DSMD Doc. #2.4-R-18

BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION
RESEARCH REPORT SERIES
No. RR-93/02

A REDUCED-SIZE TRANSPORTATION ALGORITHM FOR MAXIMIZING THE OVERLAP BETWEEN SURVEYS

by

Lawrence R. Ernst
Michael M. Ikeda
Bureau of the Census
Statistical Research Division
Washington, D.C. 20233

Report issued: January 28, 1993

Revised: January 31, 1994

ABSTRACT

When redesigning a sample with a stratified multi-stage design, it is sometimes considered desirable to maximize the number of primary sampling units retained in the new sample without altering unconditional selection probabilities. For this problem, an optimal solution which uses transportation theory exists for a very general class of designs. However, this procedure has never been used in the redesign of any survey (that the authors are aware of), in part because even for moderately-sized strata, the resulting transportation problem may be too large to solve in practice. In this paper, a modified reduced-size transportation algorithm is presented for maximizing the overlap, which substantially reduces the size of the problem. This reduced-size overlap procedure was used in the recent redesign of the Survey of Income and Program Participation (SIPP). The performance of the reduced-size algorithm is summarized, both for the actual production SIPP overlap and for earlier, artificial simulations of the SIPP overlap. Although the procedure is not optimal and theoretically, as is shown, can produce only negligible improvements in expected overlap compared to independent selection, in practice it gave substantial improvements in overlap over independent selection for SIPP, and generally provided an overlap that is close to optimal.

KEYWORDS: Linear programming; Sample redesign; Survey of Income and Program Participation.

1. INTRODUCTION

The problem of maximizing the expected number of primary sampling units (PSUs) retained in sample when redesigning a survey with a stratified design for which the PSUs are selected with probability proportional to size was introduced to the literature by Keyfitz (1951). Typically, the motivation for maximizing the overlap of PSUs is to reduce additional costs, such as the training of a new interviewer for a household survey, incurred with each change of sample PSU. Procedures for maximizing overlap do not alter the unconditional probability of selection for a

set of PSUs in a new stratum, but conditions its probability of selection in such a manner that the probability of a PSU being selected in the new sample is generally greater than its unconditional probability when the PSU was in the initial sample and less otherwise.

Overlap procedures are applicable when the redesign results in either a restratification of the PSUs or a change in their selection probabilities. Keyfitz (1951) presented an optimal procedure, but only for one-PSU-per-stratum designs in the special case when the initial and new strata are identical, with only the selection probabilities changing. Causey, Cox and Ernst (1985) obtained an optimal solution to the overlap problem under very general conditions by formulating it as a transportation problem, which is a special form of linear programming problem. This procedure imposes no restrictions on changes in strata definitions or number of PSUs per stratum. (A similar result had been independently obtained by Arthanari and Dodge (1981), although they did not discuss the issue of changes in strata definitions. Both sets of authors obtained their results by generalizing work of Raj (1968).) However, there are at least two other difficulties with the procedure of Causey, Cox and Ernst which can make it unusable in practice, one which is the focus of Ernst (1986), and the other the focus of the current paper.

The first difficulty is that, if the initial sample of PSUs was not selected independently from stratum to stratum, the information necessary to compute all the joint probabilities required by this method may not be available in practice. An alternative linear programming procedure, for use in such cases, was developed by Ernst (1986). The Bureau of the Census has used linear programming to overlap its demographic surveys on five occasions. On four of these occasions (the selection of the 1980s and 1990s Current Population Survey (CPS) designs, and the 1980s and 1990s National Crime Victimization Survey (NCVS) designs) the procedure in Ernst (1986) was used because the initial design was not selected independently from stratum to stratum. In particular, as explained in Ernst (1986), if the initial sample was itself selected by overlapping with a still earlier design then this independence assumption generally does not hold, which was the key reason why it did not hold for these four redesigns.

The second difficulty with the optimal procedure is that the transportation problem may be too

large to solve in practice. The Bureau of the Census also used linear programming to overlap the 1990s Survey of Income and Program Participation (SIPP) design with the 1980s SIPP design, both two-PSUs-per-stratum designs. The initial sample for SIPP was selected independently from stratum to stratum. However, the transportation problem for the optimal procedure would have been too large to practically solve for many strata. This is because for each new stratum to be overlapped consisting of n PSUs, the number of variables in the transportation problem for the optimal procedure can be as large as $2^n \times \binom{n}{2}$. The largest value of n for which a transportation problem with that many variables can be solved with the computer facilities that we have used is approximately n=15.

This paper presents a reduced-size formulation of the overlap procedure as a transportation problem which decreases the numbers of variables in the SIPP problem to $\left(\binom{n}{2}+n+1\right)\times\binom{n}{2}$, a striking reduction for moderate to large values of n. The procedure assumes that the initial sample was selected independently from stratum to stratum, and hence could not have been used instead of the procedure of Ernst (1986) to overlap the CPS and NCVS designs. This reduced-size procedure has been successfully run for strata with as many as 68 PSUs. In contrast, for n=68, the $2^{68}\times\binom{68}{2}$ possible number of variables for the unreduced formulation is far beyond the size of problem that can be solved by any current computer. Furthermore, though the reduced-size procedure sacrifices optimality in exchange for its size reduction, it does appear in practice to yield results fairly close to optimal, as we will show. The reduced-size procedure is the procedure that was used to overlap SIPP.

In Section 2 the procedures of Keyfitz (1951), Raj (1968), and Causey, Cox and Ernst (1985) are reviewed, to provide background for the presentation of the reduced-size procedure.

The reduced-size procedure is presented in Section 3. Although the approach has general applicability, for ease of presentation it is only described in detail for the case when both the initial and new designs are two-PSUs-per-stratum without replacement. Small, artificial examples of the reduced-size procedure are also presented in Section 3. These examples serve to illustrate the procedure; to demonstrate that the reduced-size procedure is sometimes, but not always, optimal; and to demonstrate that the ordering of the pairs of PSUs in a new design stratum, a key step in the algorithm, affects the expected overlap.

In Section 4 the reduced-size procedure is compared analytically to the optimal procedure. Upper bounds on the loss in expected overlap from using the reduced-size procedure instead of the optimal procedure are obtained. It is also demonstrated that in certain situations this loss can approach two PSUs for two-PSUs-per-stratum designs, the worst possible situation.

Finally, in Section 5, the performance of the reduced-size procedure is presented, both for the actual SIPP production overlap and for earlier, artificial simulations of the SIPP overlap. The expected overlap for this procedure is compared to that for independent selection of the new sample PSUs and to an upper bound on the optimal expected overlap. The results show that for this application, in constrast with some of the theoretical results in Section 4, the expected overlap with the reduced-size procedure was much larger than if independent selection had been used to select the new sample PSUs, and nearly as large as the optimal expected overlap. Also presented are computer running times for the reduced-size procedure as a function of stratum size.

2. REVIEW OF PRIOR OVERLAP PROCEDURES

The procedure of Keyfitz (1951) is reviewed in Section 2.1, and the transportation problem procedures of Raj (1968) and Causey, Cox and Ernst (1985) are reviewed in Section 2.2.

First, however, we present some notation that will be used throughout the paper. Let S denote a stratum in the new design consisting of n PSUs, $A_1,...,A_n$. Let the random set I denote the set of integers i for which A_i was in the initial sample, and let N be the corresponding random set

with respect to the new sample. For the simple conditions considered by Keyfitz (1951), the possible values of I and N are the n singleton sets $\{1\}, \{2\},...,\{n\}$. Possible values for I and N for more general overlap problems are discussed in Section 2.2. Finally, let p_i, π_i denote the probability that $i \in I$ and $i \in N$, respectively, and p_{ij}, π_{ij}, $i \neq j$, be the joint probability that $i,j \in I$ and $i,j \in N$, respectively.

The goal of all overlap procedures is the same, to obtain conditional probabilities of selection for the new sample PSUs which maximize the expected number of PSUs common to both samples, that is the number of elements in $N \cap I$, while preserving the unconditional selection probability for each possible value of N. For the procedure of Keyfitz (1951) this reduces to the problem of maximizing the probability that $N=I$, that is, the probability that the same PSU was selected from S for both the initial and final samples.

2.1 The Method of Keyfitz

Keyfitz (1951) presented the following simple set of conditional probabilities in the case when the initial and new designs are both one-PSU-per-stratum and the strata definitions are identical in both designs, with only the selection probabilities changing.

$$P(N=\{j\} \mid I=\{i\}) = \min\{1,\ \pi_i/p_i\} \qquad \text{if } j=i, \tag{2.1}$$

$$= (1-\min\{1,\ \pi_i/p_i\})\ \frac{\max\{\pi_j - p_j,\ 0\}}{\sum_{k=1}^{n} \max\{\pi_k - p_k,\ 0\}} \qquad \text{if } j\neq i. \tag{2.2}$$

In particular, note that if A_i was in the initial sample and $p_i \leq \pi_i$ then this PSU is retained with certainty, while otherwise the conditional probability of its retention is π_i/p_i. This fact will be used to motivate part of the reduced-size algorithm presented in Section 3.

To illustrate Keyfitz's, method, consider a stratum S with n=3 for which

$$p_1=.36,\ p_2=.24,\ p_3=.40,\ \pi_1=.50,\ \pi_2=.30,\ \pi_3=.20.$$

The conditional selection probabilities for the PSUs in the new sample, obtained from (2.1), (2.2), are presented in Table 1.

Table 1. Conditional Probabilities for Keyfitz's Procedure

	Final PSU		
Initial PSU	1	2	3
1	1.00	.00	.00
2	.00	1.00	.00
3	.35	.15	.50

Note that by examining the entries in this table row by row, we can see that when A_i was in the initial sample the conditional probability of selecting A_i in the new sample is greater than π_i, while the conditional probability of selecting A_j, $j\neq i$, is less then π_j. Also note that with this procedure, the unconditional new selection probability for each A_j, obtained by multiplying the entry in cell (i,j), i=1,2,3, by p_i and summing, does equal π_j, as required.

Furthermore, the overlap probability using the Keyfitz procedure, obtained by multiplying the

entry in cell (*i*,*i*) by p_i, *i*=1,2,3, and summing, is .8. This compares with an overlap probability of $\sum_{i=1}^{3} p_i \pi_i = .332$ if the new PSUs are selected independently of the initial PSUs.

Perkins (1970), and Kish and Scott (1971) presented extensions of Keyfitz's procedure for the more general one-PSU-per-stratum problem for which the strata definitions can change in the new design. Their procedures are not optimal, nor do they have any obvious extension to other than one-PSU-per-stratum designs. Optimal solutions to the general overlap procedure awaited the application of linear programming techniques and are discussed in the next subsection.

2.2 Formulation of the Overlap Problem As a Transportation Problem

A transportation problem is a particular form of linear programming problem, in which an objective function of the form

$$\sum_{i=1}^{a} \sum_{j=1}^{b} c_{ij} x_{ij} \tag{2.3}$$

is to be either maximized or minimized, subject to the constraints

$$\sum_{j=1}^{b} x_{ij} = \alpha_i, \qquad \text{i=1,...,a,} \tag{2.4}$$

$$\sum_{i=1}^{a} x_{ij} = \beta_j, \qquad \text{j=1,...,b,} \tag{2.5}$$

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j, \tag{2.6}$$

where the x_{ij} are nonnegative variables and α_i, β_j are constants. The variables can be viewed as internal cells of a two-dimensional tabular array, with row and column totals specified by (2.4) and (2.5), respectively. The sum of the row totals must equal the sum of the column totals for such an array, which is precisely what (2.6) states. Solution strategies that are extremely efficient computationally exist for transportation problems (Glover et al. 1974).

