated Metric Values $\hat{m}_k$, $\hat{m}_k^*$, and m_k^{Cum} based on SIPP 1996 model-adjusted weights which have been raked as in SIPP production. Partition elements A_c in the latter two metrics are closely related to the raking cells. Displayed metric values for Waves 4 and 12 for selected models have been averaged over k

Metric	Wave	ModA	ModC	ModD	ModF	ModI	ModII	Adj.Cell
m_k	4	.0099	.0057	.0047	.0057	.0046	.0045	.0083
	12	.0181	.0157	.0163	.0120	.0166	.0113	.0181
m_k^*	4	.0099	.0056	.0046	.0056	.0046	.0045	.0082
	12	.0183	.0159	.0164	.0121	.0165	.0113	.0182
m_k^{Cum}	4	.0392	.0382	.0380	.0381	.0414	.0411	.0377
	12	.0887	.0895	.0890	.0881	.0910	.0895	.0864

Table 8. Estimated metric scores $\hat{m}_k$ and $\hat{m}_k^*$ from SIPP 96 data, averaged over survey items k, for Unraked and Raked adjusted weights formed using selected Logistic-regression (D,F,I,III) models and adjustment-cell model

Metric	Raked	Wave	ModD	ModF	ModI	ModIII	Adj.Cell
m_k	No	4	.0057	.0034	.0039	.0039	.0123
		12	.0281	.0078	.0137	.0088	.0474
	Yes	4	.0047	.0057	.0046	.0045	.0083
		12	.0163	.0120	.0166	.0113	.0181
m_k^*	No	4	.0065	.0044	.0046	.0045	.0127
		12	.0304	.0110	.0163	.0113	.0493
	Yes	4	.0046	.0056	.0046	.0045	.0082
		12	.0164	.0121	.0165	.0113	.0182

the rule $1/\pi_i \mapsto \rho_i/(\hat{p}_i\pi_i)$. In practice, within government surveys such as SIPP, weights are adjusted and then raked so that population totals over certain demographic cells match the totals found through other, more accurate, censuses and surveys. For this reason, we recalculated the metrics for the cell-based adjustment model and the models displayed in Table 1 based on adjusted weights which were put through a final stage of raking. The raking method was as described by Tupek (2002) and implemented in the 1996 and 2001 SIPP panels, based on cells defined in terms of sex, age, race, family structure and Hispanic origin.

Summary results, averaged over survey items k (11 items plus population count), are presented in Table 7 for selected models and the three metrics m_k, m_k^*, and m_k^{Cum}. The effect of raking can be assessed by comparing these averaged post-raking metric values with the corresponding unraked-weight results in Tables 4–6. When the weights have been raked, the performance of the different models is not nearly so easy to distinguish as without raking. Thus, Models C, D, F, I, and III have very nearly the same performance, with respect to each of the three metrics m_k, m_k^*, m_k^{Cum}. With respect to the first two metrics, the best model (the only one that does well for both Wave 4 and 12) is Model III, and all of the good models (C,D,F, I, III in the table) clearly outperform the cell-adjustment model. However, with respect to m_k^{Cum}, no model outperforms simple cell-based adjustment after raking, although Model F comes close.

What do the tables tell us about the impact of raking, if extensive model-based adjustment is to be done? From the point of view of the metric $\hat{m}_k^{Cum}$, it is clearly better to rake and not to adjust via logistic-regression models. But this is largely because of the slightly artificial decision to define distinguished cells A_c extremely close to the cells used for population-control raking. However, for the other two metrics, we provide the comparison between model-based adjustments with and without raking in Table 8. According to the metrics m_k and m_k^*, with the best models (F,I,III) it hardly matters whether raking is done or not, but there is no real benefit and may even be some loss in adjustment accuracy, especially for Models F and III.

5. Conclusions

This article has developed metrics for nonresponse-adjustment effectiveness, calculated after randomly re-indexing the survey sample and calculating maximum discrepancies over consecutively indexed subdomains. The objective was to discount any advantage which an adjustment regression model might achieve toward eliminating whole-sample nonresponse biases by including survey attributes as predictors. That is why many of the logistic regression models we used to adjust for attrition in SIPP included some or many of the survey items. However, when applied to SIPP 96 data, the metrics developed did not have the expected effect. We expect AIC-based model assessments generally to prefer models whose cross-validated or out-of-sample predictions are better, but that was not the objective here, and metric-reducing model-based weight adjustments seem to favor overfitting of the data. Those regression models which incorporated most or all of the interesting survey attributes did exceptionally well with respect to the new metrics, even though some of those models would not have been preferred from examination of likelihood ratios or deviance. While the same adjustment strategy could not be tried if the selected set of interesting survey attributes were too large, the strategy seems to be a good one in the SIPP setting, where the selected set of attributes was still small enough to contain variables which were almost all highly predictive of response and not redundant except for the triple Emp, UnEmp, NILF which partitions the population by definition.

One important check on the usefulness of highly parameterized adjustment models is to see whether their benefits, as measured by the metrics developed here, disappear when the adjusted weights are raked as they often will be in practice. That did happen here with respect to the most conservative of the metrics, m_k^{Cum}. But what we found with respect to the other two metrics is that the best nonresponse models in the SIPP 1996 panel perform almost equally well whether their adjusted weights are raked or not, and that – from the vantage point of our metrics – there may not be much value in extensive raking when a highly effective nonresponse adjustment model is used.

6. References

Bailey, L. (2004). Weighting Alternatives to Compensate for Longitudinal Nonresponse in the Survey of Income and Program Participation. Census Bureau Internal Report, Nov., 16, 2004.

Dagum, E. and Cholette, P. (2006). Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series. Lecture Notes in Statistics, 186. Springer: New York.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Dufour, J., Gagnon, F., Morin, Y., Renaud, M., and Särndal, C.-E. (2001). A Better Understanding of Weight Transformation Through a Measure of Change. Survey Methodology, 27, 97–108.

Eltinge, J. and Yansaneh, I. (1997). Diagnostics for the Formation of Nonresponse Adjustment Cells, with an Application to Income Nonresponse in the U.S. Consumer Expenditure Survey. Survey Methodology, 23, 33–40.

Kim, J.K. and Kim, J. (2007). Nonresponse Weighting Adjustment Using Estimated Response Probability. Canadian Journal of Statistics, 35, 501–514.

Feller, W. (1948). On the Kolmogorov-Smirnov Limit Theorem for Empirical Distributions. Annals of Mathematical Statistics, 19, 177–189.

Oh, H. and Scheuren, F. (1983). Weighting Adjustment for Unit Nonresponse. In Incomplete Data in Sample Surveys, W. Madow, I. Olkin, and D. Rubin (eds). New York: Academic Press, Vol. 2, 143–184.

Pollard, D. (1980). Convergence of Stochastic Processes. New York: Springer-Verlag.

Rizzo, L., Kalton, G., Brick, M., and Petroni, R. (1994). Adjusting for Panel Nonresponse in the Survey of Income and Program Participation. Proceedings of the American Statistical Association, Section on Survey Research.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer.

Slud, E. and Bailey, L. (2006). Estimation of Attrition Biases in SIPP. Proceedings of the American Statistical Association, Section on Survey Research.

Tupek, A. (2002). SIPP 96: Specifications for the Longitudinal Weighting of Sample People. Internal U.S. Census Bureau memorandum.

Received May 2008
Revised May 2009