Raj (1968) was the first to formulate the overlap problem as a transportation problem, but he only considered the same one-PSU-per-stratum case without change in strata definitions as Keyfitz. Raj let the variables x_{ij}, $i,j=1,...,n$, denote the joint probability that $i \in I$ and $j \in N$; set $c_{ij}=1$ if $i=j$ and $c_{ij}=0$ if $i \neq j$; and let $\alpha_i = p_i$, $\beta_j = \pi_j$. Thus, the problem in the transportation formulation is to determine a set of nonnegative x_{ij}'s which maximize

$$\sum_{i=1}^{n} x_{ii} \tag{2.7}$$

subject to the constraints

$$\sum_{j=1}^{n} x_{ij} = p_i, \qquad i=1,...,n, \tag{2.8}$$

$$\sum_{i=1}^{n} x_{ij} = \pi_j, \qquad j=1,...,n. \tag{2.9}$$

The objective function (2.7) is the probability of overlap, that is the probability that the same PSU in *S* is in both samples, while constraints (2.8) and (2.9) must be met in order for the joint probabilities to sum to the correct initial and new selection probabilities, respectively.

Once an optimal set of x_{ij}'s are obtained, the conditional probability of selecting A_j in the new sample given that A_i was in the initial sample is simply x_{ij}/p_i for all i,j.

To illustrate, the optimal set of x_{ij}'s for the example in Section 2.1 are given in Table 2.

Table 2. Values of x_{ij} for Example of Section 2.1

	j		
i	1	2	3
1	.36	.00	.00
2	.00	.24	.00
3	.14	.06	.20

Upon dividing each entry in row *i* of Table 2 by p_i, Table 1 is obtained again. Furthermore, the maximum value of (2.7) is the sum of the diagonal elements of Table 2, or .8, consistent with the overlap probability given in Section 2.1.

Raj's formulation of the overlap problem as a transportation problem under the conditions considered by Keyfitz has no practical utility, since it is easier to use (2.1), (2.2) to obtain optimal conditional selection probabilities than to solve a transportation problem. The real importance of Raj's approach is that, unlike Keyfitz's, it is readily generalizable, as done in Causey, Cox and Ernst (1985), to yield formulations for optimal solutions for very general designs, with no restrictions on changes in strata definitions or number of PSUs per stratum.

We proceed to present this generalization, which requires additional notation. Let J_i, i=1,...,m*, denote the possible values for *I*, and let S_j, j=1,...,n^*, denote the possible values for *N*. Denote by $P(J_i)$, the probability that $I = J_i$ and by $P(S_j)$ the probability that $N = S_j$. In addition, let x_{ij} be the variable denoting the joint probability of these two events, and let c_{ij} denote the number of elements in $J_i \cap S_j$. Then the transportation problem to solve is to determine $x_{ij} \geq 0$ which maximize

$$\sum_{i=1}^{m^*} \sum_{j=1}^{n^*} c_{ij} x_{ij} \qquad (2.10)$$

subject to

$$\sum_{j=1}^{n^*} x_{ij} = P(J_i), \qquad i=1,...,m^*, \tag{2.11}$$

$$\sum_{i=1}^{m^*} x_{ij} = P(S_j), \qquad j=1,...,n^*. \tag{2.12}$$

The conditional probability that $N = S_j$ given that $I = J_i$ is then $x_{ij}/P(J_i)$ for all i,j.

We observe that under the conditions considered by Keyfitz and Raj, I and N consist of the singleton subsets of $\{1,...,n\}$, $m^* = n^* = n$, $p_i = P(J_i)$, $\pi_j = P(S_j)$, and thus the procedure of Causey, Cox and Ernst reduces to that of Raj.

For the more general one-PSU-per-stratum problem in which the strata definitions can change in the new design, N still consists of the singleton subsets of $\{1,...,n\}$, and hence $n^* = n$, but I depends on the stratification in the initial design. To illustrate, an example is presented in Causey, Cox and Ernst (1985) for which n=5, with A_1, A_2, A_3 in one initial stratum and A_4, A_5 in a second initial stratum; in both of these initial strata there were also additional PSUs not in S. Then $m^* = 12$, with the values of I consisting of the empty set, the 5 singleton sets, and the 6 sets of size 2 for which one element is from $\{1,2,3\}$ and the other from $\{4,5\}$. In general, the maximum value for m^* is 2^n, the number of subsets of $\{1,...,n\}$. For the case when both designs are one-PSU-per-stratum, $m^* = 2^n$ if and only if the n PSUs in S were in n different initial strata.

We next consider the case where both the initial and new designs are two-PSUs-per-stratum without replacement. We present an example to illustrate the use of the formulation (2.10)-(2.12).

Consider a final stratum S with n=3. All of the PSUs were in different initial strata.

Let p_1=.6, p_2=.75, p_3=.7, π_1=.5, π_2=.8, π_3=.7.

Then p_{12}=.45, p_{13}=.42, p_{23}=.525, π_{12}=.30, π_{13}=.20, π_{23}=.50.

Since the PSUs were all in different initial strata, there are 8 different possibilities for I, with probabilities given in Table 3.

Table 3. Probabilities for Possible Sets of Initial Sample PSUs

i	1	2	3	4	5	6	7	8
J_i	{1,2,3}	{1,2}	{1,3}	{2,3}	{1}	{2]	{3}	∅
P(J_i)	.315	.135	.105	.21	.045	.09	.07	.03

Since the new design is two-PSUs-per-stratum without replacement, there are 3 different possibilities for N, namely the pairs S_1={1,2}, S_2={1,3}, S_3={2,3}, and hence $P(S_1)$=.30, $P(S_2)$=.20, $P(S_3)$=.50.

Furthermore, the values of c_{ij} are then as given in Table 4. Upon maximizing (2.10) subject to (2.11) and (2.12) with the given $P(J_i)$'s, $P(S_j)$'s and c_{ij}'s, an optimal set of x_{ij}'s, presented in Table 5, is obtained. Finally, by dividing each of the entries in row i of Table 5 by $P(J_i)$, an optimal set of conditional probabilities $P(S_j \mid J_i)$, in Table 6, is obtained.

Table 4. Values of c_{ij} for Optimal Procedure

	j		
i	1	2	3
1	2	2	2
2	2	1	1

3	1	2	1
4	1	1	2
5	1	1	0
6	1	0	1
7	0	1	1
8	0	0	0

Table 5. Values of x_{ij} that Maximize Overlap for Optimal Procedure

	j		
i	1	2	3
1	.000	.025	.290
2	.135	.000	.000
3	.000	.105	.000
4	.000	.000	.210
5	.045	.000	.000
6	.090	.000	.000
7	.000	.070	.000
8	.030	.000	.000

Table 6. Conditional Probabilities for Optimal Procedure

	Final Sample PSUs		
Initial Sample PSUs	{1,2}	{1,3}	{2,3}
{1,2,3}	0	5/63	58/63
{1,2}	1	0	0
{1,3}	0	1	0
{2,3}	0	0	1
{1}	1	0	0
{2}	1	0	0

{3}	0	1	0
∅	1	0	0

Note that, as previously mentioned and as illustrated by this example, when the number of PSUs per stratum in the new design is greater than 1, objective function (2.10) has a more general meaning than (2.7). While (2.7) is the probability that the new sample PSU was in the initial sample, (2.10) is the expected number of new sample PSUs that were in the initial sample. For this example, the expected overlap under the optimal procedure is 1.735 PSUs. In comparison, the expected overlap if the initial and final designs are selected independently is $p_1\pi_1 + p_2\pi_2 + p_3\pi_3 = 1.39$ PSUs. Also observe that it can readily be verified that the set of conditional probabilities in Table 6 is optimal, since the conditional expected overlap is 2 whenever at least a pair of PSUs are in *I* and the conditional expected overlap is 1 whenever *I* consists of exactly one PSU.

For two-PSU-per-stratum without replacement problems, the possible values for *N* are always the $\binom{n}{2}$ subsets of {1,..,*n*} of size 2, that is $n^* = \binom{n}{2}$. However m^* can vary widely. $m^* = \binom{n}{2}$ when the PSUs in *S* comprise a single initial stratum. The upper bound of 2^n on m^* is attained when all the PSUs in *S* were in different initial strata, as illustrated by the previous example, and in some other situations, as will be shown. To obtain a general, exact expression for m^*, let G_i', *i*=1,...,*r*, denote the set of initial strata with PSUs in common with *S*; let $G_i = G_i' \cap S$; and let n_i, n_i' denote the number of PSUs in G_i, G_i', respectively. Assume that each G_i' is a nonselfrepresenting stratum. Then the following results hold:

Theorem 2.1

$$m^* = \prod_{i=1}^{r} m_i, \tag{2.13}$$

where

$$m_i = 2 \quad \text{if } n_i = 1, \tag{2.14}$$

$$= \binom{n_i}{2} \quad \text{if} \quad n_i' = n_i \geq 2, \tag{2.15}$$

$$= \binom{n_i}{2} + n_i \quad \text{if} \quad n_i' - 1 = n_i \geq 2, \tag{2.16}$$

$$= \binom{n_i}{2} + n_i + 1 \quad \text{if} \quad n_i' - 2 \geq n_i \geq 2, \tag{2.17}$$

Furthermore, $m^* = 2^n$ if and only if $n_i \leq 2$ for all i and, in addition, $n_i' \geq 4$ whenever $n_i = 2$. This theorem is proven in Appendix A.

For the two-PSUs-per-stratum without replacement overlap problem, the number of variables in the transportation problem for the optimal procedure is $m^* n^*$ which, by Theorem 2.1, can be as large as $2^n \binom{n}{2}$. For n=15, $2^n \binom{n}{2} = 3{,}440{,}640$, which is about as large a transportation problem as can be solved with the computer facilities that we used. However, n>15 for nearly half the nonselfrepresenting strata in our SIPP application, and consequently it was necessary to develop a procedure, described in the next section, which reduces the size of the transportation problem, while still producing nearly maximal expected overlap in practice.

Aragon and Pathak (1990) present a different approach to the problem of reducing the size of the transportation problem than the procedure to be presented in Section 3. Their approach retains optimality and reduces the size of the problem by 75 percent when $m^* = n^*$. Unfortunately, when m^* is much larger than n^*, which is when size reduction is most needed,

their approach produces negligible size reduction in relative terms.

3. THE ALGORITHM FOR THE REDUCED-SIZE PROCEDURE

The reduced-size procedure is applicable whenever PSUs in the initial and new designs are selected without replacement. However, the procedure will be described in detail, in Section 3.1, only for the case when both the initial and new designs are two-PSUs-per-stratum. Then, in Section 3.2, the changes necessary to apply this procedure for other initial and new designs will be sketched. It is assumed that PSUs in the initial sample were selected independently from stratum to stratum.

3.1 Reduced-Size Procedure When Both Designs Are Two-PSUs-Per-Stratum

The general outline of the procedure for this case is as follows. First, the $\binom{n}{2}$ subsets of $\{1,...,n\}$ of size 2 are ordered in a manner to be described later. (For now, we simply note that any ordering can be used to reduce the size of the transportation problem. The specific one used is for the purpose of accomplishing the size reduction while also attempting to give up as little as possible of the gains in overlap that the optimal procedure yields.) We let I_i, $i=1,...,\binom{n}{2}$, denote the i-th element in the ordering; let $I_{\binom{n}{2}+1},...,I_{\binom{n}{2}+n}$ be the n singleton subsets; and set $I_{\binom{n}{2}+n+1}=\varnothing$. Thus, the I_i's constitute all subsets of $\{1,...,n\}$ of 2 or fewer elements. For each possibility for I, a unique set I^* is associated among these $\binom{n}{2}+n+1$ subsets and the new selection probabilities conditioned on the associated I^*, rather than on I itself. Therefore, the new selection probabilities are conditioned on $\binom{n}{2}+n+1$ events instead

of a possible 2^n events, which is the reason for the size reduction. The associated I^* is the first I_i for which $I_i \subset I$. That is, if I consists of at least two integers, the associated I^* is the first pair in the ordering contained in I, while if I is a singleton set or empty then $I^* = I$.

The reduced-size transportation problem attempts to retain the PSUs corresponding to elements in the associated set I^* in the new sample, but does not use information on elements in $I \sim I^*$. The form of this reduced-sized transportation problem based on the set of I_i's is as follows. Let p_i^* be the probability that $I^* = I_i$, $i=1,...,\binom{n}{2}+n+1$, and abbreviate $\pi_j^* = P(S_j)$, $j=1,...,\binom{n}{2}$. For each i,j, the variable x_{ij} is the joint probability that $I^* = I_i$ and that $N=S_j$, while c_{ij} is the expected number of elements in $I \cap S_j$ given $I^* = I_i$. The problem to solve is to determine $x_{ij} \geq 0$ that maximize

$$\sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} c_{ij} x_{ij}, \tag{3.1}$$

subject to

$$\sum_{j=1}^{\binom{n}{2}} x_{ij} = p_i^*, \qquad i=1,...,\binom{n}{2}+n+1, \tag{3.2}$$

$$\sum_{i=1}^{\binom{n}{2}+n+1} x_{ij} = \pi_j^*, \qquad j=1,...,\binom{n}{2}. \tag{3.3}$$

Once the optimal x_{ij}'s have been obtained, then the conditional new selection probabilities for S_j, $j=1,...,\binom{n}{2}$, given $I^* = I_i$, are x_{ij}/p_i^*. Note that the number of variables, x_{ij}, in the formulation (3.1)-(3.3) is $\left(\binom{n}{2}+n+1\right)\times\binom{n}{2}$, in comparison with a maximum of $2^n\times\binom{n}{2}$ in the formulation (2.10)-(2.12).

It remains to explain the general method for obtaining the ordering of the $\binom{n}{2}$ pairs and the procedures for computing the p_i^*'s and c_{ij}'s. Before doing this, we present an example of the reduced-size procedure, namely the two-PSUs-per-stratum example used in Section 2.2 to illustrate the transportation problem formulation for the optimal procedure.

The ordering of the pairs for this example, as will be shown later, is {2,3}, {1,2}, {1,3}. Consequently, the I_i's, are as given in Table 7. Note that if I={1,2,3} or I={2,3}, then the associated set is I_1={2,3}. For the other six possibilities for I the associated set is I itself.

Consequently, from Table 3 we obtain that

$$p_1^* = P(I=\{1,2,3\}) + P(I=\{2,3\}) = .525, \tag{3.4}$$

$p_i^* = P(J_i)$, i=2,3, and $p_i^* = P(J_{i+1})$, i=4,...,7, yielding the values in Table 7. Since $\pi_j^* = P(S_j)$, we have $\pi_1^* = .30$, $\pi_2^* = .20$, $\pi_3^* = .50$.

Table 7. Probabilities of Associated Sets: Reduced-Size Procedure

	i						
	1	2	3	4	5	6	7
I_i	{2,3}	{1,2}	{1,3}	{1}	{2}	{3}	∅

p_i^*	.525	.135	.105	.045	.09	.07	.03

The c_{ij} values for this example are given in Table 8. In order to obtain these values, we simplified the computation by letting

$$b_{it} = P(t \in I \mid I^* = I_i), \quad i = 1,\ldots,\binom{n}{2} + n + 1, \quad t = 1,\ldots,n, \tag{3.5}$$

and noting that if $S_j = \{s,t\}$ then

$$c_{ij} = b_{is} + b_{it}. \tag{3.6}$$

That is, the expected number of elements in $I \cap S_j$ given $I^* = I_i$ is simply the sum of the probabilities that each of the two elements in S_j was in I given $I^* = I_i$. Also observe that while the transportation problem for the optimal procedure knows the exact value for I and hence knows with certainty whether each element in S_j was in I, this is not the case for the reduced-size procedure, since only the associated set I_i is known. To illustrate, consider the first row of Table 8. Since $I_1 = \{2,3\}$, we know that $2 \in I$ and $3 \in I$, and hence $b_{12} = b_{13} = 1$. However, we do not with certainty whether $1 \in I$ since I_1 is the associated set for both $I = \{1,2,3\}$ and $I = \{2,3\}$. In fact, from Table 3,

$$b_{11} = \frac{P(I = \{1,2,3\})}{P(I = \{1,2,3\}) + P(I = \{2,3\})} = .6.$$

Then $c_{11} = b_{11} + b_{12} = 1.6$, with c_{12}, c_{13} computed similarly. For the remaining six rows in Table 8, $I_i = I$ and hence it is known with certainty which integers were in I. Consequently, the c_{ij}'s for these six rows are easily computed.

Finally, we maximize the expected overlap (3.1) subject to (3.2) and (3.3), obtaining the x_{ij} values in Table 9. The conditional probabilities $P(N = S_j \mid I^* = I_i)$ in Table 10 are then obtained by dividing the *i*-th row of Table 9 by $p^*_{i\cdot}$.

Table 8. Values of c_{ij} for the Reduced-Size Procedure

		j		
i	I_i	1	2	3
1	{2,3}	1.6	1.6	2.0
2	{1,2}	2.0	1.0	1.0
3	{1,3}	1.0	2.0	1.0
4	{1}	1.0	1.0	0.0
5	{2}	1.0	0.0	1.0
6	{3}	0.0	1.0	1.0
7	∅	0.0	0.0	0.0

Table 9. Values of x_{ij} for the Reduced-Size Procedure

		j		
i	I_i	1	2	3
1	{2,3}	0.000	0.025	0.500
2	{1,2}	0.135	0.000	0.000
3	{1,3}	0.000	0.105	0.000
4	{1}	0.045	0.000	0.000
5	{2}	0.090	0.000	0.000
6	{3}	0.000	0.070	0.000
7	∅	0.030	0.000	0.000

Table 10. Conditional Probabilities for the Reduced-Size Procedure

		j		
i	I_i	1	2	3
1	{2,3}	0	1/21	20/21
2	{1,2}	1	0	0
3	{1,3}	0	1	0
4	{1}	1	0	0
5	{2}	1	0	0
6	{3}	0	1	0
7	$\varnothing$	1	0	0

The expected overlap for the reduced-size procedure is .01 less than optimal, that is 1.725 PSUs. The deviation from optimality arises solely because the expected overlap is 1.6 for the joint event that $I^* = \{2,3\}$ and $N = \{1,3\}$. Since the probability of this joint event is .025, and the optimal procedure for this example always produces an overlap of 2 when at least 2 of the PSUs were in the initial sample, the deviation from optimality is .025(2-1.6)=.01.

The reason that the reduced-size procedure is not able to obtain optimality is that the pair {2,3} has a smaller probability of selection in the new sample than in the initial sample. As a result, both the optimal procedure and the reduced-size procedure must sometimes select another pair (always {1,3} for both procedures in this example) when {2,3} was in the initial sample. The distinction between the two procedures is that the optimal procedure only selects {1,3} when $1 \in I$. The reduced-size procedure is unable to use the information about whether $1 \in I$. As a result, when $\{2,3\} \subset I$, $1 \in N$ independently of whether $1 \in I$. This results in a deviation from the optimal overlap.

Although, as illustrated by this example the reduced-size algorithm does not always yield the optimal expected overlap, in practice it often does. To illustrate, consider the previous example with the single modification that p_2=.50 instead of .75. The conditional probabilities for the optimal procedure are presented in Table 11.

Table 11. Conditional Probabilities for Optimal Procedure with p_2 = .50

	Final Sample PSUs		
Initial Sample PSUs	{1,2}	{1,3}	{2,3}
{1,2,3}	0	0	1
{2,3}	0	0	1
{1,2}	1	0	0
{1,3}	0	20/21	1/21
{1}	1	0	0
{2}	1	0	0
{3}	0	0	1
∅	1	0	0

The corresponding table for the reduced-size procedure is identical to Table 11 except that the first row is omitted, with the rows for {1,2,3} and {2,3} combined into a single row for {2,3}. The expected overlap is 1.58 for both procedures for this example. The conditional expected overlap is 1 whenever there is exactly 1 PSU in *I* and the conditional expected overlap is 2 whenever there are at least 2 PSUs in *I*, except if *I*={1,3}. If *I*={1,3} then the conditional probability of {1,3} being selected in the new sample is 20/21 for both procedures. Since $P(I=\{1,3\})$=.21 and π_{13}=.20, no procedure can yield a higher conditional probability of retaining {1,3} when *I*={1,3}.

We now proceed to show in general how the ordering of the pairs is obtained and the p_i^*'s and c_{ij}'s are computed.

We first consider the ordering of the pairs. The motivation for the ordering is as follows. If the *i*-th pair in the ordering is {*s*,*t*} then it would be possible for the transportation problem to retain this pair in the new sample when $I^* = I_i$ with conditional probability $\min\{1, \pi_{st}/p_i^*\}$. This is analogous to (2.1) in Keyfitz's procedure. Therefore, roughly the goal in the ordering is to make these conditional probabilities as large as possible on average over all pairs.

To illustrate how the ordering of the pairs affects the expected overlap we consider the example of Table 7. Our ordering procedure, as will be shown later, produces the indicated ordering and yields an expected overlap of 1.725 PSUs. Next consider the following alternative ordering for this example. Let the first pair in the ordering be {1,3}, the second pair be {1,2} and the last pair be {2,3}. With this alternative ordering, $I^* = \{1,3\}$ whenever either I={1,2,3} or I={1,3}. Therefore, for this ordering p_1^* is the probability that $I^* = \{1,3\}$, which is now .42. Furthermore, for this alternative ordering $p_3^* = P(I^* = \{2,3\}) = P(I = \{2,3\}) = .21,$ while the other 5 columns in Table 7 remain unchanged. The alternative ordering results in the conditional probabilities in Table 12.

Table 12. Conditional Probabilities for Alternative Ordering

		j		
i	I_i	1	2	3
1	{1,3}	0	10/21	11/21
2	{1,2}	1	0	0
3	{2,3}	0	0	1
4	{1}	1	0	0
5	{2}	1	0	0
6	{3}	0	0	1
7	∅	1	0	0

It can be calculated, using the same approach used for Table 10, that the expected overlap for this example is 0.055 less than optimal, that is 1.68 PSUs. The reason that this alternative ordering results in a lower expected overlap is as follows. In general a later placement of a pair in the ordering, results in a lower value for the corresponding p_i^*, and hence a higher conditional retention probability when $I^* = I_i$. That is, with {1,3} first in the ordering, $\pi_{13}/p_1^* = 10/21$, which is the conditional retention probability for this pair when $I^*=\{1,3\}$; while when {1,3} is third in the ordering $\pi_{13}/p_3^* > 1$ and this pair is retained with certainty. Now the conditional retention probability for the pair {2,3} when $I^* = \{2,3\}$ also increases to 1 when {2,3} is moved from first to third in the ordering, but the increase is only from 20/21, and hence the ordering in Table 7 produces a higher expected overlap than the ordering in Table 12.

Thus, as this example illustrates, the goal of the ordering is to place pairs earlier in the ordering that have a relatively high conditional retention probability even with an early placement. To obtain the desired ordering of the pairs of integers, an ordering $f(1),...,f(n)$ of $\{1,...,n\}$ will first be obtained by recursion. Then corresponding to each k=1,...,n-1, an ordering $g_k(1),...,g_k(n-k)$ of $\{1,...,n\}\sim\{f(1),...,f(k)\}$ will be constructed by recursion. A linear ordering of the distinct pairs in $\{1,...,n\}$ would then be determined as follows. Each such pair can be represented uniquely as an ordered pair $(f(k), g_k(\ell))$ for some $k \in \{1,...,n-1\}$, $\ell \in \{1,...,n-k\}$. A second pair representable in the form $(f(k'), g_{k'}(\ell'))$ precedes $(f(k), g_k(\ell))$ if and only if either $k'<k$, or $k'=k$ and $\ell'<\ell$. To illustrate, for the example just considered it will be shown later that $f(1)=2$, $f(2)=3$, $f(3)=1$, $g_1(1)=3$, $g_1(2)=1$, $g_2(1)=1$, and hence the ordering of the pairs is {2,3}, {2,1}, {3,1}. Both the f ordering and the g_k ordering will be constructed to meet the goal stated at the beginning of this paragraph.

To obtain the ordering $f(1),...,f(n)$, recursively define $f(k)$, k=1,...,n, by choosing $f(k) \in T_k$

satisfying

$$\pi_{f(k)}/p_{f(k)}^{(k)} = \max\{\pi_i/p_i^{(k)}: i \in T_k\},$$

where

$$T_1 = \{1,...,n\}, \quad T_k = T_{k-1} \sim \{f(k-1)\}, \quad k=2,...,n,$$

$$p_i^{(k)} = P(i \in I \text{ and } I \subset T_k), \quad k=1,...,n, \quad i \in T_k. \tag{3.7}$$

Since $p_i^{(1)} = p_i$, the ordering just defined corresponds to placing first a PSU with the greatest value of π_i/p_i. For all k, $p_{f(k)}^{(k)}$ is the probability that $f(k)$ was in I and none of the k-1 elements proceeding $f(k)$ in the f ordering were in I, and hence $p_{f(k)}^{(k)}$ is the probability that an attempt is made to retain $A_{f(k)}$ in the new sample either as the first member of an ordered pair of initial sample PSUs or as the only initial sample PSU in S. Generally, the larger $\pi_{f(k)}/p_{f(k)}^{(k)}$ is, the greater the probability that this attempt would be successful. Thus, the motivation for the f ordering of the individual PSUs is the analog of the motivation for the ordering of the pairs of PSUs that we previously discussed.

It remains to explain how to compute $p_i^{(k)}$ for $k \geq 2$. To this end, let r denote the number of initial strata with PSUs in common with S and let F_α, $\alpha=1,...,r$, denote a partition of $\{1,...,n\}$ such that i and j are in the same F_α if and only if A_i and A_j were in the same initial stratum. Then let

$$p'_\alpha(T) = P(I \cap F_\alpha \subset T), \quad \alpha=1,...,r, \quad T \subset \{1,...,n\}, \tag{3.8}$$

$$p''_{i\alpha}(T) = P(i \in I \text{ and } I \cap F_\alpha \subset T), \quad \alpha=1,...,r, \quad T \subset \{1,...,n\}, \quad i \in F_\alpha \cap T, \tag{3.9}$$

and observe that

$$p'_{\alpha}(T) = 1 - \sum_{i \in F_{\alpha} \sim T} p_i + \sum_{\substack{i,j \in F_{\alpha} \sim T \\ i<j}} p_{ij}, \tag{3.10}$$

$$p''_{i\alpha}(T) = p_i - \sum_{j \in F_{\alpha} \sim T} p_{ij}, \tag{3.11}$$

and finally that

$$p_i^{(k)} = p''_{i\alpha}(T_k) \prod_{\substack{\ell=1 \\ \ell \neq \alpha}}^{r} p'_{\ell}(T_k), \quad k=1,...,n, \quad i \in F_{\alpha} \cap T_k. \tag{3.12}$$

This last formula is obtained by noting that since $p'_{\ell}(T_k)$, $\ell \neq \alpha$, is the probability that all of the elements in $I \cap F_{\ell}$ are in T_k, while $p''_{i\alpha}(T_k)$ is the joint probability that $i \in I$ and that all the elements in $I \cap F_{\alpha}$ are in T_k, then since the selection of the PSUs in the initial design is assumed independent from stratum to stratum, the product of terms on the right hand side of equation (3.12) is the probability that $i \in I$ and $I \subset T_k$, which is precisely the definition of $p_i^{(k)}$ given by (3.7).

Next, for each $k=1,...,n-1$, the ordering $g_k(\ell)$, $\ell=1,...,n-k$, is recursively defined by choosing $g_k(\ell) \in T_{k\ell}$ satisfying

$$\pi_{f(k),g_k(\ell)} / p^{(\ell)}_{f(k),g_k(\ell)} = \max \{ \pi_{f(k),j} / p^{(\ell)}_{f(k),j} : j \in T_{k\ell} \},$$

where

$$T_{k1} = \{1,...,n\} \sim \{f(1),...,f(k)\},$$

$$T_{k\ell} = T_{k(\ell-1)} \sim \{g_k(\ell-1)\}, \quad \ell=2,...,n-k,$$

$$T^*_{k\ell} = T_{k\ell} \cup \{f(k)\}, \quad \ell=1,...,n-k,$$

$$p^{(\ell)}_{f(k),j} = P(f(k),\ j \in I \text{ and } I \subset T^*_{k\ell}), \quad \ell=1,...,n-k, \quad j \in T_{k\ell}. \tag{3.13}$$

Note that $p^{(\ell)}_{f(k),j}$ is thus the joint probability that $f(k)$ is the first integer in the f ordering in I, that none of the first ℓ-1 integers in the g_k ordering are in I, and that $j \in I$. Consequently, $p^{(\ell)}_{f(k),g_k(\ell)}$ is the probability that $I^* = \{f(k),\ g_k(\ell)\}$. Furthermore, if $I_i = \{f(k),\ g_k(\ell)\}$ then $p_i^* = p^{(\ell)}_{f(k),g_k(\ell)}$, and hence the choice of $g_k(\ell)$ results in the largest value of $\pi_{f(k),\,g_k(\ell)}/p_i^*$ among the elements in $T_{k\ell}$, in accordance with the previously stated goal for the ordering of the pairs of PSUs.

To compute $p^{(\ell)}_{f(k),j}$, observe that if $f(k) \in F_\alpha$, $j \in F_\beta$, then

$$\begin{aligned} p^{(\ell)}_{f(k),j} &= p_{f(k),j} \prod_{\substack{t=1 \\ t\neq\alpha}}^{r} p_t'(T^*_{k\ell}) \quad \text{if } \alpha=\beta, \\ &= p''_{f(k),\alpha}(T^*_{k\ell}) p''_{j\beta}(T^*_{k\ell}) \prod_{\substack{t=1 \\ t\neq\alpha,\beta}}^{r} p_t'(T^*_{k\ell}) \text{ if } \alpha\neq\beta. \end{aligned} \tag{3.14}$$

These formulas can be obtained by first noting that in (3.14), $p_t'(T^*_{k\ell})$ is the probability that $I \cap F_t \subset T^*_{k\ell}$. If $\alpha=\beta$ then $p_{f(k),j}$ is the probability that $f(k)$ and j were in I, and hence the only elements in $I \cap F_\alpha$; while if $\alpha\neq\beta$, then $p''_{f(k),\alpha}(T^*_{k\ell})\, p''_{j\beta}(T^*_{k\ell})$ is the probability that $f(k)$ and j were

in I, and $I \cap (F_\alpha \cup F_\beta) \subset T_{k\ell}^*$. The product of the terms in both formulas in (3.14) is thus the probability that $f(k)$, $j \in I$, and $I \subset T_{k\ell}^*$, which is precisely the definition of $p_{f(k),j}^{(\ell)}$ in (3.13).

We illustrate the computations used in obtaining the ordering for the example that we have been considering. First note that $f(1)=2$ since the largest value of π_i/p_i occurs for $i=2$. Next we find $g_1(1)$ which, since $f(1)=2$, is the $j \in \{1,3\}$ with the maximum value of $\pi_{2j}/p_{2j}^{(1)}$. To find this j, first let $F_\alpha=\{\alpha\}$, $\alpha=1,2,3$, and note that $T_{11}^*=\{1,2,3\}$. From (3.14) with $\alpha=2$, $\beta=1$, it then follows that

$$p_{21}^{(1)} = p_{22}''\{1,2,3\}p_{11}''\{1,2,3\}p_3'\{1,2,3\} = p_2 p_1 \cdot 1 = .45,$$

and similarly it can be obtained that $p_{23}^{(1)}=.525$. Hence $g_1(1) = 3$, since .5/.525>.3/.45. Therefore, the first pair in the ordering is $\{f(1),\ g_1(1)\} = \{2,3\}$. Then $g_1(2) = 1$, since 1 is the only integer remaining to be used in the g_1 ordering, and consequently the second pair in the ordering is $\{f(1),\ g_1(2)\} = \{2,1\}$. It is not really necessary to determine $f(2)$, since $\{1,3\}$ is the only remaining pair, and hence the last pair, but to further illustrate the computations, observe that $T_2=\{1,3\}$, $p_1^{(2)} = p_{11}''\{1,3\}p_2'\{1,3\}p_3'\{1,3\} = p_1(1-p_2)\cdot 1 = .15$ by (3.12), and similarly $p_3^{(2)} = p_3(1-p_2)\cdot 1 = .175$. Hence $f(2)=3$, since .7/.175 > .5/.15. Consequently, $g_2(1)=1$, $f(3)=1$.

Next we explain the computation of the p_i^*'s. If I_i consists of the pair of integers $I_i = \{f(k), g_k(\ell)\}$ then, as previously noted, $p_i^* = p_{f(k),g_k(\ell)}^{(\ell)}$. Consequently, p_i^* can be computed from (3.14) with $j = g_k(\ell)$.

If I_i is a singleton set $\{t\}$ for some $t \in F_\alpha$, then

$$p_i^* = p''_{t\alpha}(\{t\}) \prod_{\substack{u=1 \\ u\neq\alpha}}^{r} p'_u(\varnothing). \tag{3.15}$$

This expression holds since $p'_u(\varnothing)$, $u\neq\alpha$, is the probability that $F_u \cap I = \varnothing$, while $p''_{t\alpha}(\{t\})$ is the probability that $F_\alpha \cap I = \{t\}$, and hence the right hand side of (3.15) is the probability that $I = I_i = \{t\}$.

Finally, if $I_i = \varnothing$, then it can be similarly shown that $p_i^* = \prod_{u=1}^{r} p_u'(\varnothing)$.

To illustrate the computations of the p_i^*'s for our example, note that since $I_1 = \{f(1),\ g_1(1)\} = \{2,3\}$ and, as previously computed, $p_{23}^{(1)} = .525$, it follows that $p_1^* = p_{f(1),g_1(1)}^{(1)} = p_{23}^{(1)} = .525$. Note that we have also computed p_1^* by means of (3.4). That approach to computing p_i^*, requiring summing $P(I)$ over all possible I for which I_i is the associated set, is not practical in general, since there can be as many as 2^{n-2} probabilities summed in the computation of p_1^*.

It remains only to explain how to compute the c_{ij}'s which, by (3.5) and (3.6), reduces to computing b_{it}, $i=1,\ldots,\binom{n}{2}+n+1$, $t=1,\ldots,n$.

To compute b_{it}, observe that

$b_{it} = 0$ if $I_i = \emptyset$,

$= 1$ if $I_i = \{v\}$ and $t=v$,

$= 0$ if $I_i = \{v\}$ and $t \neq v$,

while if $I_i = \{f(k), g_k(\ell)\}$ and $f(k) \in F_\alpha$, $g_k(\ell) \in F_\beta$, $t \in F_\gamma$, then

$b_{it} = 1$ if $t=f(k)$ or $t=g_k(\ell)$, (3.16)

$= 0$ if $t \notin T^*_{k\ell}$, (3.17)

$= 0$ if $t \in T_{k\ell} \sim \{g_k(\ell)\}$ and $\gamma=\alpha=\beta$, (3.18)

$$= \frac{p_{f(k),t}}{p''_{f(k),\alpha}(T^*_{k\ell})} \quad \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \text{ and } \gamma=\alpha\neq\beta, \tag{3.19}$$

$$= \frac{p_{g_k(\ell),t}}{p''_{g_k(\ell),\beta}(T^*_{k\ell})} \quad \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \text{ and } \gamma=\beta\neq\alpha, \tag{3.20}$$

$$= \frac{p''_{t\gamma}(T^*_{k\ell})}{p'_{\gamma}(T^*_{k\ell})} \quad \text{if } t \in T_{k\ell} \sim \{g_k(\ell)\} \text{ and } \gamma \neq \alpha,\ \gamma \neq \beta. \tag{3.21}$$

In Appendix B it is demonstrated how (3.16)-(3.21) were obtained.

In the actual implementation for the SIPP application, modifications of the reduced-size procedure were needed to overlap the 1990s SIPP design with the 1980s SIPP design. The modifications were necessary because the PSU definitions in the 1980s and 1990s designs were not identical. As a result, some PSUs in the 1990s design could intersect more than one 1980s design PSU. These modifications are detailed in Appendix C.

3.2 Modifications of Reduced-Sized Procedure for Other Designs

In general, consider any m'-PSU-per-stratum without replacement initial design and any m-PSUs-per-stratum without replacement final design, where m', m are any positive integers. Although the reduced-size procedure in Section 3.1 was only presented for the case $m = m' = 2$, it is actually applicable for any m, m'. We will sketch the modifications necessary when $m \neq 2$ or $m' \neq 2$.

A different value of m' only requires modification of some of the computations. For example, if m=2, but $m' \neq 2$, then the computations for $p_i^{(k)}$, $p_{f(k),j}^{(\ell)}$ and c_{ij} would be different but their definitions would not change.

If m=3, then, regardless of the value of m', the set of all distinct triples, instead of pairs, of integers in {1,..,n}, is ordered. If I consists of at least three integers, then the new selection probabilities are conditioned only on the first listed triple in the ordering contained in I.

Otherwise, the new selection probabilities are conditioned on I itself. Thus the new selection probabilities are conditioned on $\binom{n}{3}+\binom{n}{2}+n+1$ events.

To obtain the desired ordering of the triples of integers, first the orderings $f(1),...,f(n)$ and $g_k(1),...,g_k(n-k)$ are constructed exactly as in the case $m=2$. Then, corresponding to each $k=1,...,n-2$, $\ell=1,...,n-k-1$, an ordering $h_{k\ell}(1),...,h_{k\ell}(n-k-\ell)$ of $\{1,...,n\}\sim\{f(1),...,f(k), g_k(1),...,g_k(\ell)\}$ is constructed in a manner similar to the construction of $g_k(1),...,g_k(n-k)$. For example, in defining $h_{k\ell}(v)$ for $v\geq 2$, $p^{(\ell)}_{f(k),j}$ in the definition of $g_k(\ell)$ is replaced by

$$P(f(k),\ g_k(\ell),\ j \in I \text{ and } I \subset (T^*_{k\ell} \cup g_k(\ell))\sim\{h_{k\ell}(1),...,h_{k\ell}(v-1)\}).$$

A linear ordering of the distinct triples in $\{1,...,n\}$ is then determined by representing each triple uniquely as an ordered triple of the form $(f(k), g_k(\ell), h_{k\ell}(v))$. A second triple $(f(k'), g_{k'}(\ell'), h_{k'\ell'}(v'))$ precedes the first if and only if either $k'<k$, or $k'=k$ and $\ell'<\ell$, or $k'=k$ and $\ell'=\ell$ and $v'<v$.

For $m\geq 4$, ordered m-tuples would be defined in a similar manner and the new selection probabilities conditioned on $\binom{n}{m}+\binom{n}{m-1}...+n+1$ events.

For $m=1$, the new selection probabilities are conditioned on the first member of the ordering $f(1),...,f(n)$ in I if $I \neq \emptyset$, or on $\emptyset$ if $I = \emptyset$.

Note that if $m>m'$, it is possible that at least some ordered m-tuples cannot be subsets of I, in which case all such subsets should be excluded from the ordering and the set of events on which the new selection probabilities are conditioned. If no m-tuple can be a subset of I, then

the new selection probabilities are conditioned on I itself.

It is not necessary to limit the initial events used in the transportation problem to subsets of I of size m or less. For example, if m=2 and $\binom{n}{3}+\binom{n}{2}+n+1$ is sufficiently small, then a procedure conditioned on subsets of three or less can be used, resulting in a generally higher expected overlap. Conversely, if $\binom{n}{m}+\binom{n}{m-1}..+n+1$ is too large, the new selection probabilities can be conditioned on subsets of I of size m'' or less, where $m''<m$, although with a generally smaller expected overlap.

4. RELATIONSHIP BETWEEN EXPECTED OVERLAP FOR THE REDUCED-SIZE PROCEDURE, THE OPTIMAL PROCEDURE AND INDEPENDENT SELECTION

Let Ω_I, Ω_R, Ω_O denote the expected overlap for independent selection, the reduced-size procedure, and the optimal procedure, respectively. In this section we explore the relation between these quantities.

We are unaware of any way of determining the values of any of these quantities for a specific problem without actually doing the computations, which in the case of the reduced-size and the optimal procedures requires the solution of the appropriate transportation problem. However, we prove that $\Omega_I \leq \Omega_R$ always. In addition, we obtain upper bounds on Ω_O, lower bounds on Ω_R, and hence upper bounds on $\Omega_O - \Omega_R$. Although these bounds are functions of the overlap problem, they can be computed without solving any transportation problems. These bounds have somewhat limited utility, since they may not be very close to the actual values of Ω_O, Ω_R and $\Omega_O - \Omega_R$. However, as we demonstrate by example, tightening these bounds would not be a simple matter, since under certain conditions Ω_O, Ω_R and $\Omega_O - \Omega_R$

can either equal or get arbitrarily close to these bounds.

In addition, we show by example that in the worst case, in terms of the performance of the reduced-size procedure for two-PSU-per-stratum designs, Ω_O can be arbitrarily close to 2, while Ω_R is arbitrarily close to 0. Thus, at least in theory, the reduced-size procedure can be ineffective. However, in practice, as will be shown in Section 5, Ω_R is much closer to Ω_O than to Ω_I, at least for the SIPP application.

The key results on the relationship between Ω_I, Ω_R and Ω_O will be stated as theorems. Theorem 4.1 holds for any m, m', where m, m' are as in Section 3.2, while the remaining three theorems are only for the case that we have been focusing on, $m = m' = 2$.

Theorem 4.1. $\Omega_I \le \Omega_R \le \Omega_O$.

Proof: See Appendix D.

The next three theorems require the following additional notation. Let μ_2 denote the probability that there are at least two elements in I, and μ_1 denote the probability that I is a singleton set. Let

$$\lambda = \min\{\min\{\pi_i/p_i:\ i=1,...,n\},\ \min\{\pi_{ij}/p_{ij}:\ i,j=1,...,n,\ i\neq j\},\ 1\},$$

$$\lambda' = \min\{\min\{\pi_i/p_i:\ i=1,...,n\},\ \min\{\pi_{st}/p_i^*:\ i=1,...,\binom{n}{2},\ I_i=\{s,t\}\},\ 1\}.$$

Note that $\lambda' \ge \lambda$, since $p_i^* \le p_{st}$ by definition of p_i^* and hence $\pi_{st}/p_i^* \ge \pi_{st}/p_{st}$.

Theorem 4.2. $\Omega_O \le 2\mu_2 + \mu_1$.

Proof: The number of PSUs overlapped cannot exceed 2 when I consists of at least 2 elements, cannot exceed 1 when I is a singleton set, and must be 0 when $I=\varnothing$.

Theorem 4.3.

(*a*) $\Omega_R \geq \lambda(2\mu_2 + \mu_1/2)$,

(*b*) $\Omega_R \geq \lambda'(2\mu_2 + \mu_1/2)$.

Proof: See Appendix D.

Note that since $\Omega_R \leq \Omega_O$, the bound in Theorem 4.2 is also an upper bound on Ω_R and the bounds in Theorem 4.3 are also lower bounds on Ω_O. Also note that (b) in Theorem 4.3 is a tighter bound than (a) since $\lambda' \geq \lambda$. However, it is easier to compute λ since, unlike λ', this does not require computation of the p_i^*'s.

Theorem 4.4.

(a) $\Omega_O - \Omega_R \leq 2(1-\lambda)\mu_2 + (1-\lambda/2)\mu_1$,

(b) $\Omega_O - \Omega_R \leq 2(1-\lambda')\mu_2 + (1-\lambda'/2)\mu_1$.

Proof: Combine Theorems 4.2 and 4.3.

In particular, if $\lambda=1$ or $\lambda'=1$, which will occur when all the relevant probabilities in the definition of λ or λ' are greater in the new design than in the initial design, then $\Omega_O - \Omega_R \leq \mu_1/2$. If, in addition, μ_1 is small then Ω_R must be close to Ω_O.

To illustrate Theorems 4.2 - 4.4 and the fact the bounds that they give may not always be useful, consider the example of Section 2.2 and Section 3.1. From Table 3 it can be seen that

$\mu_2 = .765$, $\mu_1 = .205$. We also have from Tables 3 and 7 that $\lambda = .476$, $\lambda' = .833$. Then the bound in Theorem 4.2 is 1.735. In Theorem 4.3 (a) and (b) the bounds are .777 and 1.360, respectively. In Theorem 4.4 (a) and (b) they are .958 and .375, respectively. This compares to exact values of 1.735 for Ω_O, 1.725 for Ω_R, and .01 for $\Omega_O - \Omega_R$. While the bound on Ω_O is equal to its exact value in this example, the bounds on Ω_R and $\Omega_R - \Omega_O$ are not close to their exact values. In fact, since $\Omega_I = 1.39$ for this example, the lower bound on Ω_R of 1.39 guaranteed by Theorem 4.1 is greater than the bounds given by Theorem 4.3.

The bounds of Theorem 4.3 are examined further in Section 5, using SIPP data. The results there also indicate that the bound in (a) is of little practical utility, but the bound in (b) may be of some use since it guaranteed a mean expected overlap for the reduced-size procedure of .5 PSUs/stratum more than independent selection.

The previous example illustrates the difficulty in improving on the bound of Theorem 4.2, since Ω_O is equal to the bound for this example. (This is not always the case. For the example in Table 11, Ω_O is less than this upper bound, since if I={1,3} it is not always possible to retain {1,3} in the final sample.)

The following example illustrates the difficulty in improving upon the bounds of Theorem 4.3. Consider a new stratum S with n=4. All of the PSUs were in different initial strata. Let $p_1 = p_2 = .5$ and $p_3 = p_4 = \varepsilon$ for a small value ε. Let $\pi_{12} = .5$, $\pi_{13} = \pi_{14} = \pi_{23} = \pi_{24} = \varepsilon$, $\pi_{34} = .5 - 4\varepsilon$. It can be shown that for this example, $\mu_2 = .25 + \varepsilon - .25\varepsilon^2$, $\mu_1 = .5 - .5\varepsilon$, and $\lambda = \lambda' = 1$. Consequently, the lower bound on Ω_R in both Theorem 4.3 (a) and (b) is $.75 + 1.75\varepsilon - .5\varepsilon^2$. However, in Appendix E it is established that

$$\Omega_O \leq .75 + 12\varepsilon. \qquad (4.1)$$

Consequently, since $\Omega_R \leq \Omega_O$, the actual value of Ω_R exceeds the lower bound on Ω_R by at most $10.25\varepsilon + .5\varepsilon^2$, which for small ε is a negligible amount both in absolute and relative terms.

Finally, the following example illustrates a worst case situation for Ω_R in relation to both Ω_I and Ω_O. Let *S* consist of *n* PSUs, in *n* different initial strata, with $\pi_i = 2/n$, $\pi_{ij} = 2/[n(n-1)]$, $p_i = c$, $i,j = 1,...,n$, $i<j$, where $c<1$ is a constant. For independent selection for this example, the probability of overlap for each PSU selected in the new sample is *c* and, therefore, $\Omega_I = 2c$. Furthermore, in Appendix E it is shown that

$$\lim_{n\to\infty} \Omega_0 = 2, \tag{4.2}$$

$$\lim_{n\to\infty} \Omega_R = 2c. \tag{4.3}$$

Thus Ω_R exceeds Ω_I by a negligible amount for large *n*. Furthermore, since $\lim_{n\to\infty} (\Omega_O - \Omega_R) = 2 - 2c$ by (4.2) and (4.3), $\Omega_O - \Omega_R$ can be made arbitrarily close to 2 by making *c* small enough and *n* large enough. In addition, since neither of the bounds in Theorem 4.4 can exceed 2, this example demonstrates that it is possible for the exact value of $\Omega_O - \Omega_R$ to be arbitrarily close to the upper bounds of Theorem 4.4.

5. APPLICATION OF REDUCED-SIZE PROCEDURE TO SIPP

Results from simulations of the SIPP overlap, done prior to production for research and testing purposes, are presented in 5.1. Results from the actual SIPP production overlap are presented in 5.2. Further details are given in Ernst and Ikeda (1992b).

5.1 Simulation Results

In the implementation of the reduced-size overlap procedure, minimum cost flow (MCF) optimization software, written by Darwin Kingman and John Mote at the University of Texas at Austin, was used to solve the required transportation problem. A FORTRAN program was written to produce input to and process output from the MCF software.

To test the software prior to production, the program was used to overlap two stratifications, based on 1970 census data, of the SIPP Midwest region with the actual 1980s design stratification for the SIPP Midwest region. (At the time of this test, 1990 census data was not yet available.) The 1970-based stratifications were produced by stratifying the 1980s SIPP noncertainty PSUs in the Midwest region using 1970 data. Both of the 1970-based stratifications partitioned the noncertainty PSUs into 31 strata, using different sets of stratification variables. The stratifications based on 1980 and 1970 data were treated as "initial" and "final" stratifications for the purposes of the overlap algorithm.

The expected overlap was calculated for the reduced-size maximum overlap algorithm, for independent selection of final PSUs, and for the upper bound to the expected overlap for the optimal procedure given by Theorem 4.2. That upper bound was calculated instead of the actual optimal overlap, since the optimal overlap cannot be calculated for the larger strata.

The results from the two final stratifications in the simulation were generally similar to each other. Combining the results from both stratifications, the reduced-size maximum overlap algorithm had a mean expected overlap of 1.552 PSUs/stratum for this set of 62 strata, with a range from 1.257 to 1.762. The upper bound to the expected overlap had a mean of 1.569 PSUs/stratum, with a range from 1.260 to 1.809. The largest difference between the expected overlap under the reduced-size maximum overlap algorithm and the upper bound to the expected overlap was .084 PSUs. The expected overlap for independent selection had a mean of .480 PSUs, with a range from .088 to 1.214. The reduced-size maximum overlap algorithm always gave substantial improvement over independent selection, with a range of

increase from .455 to 1.464 PSUs. Thus, for this set of 62 strata, the expected number of PSUs overlapped is 29.8 for independent selection, 96.2 for the reduced-size procedure, and at most 97.3 for the optimal procedure.

We also computed the lower bounds for Theorem 4.3 for those 62 strata. The mean lower bound was .416 for the bound in (a) of that theorem and .980 for (b). This further illustrates that the bound in (a) appears to be of no practical value, since it is below, on average, that of independent selection. The bound in (b) is of some use, since it does guarantee an expected gain of .5 PSUs/stratum over independent selection by using the reduced-size procedure, which was 46.6 percent of the expected gain actually attained.

The reduced-size algorithm took a fairly short time to run on most strata. The CPU times for final strata with different numbers of PSUs are given below. The reduced-size program was run on a Solbourne 5/605 computer. The median number of PSUs in a stratum, for the entire group of 62 strata, was 17 PSUs. The 37 PSUs stratum had the 6th largest number of PSUs. The 68 PSUs stratum was the largest stratum.

Table 12. CPU Times for Reduced-Size Procedure

Number of PSUs	CPU Time (hrs:min:sec)
18	0:36
37	5:44
49	24:05
68	2:23:43

5.2 Implementation in the 1990s SIPP Design

In the actual implementation, as noted in Section 3.1 and detailed in Appendix C, a modification of the reduced-size procedure was used to overlap the 1990s SIPP design with the 1980s SIPP design, because the PSU definitions in the 1980s and 1990s designs were not identical.

The modified reduced-size procedure was used to overlap 103 final (1990s design) nonselfrepresenting strata in SIPP. The average expected overlap was 1.523 PSUs/stratum compared to 0.582 for independent selection. Two strata (with 69 and 72 PSUs) were not overlapped since they exceeded the cutoff of 57 PSUs used during production, which employed a different computer than used in the simulations, with a more restricted memory allocation. There are also 112 selfrepresenting strata in the 1990s design for which overlap procedures are not applicable. Thus, there are a total of 322 sample PSUs in this design. In addition, there are 1606 nonsample PSUs in the design.

As we did for the simulation study described in the previous subsection, we calculated an upper bound for the expected overlap for each production SIPP final stratum that was overlapped. The mean upper bound for the 103 strata was 1.647 PSUs/stratum, reasonably close to the mean expected overlap of 1.523 using the production overlap procedure. Thus, among the 103 strata that were overlapped using the reduced-size procedure, the expected number of PSUs overlapped with this procedure was 156.9, compared to 59.9 for independent selection and at most 169.6 for the optimal procedure.

Because of the changes resulting from the fact that the two designs did not have identical PSUs definitions, it was necessary to modify the upper bound given by Theorem 4.2 to obtain the 1.647 mean upper bound on the expected overlap. This is because, as noted in Appendix *C*, if both new sample PSUs intersect the same initial sample PSU, this event is counted as two successful overlaps. As a result, when *I* is a singleton set it is possible that there can be 2 PSUs overlapped, which is not the case when the PSU definitions are the same in the new design. Consequently, the μ_1 term in the upper bound $2\mu_2 + \mu_1$ is no longer valid. Instead, we let μ_1' denote the probability that *I* is a singleton set corresponding to an initial PSU which intersects at least two final PSUs in *S*, and μ_1'' denote the probability that *I* is a singleton set corresponding to an initial PSU which intersects exactly one final PSU. We then used the valid upper bound $2\mu_2 + 2\mu_1' + \mu_1''$.

The computer time during production for the modified reduced-size overlap program was reasonably short. Production was done in four computer runs, one for each region of the country. The maximum clock time for a region (44 strata, the largest consisting of 46 PSUs) was 1 hour and 40 minutes. The CPU time is not known, but believed to be only slightly less than the clock time.

We also calculated that of the 103 final strata overlapped by the modified reduced-size procedure, 41 would not have run under the optimal procedure. This calculation was based on our estimate that the maximum size transportation problem, in terms of number of variables, that could have run in production was 4×10^6. The number of variables for the optimal procedure was less than 4×10^6 for all 56 strata for which $n\leq 14$, but exceeded this limit for all but 6 of the 47 strata with $n\geq 15$, including two with n=15. The largest strata for which the optimal procedure could have run has 19 PSUs. Of the 41 strata for which the optimal procedure would not have run, 37 had transportation problems for the optimal procedure with more than 10^7 variables, 33 with more than 10^8 variables and 23 with more than 10^9 variables. The maximal size of the transportation for the optimal procedure among the 103 strata occurred for a stratum with n=46, for which there were 3.61×10^{12} variables. In contrast, there were 1.03×10^6 variables for the modified reduced-size procedure for this stratum.

In performing these size calculations for the optimal procedure, Theorem 2.1 was used with the following modification to account for different PSU definitions in the two designs. In computing m^*, n_i is now the number of PSUs in initial stratum i that intersect PSUs in S, rather than the number of PSUs in the stratum that are in S. In particular, the maximum value of m^* is now $2^{n''}$, where n'', using the notation in Appendix C, is the number of initial PSUs that intersect PSUs in S. Furthermore, for the modified reduced-size procedure the number of variables is $\left[\binom{n'}{2}+n'+1\right]\times\binom{n}{2}$, where n', as explained in Appendix C, is the

number of PSUs in S matched to initial PSUs.

From Theorem 2.1, it may be surmized that for fixed n, m^* tends to increase with the number of initial strata r that have PSUs which intersect PSUs in S. A rather striking example of this relationship occurred for two of the SIPP strata. For one of these strata n=25, r=4 and the number of variables for the optimal procedure was 9.01×10^6, while for the other stratum n=24, r=18, and the number of variables was 4.29×10^{11}. We also had that n''=24 for the former stratum and n''=33 for the latter stratum, which would explain part, but not all of the large difference in number of variables for the two strata.

Another question of interest is the overlap effectiveness of the reduced-size procedure in comparison with the overlap procedure of Ernst (1986). In general it is believed that the reduced-size procedure should produce a higher overlap in situations when both are usable, since the reduced-size procedure makes use of the stratum-to-stratum independence in the initial design. However, although the procedure in Ernst (1986) is applicable to two-PSU-per-stratum designs, no computer program has ever been written at the Census Bureau (or anywhere else that the authors are aware of) to implement this procedure for such designs, since there has not yet been a production application for this program. Consequently, we cannot make a direct comparison of these two methods on the same data. However, a crude comparison can be made from the results of the reduced-size overlap procedure for SIPP data and the results of the overlap using the procedure in Ernst (1986) for the overlap of 1990s CPS and NCVS designs with their respective 1980s designs. (Both the 1980s and 1990s designs for CPS and NCVS are one-PSU-per-stratum designs.)

For CPS, the overlap procedure resulted in an average increase in expected overlap, in comparison with independent selection, of .26 PSUs/stratum, and for NCVS the overlap procedure resulted in an average increase in expected overlap of .30 PSUs/stratum. This compares with an increase of .94 PSUs/stratum for the reduced-size procedure over independent selection for SIPP. If the two overlap procedures are equally effective, then one

might expect that the increase in overlap per stratum for SIPP would be roughly twice as large as for CPS and NCVS, since SIPP has a two-PSUs-per-stratum design. By this standard, the reduced-size procedure program performs better than the procedure in Ernst (1986). However, since the stratifications were quite different for these three surveys, the validity of this comparison is open to question.

For the examples in Tables 6 and 11, a valid comparison of the different overlap procedures can be made, since the expected overlap values for the procedure in Ernst (1986), 1.625 for the Table 6 example and 1.425 for the Table 11 example, were easily calculated by hand. For the reduced-size procedure the corresponding overlap values are 1.725 and 1.58 respectively, and for the optimal procedure, 1.735 and 1.58, respectively.

In summary, we believe the reduced-size procedure to be a practical procedure which, although in theory can be ineffective in increasing overlap, yields results reasonably close to optimal in practice. It can be only used when the PSUs in the initial design are selected independently from stratum to stratum, but when this condition is met we believe it is the overlap procedure of choice for large strata.

APPENDIX A: PROOF OF THEOREM 2.1

Let $F_i = \{j: A_j \in G_i\}$, i=1,...,r. Since the sampling of PSUs is assumed to have been independent from stratum to stratum in the initial design, to establish (2.13) it suffices to show that m_i is the number of possible values for $F_i \cap I$ in cases (2.14) - (2.17). Now (2.14) holds, since it is always possible for each element in F_i to have either been in I or not have been in I. (2.15) is the case when $G_i = G_i'$ and hence $F_i \cap I$ can be any of the $\binom{n_i}{2}$ subsets of F_i of size 2. To obtain (2.16), note that $F_i \cap I$ can either be one of the $\binom{n_i}{2}$ subsets of size 2 of F_i or one of the n_i singleton subsets of F_i, the latter event occurring when the one PSU in $G_i' \sim G_i$ was in the intial sample. (2.17) is similar, except that since $G_i' \sim G_i$ now consists of at least two PSUs, both initial sample PSUs in G_i' can be in $G_i' \sim G_i$, in which case $F_i \cap I = \varnothing$, creating one additional possibility for $F_i \cap I$.

To show under what conditions the upper bound of 2^n for m^* is obtained, observe that the number of possible values for $F_i \cap I$ cannot exceed 2^{n_i}. Since $\prod_{i=1}^{n} 2^{n_i} = 2^n$, it follows that $m^* = 2^n$ if and only if $m_i = 2^{n_i}$ for all i. Now if n_i = 1, then $m_i = 2^1$ by (2.14). If n_i = 2, then $m_i = 2^2$ if and only if $m_i = \binom{n_i}{2} + n_i + 1$, which by (2.15) - (2.17) occurs if and only if $n_i' \geq 4$. Finally, if $n_i \geq 3$ for some i, then F_i itself is not a possible value for I. Consequently, the possible values of I do not include all subsets of $\{1,...,n\}$ and, therefore, $m^* < 2^n$.

APPENDIX B: PROOF OF (3.16) - (3.21)

Note that since $\{f(k), g_k(\ell)\}$ is the first pair in the ordering contained in I, we have $f(k),\ g_k(\ell) \in I,$ but none of $f(1),...,f(k\text{-}1),\ g_k(1),...,g_k(\ell\text{-}1)$ are in I, and hence (3.16), (3.17) follow. Thus b_{it} is determined for all t except those in $T_{k\ell} \sim \{g_k(\ell)\},$ for which b_{it} is computed by one of (3.18) - (3.21). Now if $f(k)$, $g_k(\ell)$ and t were all in F_α then $t \notin I$ since there are only two initial sample PSUs in each initial stratum and hence (3.18) holds. To obtain (3.19), observe that if $f(k),\ t \in F_\alpha,$ but $g_k \notin F_\alpha,$ then we know that $f(k) \in I,\ I \cap F_\alpha \subset T_{k\ell}^*,$ and hence

$$b_{it} = P(t \in I \mid f(k) \in I \text{ and } I \cap F_\alpha \subset T_{k\ell}^*)$$

$$= \frac{P(t, f(k) \in I \text{ and } I \cap F_\alpha \subset T_{k\ell}^*)}{P(f(k) \in I \text{ and } I \cap F_\alpha \subset T_{k\ell}^*)} = \frac{p_{f(k),t}}{p''_{f(k),\alpha}(T_{k\ell}^*)}.$$

(3.20) is obtained similarly to (3.19), while (3.21) follows since

$$b_{it} = P(t \in I \mid I \cap F_\gamma \subset T_{k\ell}^*) = \frac{P(t \in I \text{ and } I \cap F_\gamma \subset T_{k\ell}^*)}{P(I \cap F_\gamma \subset T_{k\ell}^*)} = \frac{p''_{t\gamma}(T_{k\ell}^*)}{p'_\gamma(T_{k\ell}^*)}.$$

APPENDIX C: MODIFICATION OF REDUCED SIZE PROCEDURE WHEN PSU DEFINITIONS CHANGE

The procedure described in Section 3.1 requires two modifications when there are different PSU definitions in the two designs.

The first modification is a procedure for establishing a one-to-one correspondence between a subset of the n PSUs in a final stratum S and a subset of the PSUs in the initial design, which is needed in ordering pairs of PSUs. The natural one-to-one correspondence between all the PSUs in S and the set of PSUs in the initial design that exists when the PSU definitions are the same in the two designs no longer holds, and such terminology as the initial and new selections probability for A_i no longer would make sense unless this correspondence is restored.

The second modification is a change in the calculation of the b_{it}'s to account for the possibility of several initial PSUs intersecting with one final PSU.

The one-to-one correspondence is created as follows. Order the n PSUs in S, that is $A_1,...,A_n$, in descending order of new selection probability. Match A_1 to the initial PSU that makes up the largest portion (using the measure of size for the new design) of A_1. Then proceed to match $A_2,...,A_n$, where A_i is matched to the initial PSU that makes up the largest portion of A_i among those initial PSUs intersecting A_i that have not already been matched. If no such initial PSU exists then A_i is not matched. For example, if $j<i$ and A_j, A_i were formed by splitting an initial PSU, then A_j is matched to this PSU and A_i is unmatched. Let $B_1,...,B_{n''}$ denote the initial PSUs that intersect at least one of the A_i's. Assume that n' of the PSUs in S are matched by this process, with A_i matched to B_i, $i=1,...,n'$.

The ordering of the pairs of PSUs and the formulation of the reduced-size transportation problem proceeds as in Section 3.1 with the following modifications. I is now the set of initial PSUs in

$\{B_1,...,B_{n''}\}$ that were in the initial sample, while $I' = I \cap \{B_1,...,B_{n'}\}$. The ordering is of pairs of PSUs in I', that is for matched PSUs only. Let F_α, α=1,...,r, denote a partition of $\{B_1,...,B_{n''}\}$ according to their initial stratum, with F'_α, α=1,...,r', the analogous partition for $\{B_1,...,B_{n'}\}$. The I_i's now consist of the subsets of $\{B_1,...,B_{n'}\}$ of two or fewer elements, and associated with each I' is a subset I^*, namely the first I_i contained in I'. That is, there are now $\binom{n'}{2} + n' + 1$ sets on which the new selection probabilities are conditioned. As a result, in determining the ordering of the pairs of PSUs and the calculation of p_i^*, the definitions of $f(k)$, $g_k(\ell)$, T_k, $T_{k\ell}$, $T^*_{k\ell}$, $p_i^{(k)}$, $p'_\alpha(T)$, $p''_{i\alpha}(T)$, $p^{(\ell)}_{f(k),j}$, and p_i^* are modified by replacing n, I, F_α, r by n', I', F'_α, r', respectively. Also n is replaced by n' in the i index in (3.2), (3.3), but the j index remains unchanged.

It remains only to explain how the c_{ij}'s are defined and calculated under this modified procedure. A PSU A_t, t=1,...,n, selected in the new sample is considered a successful overlap if any of the B_j, j=1,...,n'', that intersect it, even if not matched to A_t, were in I. (In particular, if two final sample PSU intersect the same initial sample PSU, this counts as two successful overlaps. Some may prefer to count this as only one successful overlap, since there generally can be only one interviewer retained in this case. However, this would complicate the calculation of the c_{ij}'s, since the relation $c_{ij} = b_{is} + b_{it}$ given below would no longer hold.) Consequently, c_{ij} is the expected number of PSUs in S_j that intersect PSUs in I given $I^*=I_i$. Hence, if

$$H_t = \{k:\ B_k \cap A_t \neq \varnothing,\ k=1,...,n''\}, \quad t=1,...,n,$$

$$b_{it} = P(H_t \cap I \neq \varnothing \mid I^* = I_i), \quad i=1,...,\binom{n'}{2} + n' + 1, \quad t=1,...,n,$$

and $S_j = \{s,t\}$, then $c_{ij} = b_{is} + b_{it}$.

The calculation of b_{it} is more complex here than in Section 3.1, because of the possible multiple intersections. For $i=1,...,\binom{n'}{2}+n'+1$, $t=1,...,n$, $\alpha=1,...,r$,

let $H_{t\alpha} = H_t \cap F_\alpha$,

$$b_{it\alpha} = P(H_{t\alpha} \cap I \neq \varnothing \mid I^* = I_i), \qquad b_{it\alpha j} = P(j \in I \mid I^* = I_i), \quad j \in H_{t\alpha},$$

and

$$b_{it\alpha jk} = P(j,k \in I \mid I^* = I_i), \quad j,k \in H_{t\alpha}, \quad j \neq k.$$

Observe that

$$b_{it} = 1 - \prod_{\alpha=1}^{r} (1 - b_{it\alpha})$$

and

$$b_{it\alpha} = \sum_{j \in H_{t\alpha}} b_{it\alpha j} - \sum_{\substack{j,k \in H_{t\alpha} \\ j<k}} b_{it\alpha jk}.$$

Thus we have reduced the problem to the calculation of $b_{it\alpha j}$ and $b_{it\alpha jk}$. To do this, let $n_{i\alpha}$ be the number of elements in $I_i \cap F_\alpha$,

$$T' = \{n'+1,...,n''\} \cup I_i \quad \text{if } n_{i\alpha} \leq 1,$$

$$= \{n'+1,...,n''\} \cup T^*_{k\ell} \quad \text{if } I_i = \{f(k),\ g_k(\ell)\},$$

$$T'' = T' \sim I_i.$$

Then with $p'_\alpha(T)$, $p''_{i\alpha}(T)$ as in their original definition in Section 3.1 except that n is replaced by n'', we have

$$
\begin{aligned}
b_{it\alpha j} &= 1 \quad \text{if } j \in I_i, \\
&= 0 \quad \text{if } j \notin T', \\
&= 0 \quad \text{if } j \in T'' \text{ and } n_{i\alpha} = 2, \\
&= \frac{p_{vj}}{p''_{v\alpha}(T')} \quad \text{if } j \in T'' \text{ and } I_i \cap F_\alpha = \{v\}, \\
&= \frac{p''_{j\alpha}(T')}{p'_\alpha(T')} \quad \text{if } j \in T'' \text{ and } n_{i\alpha} = 0,
\end{aligned}
$$

and

$$
\begin{aligned}
b_{it\alpha jk} &= 1 \quad \text{if } j,k \in I_i, \\
&= 0 \quad \text{if } j \notin T' \text{ or } k \notin T', \\
&= 0 \quad \text{if } j,k \in T'' \text{ and } n_{i\alpha} \geq 1, \\
&= \frac{p_{jk}}{p'_\alpha(T')} \quad \text{if } j,k \in T'' \text{ and } n_{i\alpha} = 0 \\
&= 0 \quad \text{if } j \in T'', \ k \in I_i \text{ and } n_{i\alpha} = 2 \\
&= \frac{p_{jk}}{p''_{k\alpha}(T')} \quad \text{if } j \in T'', \ k \in I_i \text{ and } n_{i\alpha} = 1.
\end{aligned}
$$

APPENDIX D: PROOF OF THEOREMS 4.1 AND 4.3

Proof of Theorem 4.1. Since by (2.10), the optimal procedure maximizes the expected number of PSUs that are in both the initial and new samples, we must have $\Omega_R \le \Omega_0$.

To show that $\Omega_I \le \Omega_R$, consider the formulation (3.1) - (3.3). Let

$$x_{ij} = p_i^* \pi_j^*, \quad i=1,...,\binom{n}{2}+n+1, \quad j=1,...,\binom{n}{2}. \tag{D.1}$$

(D.1) satisfies (3.2), (3.3). Since Ω_R maximizes (3.1) over all such sets of x_{ij}'s, Ω_R must be at least the expected overlap for (D.1). However, for this set of x_{ij}'s, we have $P(N = S_j \mid I^* = I_i) = \pi_j^*$ for all i,j; that is the conditional selection probabilities equals the unconditional. Therefore, (D.1) corresponds to independent selection of the new sample PSUs, and hence for this set of x_{ij}'s, (3.1) equals $\sum_{i=1}^{n} p_i \pi_i$. Consequently, $\Omega_I \le \Omega_R$.

Proof of Theorem 4.3. We will prove (b) of this Theorem. Then (a) immediately follows, since $\lambda' \ge \lambda$. Label the S_j's so that $S_j = I_j$, $j=1,...,\binom{n}{2}$. Let

$$\delta_i = \{j: S_j \cap I_i \neq \varnothing\}, \quad i = \binom{n}{2}+1,...,\binom{n}{2}+n;$$

that is for each i for which I_i is a singleton set, δ_i is the set of all j for which the element in I_i is in the pair S_j. δ_i consists of n-1 elements.

Next let

$$x'_{ii} = \lambda' p_i^*, \quad i=1,\ldots,\binom{n}{2}, \tag{D.2}$$

$$x'_{ij} = \frac{\lambda' p_i^* (\pi_j^* - \lambda' p_j^*)}{2\sum_{k\in\delta_i} (\pi_k^* - \lambda' p_k^*)} \quad \text{if } \sum_{k\in\delta_i} (\pi_k^* - \lambda' p_k^*) \neq 0,$$

$$= 0 \quad \text{otherwise,} \quad i=\binom{n}{2}+1,\ldots,\binom{n}{2}+n, \quad j\in\delta_i, \tag{D.3}$$

$$x'_{ij} = 0 \quad \text{for all other } i,j, \quad i=1,\ldots,\binom{n}{2}+n+1, \quad j=1,\ldots,\binom{n}{2}, \tag{D.4}$$

$$p_i^{**} = p_i^* - \sum_{j=1}^{\binom{n}{2}} x'_{ij}, \quad i=1,\ldots,\binom{n}{2}+n+1, \tag{D.5}$$

$$\pi_j^{**} = \pi_j^* - \sum_{i=1}^{\binom{n}{2}+n+1} x'_{ij}, \quad j=1,\ldots,\binom{n}{2}, \tag{D.6}$$

$$x''_{ij} = \frac{p_i^{**}\pi_j^{**}}{\sum_{k=1}^{\binom{n}{2}} \pi_k^{**}}, \quad \text{if } \sum_{k=1}^{\binom{n}{2}} \pi_k^{**} \neq 0,$$

$$= 0 \quad \text{otherwise,} \quad i=1,\ldots,\binom{n}{2}+n+1, \quad j=1,\ldots,\binom{n}{2}, \tag{D.7}$$

$$x_{ij} = x'_{ij} + x''_{ij}, \quad i=1,\ldots,\binom{n}{2}+n+1, \quad j=1,\ldots,\binom{n}{2}. \tag{D.8}$$

To establish Theorem 4.3(b), it suffices to show that $x_{ij} \geq 0$ for all i,j; that (D.8) satisfies (3.2) and

(3.3); and that (3.1) is at least $\lambda'(2\mu_2 + \mu_1/2)$ for (D.8).

To show that $x_{ij} \geq 0$, observe that $x'_{ij} \geq 0$ by (D.2) - (D.4) and the definition of λ'. Hence it suffices to show that $x''_{ij} \geq 0$, which by (D.7) will follow if it is established that $p_i^{**} \geq 0$ for all i and $\pi_j^{**} \geq 0$ for all j. Now by (D.2) - (D.5) and the fact that $\lambda' \leq 1$, we have

$$p_i^{**} = p_i^{*} - \sum_{j=1}^{\binom{n}{2}} x'_{ij} = p_i^{*} - x'_{ii} = p_i^{*} - \lambda' p_i^{*} \geq 0, \qquad i,\ldots,\binom{n}{2},$$

$$p_i^{**} = p_i^{*} - \sum_{j \in \delta_i} x'_{ij} \geq p_i^{*} - \lambda' p_i^{*}/2 \geq 0, \qquad i = \binom{n}{2} + 1,\ldots,\binom{n}{2} + n,$$

and

$$p_i^{**} = p_i^{*} \geq 0, \qquad i = \binom{n}{2} + n + 1.$$

To show that $\pi_j^{**} \geq 0$ for all j, we first establish that

$$x'_{ij} \leq \frac{\pi_j^{*} - \lambda' p_j^{*}}{2}, \qquad i = \binom{n}{2} + 1,\ldots,\binom{n}{2} + n, \quad j \in \delta_i. \tag{D.9}$$

To obtain (D.9), let s denote the single element in I_i; observe that

$$\lambda'\left(p_i^{*} + \sum_{k \in \delta_i} p_k^{*}\right) \leq \lambda' p_s \leq \pi_s = \sum_{k \in \delta_i} \pi_k^{*}.$$

hence

$$\lambda' p_i^{*} \leq \sum_{k \in \delta_i} (\pi_k^{*} - \lambda' p_k^{*}). \tag{D.10}$$

Then combine (D.10) with (D.3) to complete the proof of (D.9).

Then for $j=1,\ldots,\binom{n}{2}$, let I_{i_1}, I_{i_2} be the two singleton subsets of S_j. From (D.2) - (D.4), (D.6), (D.9) we conclude

$$\pi_j^{**} = \pi_j^* - \sum_{i=1}^{\binom{n}{2}+n+1} x'_{ij} = \pi_j^* - (x'_{jj} + x'_{i_1 j} + x'_{i_2 j}) \geq 0,$$

and therefore $x_{ij} \geq 0$ for all i,j.

Next, to show that (D.8) satisfies (3.2), first observe that

$$\sum_{i=1}^{\binom{n}{2}+n+1} p_i^{**} = \sum_{i=1}^{\binom{n}{2}+n+1} p_i^* - \sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} x'_{ij}$$

$$= 1 - \sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} x'_{ij} = \sum_{j=1}^{\binom{n}{2}} \pi_j^* - \sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} x'_{ij} = \sum_{j=1}^{\binom{n}{2}} \pi_j^{**}. \qquad \text{(D.11)}$$

Then if $\sum_{k=1}^{\binom{n}{2}} \pi_k \neq 0$, combine (D.8), (D.7), (D.5) to obtain

$$\sum_{j=1}^{\binom{n}{2}} x_{ij} = \sum_{j=1}^{\binom{n}{2}} (x'_{ij} + x''_{ij}) = \sum_{j=1}^{\binom{n}{2}} x'_{ij} + p_i^{**} = p_i^{*}, \tag{D.12}$$

while otherwise use the same relations and (D.11) to conclude (D.12).

To establish that (D.8) satisfies (D.3), note that by (D.11) we can substitute

$\sum_{k=1}^{\binom{n}{2}+n+1} p_k^{**}$ for $\sum_{k=1}^{\binom{n}{2}} \pi_k^{**}$ in (D.7), and then proceed to establish (3.3) analogously to (3.2).

Finally, to show that (3.1) is at least $\lambda'(\mu_2 + \mu_1/2)$ for (D.8), first observe that for

$i = \binom{n}{2}+1, \ldots, \binom{n}{2}+n$, $\sum_{j \in \delta_i} x'_{ij} = \lambda' p_i^{*}/2$ by (D.3) if $\sum_{k \in \delta_i} (\pi_k - \lambda' p_k^{*}) \neq 0$, while $\sum_{j \in \delta_i} x'_{ij} = \lambda' p_i^{*}/2 = 0$ by (D.3),

(D.10) otherwise. We then combine this last result with (D.2) - (D.4) to conclude

$$\sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} c_{ij} x_{ij} \geq \sum_{i=1}^{\binom{n}{2}+n+1} \sum_{j=1}^{\binom{n}{2}} c_{ij} x'_{ij}$$

$$\geq 2\sum_{i=1}^{\binom{n}{2}} x'_{ii} + \sum_{i=\binom{n}{2}+1}^{\binom{n}{2}+n} \sum_{j \in \delta_i} x'_{ij} = 2\lambda' \sum_{i=1}^{\binom{n}{2}} p_i^{*} + \frac{\lambda'}{2} \sum_{i=\binom{n}{2}+1}^{\binom{n}{2}+n} p_i^{*} = \lambda'(2\mu_2 + \mu_1/2).$$

APPENDIX E: PROOF OF (4.1) - (4.3)

Proof of (4.1). To establish the inequality we simply show that

$$\sum_{i=1}^{m^*} \sum_{j=1}^{n^*} c_{ij} x_{ij} \le .75 + 12\varepsilon \tag{E.1}$$

for any nonnegative x_{ij}'s satisfying (2.11), (2.12), where c_{ij}, m^*, n^* are as in (2.10). Note that for this example m^*=16, n^*=6. Let S_1 = {1,2}, S_2 = {1,3}, S_3 = {1,4}, S_4 = {2,3}, S_5 = {2,4}, S_6 = {3,4}. We establish (E.1) by considering the contributions to the left hand side of (E.1) from j=2,3,4,5, j=6, and j=1, separately.

We first use (2.12) to obtain

$$\sum_{i=1}^{m^*} c_{ij} x_{ij} \le 2 \sum_{i=1}^{m^*} x_{ij} = 2P(S_j) = 2\varepsilon, \quad j = 2,3,4,5. \tag{E.2}$$

Next let $D = \{i: \{3,4\} \cap J_i \neq \varnothing\}$. Since c_{i6}=0 if $i \notin D$, it follows from (2.11) that

$$\sum_{i=1}^{m^*} c_{i6} x_{ij} \le 2 \sum_{i \in D} x_{i6} \le 2 \sum_{i \in D} P(J_i) \le 2(p_3 + p_4) = 4\varepsilon. \tag{E.3}$$

Finally, let $D_1 = \{i: 1 \in J_i$ or $2 \in J_i$, but $\{1,2\} \not\subset J_i\}$ and $D_2 = \{i: \{1,2\} \subset J_i\}$. Then

$$\sum_{i=1}^{m^*} c_{i1} x_{i1} = \sum_{i \in D_1} x_{i1} + 2 \sum_{i \in D_2} x_{i1}. \tag{E.4}$$

Now by (2.12),

$$\sum_{i \in D_1} x_{i1} + \sum_{i \in D_2} x_{i1} \le P(S_1) = .5, \tag{E.5}$$

while by (2.11),

$$\sum_{i\in D_2} x_{i1} \le \sum_{i\in D_2} P(J_i) = p_1 p_2 = .25 \tag{E.6}$$

Then we combine (E.4) - (E.6) to obtain

$$\sum_{i=1}^{m^*} c_{i1} x_{i1} \le .75, \tag{E.7}$$

and then combine (E.2), (E.3) and (E.7) to conclude (E.1).

Proof of (4.2). The following sampling procedure for the new sample will yield the optimal overlap for this example. When there are at least two PSUs in *I*, select a pair among all pairs in *I* with equal probabilities; if *I* is a singleton set, then select with equal probability among the *n*-1 pairs that contain *I*. If $I=\varnothing$ then select with equal probability among all $n(n-1)/2$ pairs in *S*. By symmetry, this procedure will yield $\pi_{ij} = 2/[n(n-1)]$ for all i,j, $i<j$ as required. Furthermore, the procedure is optimal since the expected overlap for it is $2\mu_2 + \mu_1$, which is an upper bound on Ω_0 by Theorem 4.2. In addition, for this procedure, $P(I=\varnothing) = (1-c)^n$ and $\mu_1 = nc(1-c)^{n-1}$. Consequently, $\mu_2 = 1 - [nc(1-c)^{n-1} + (1-c)^n]$ and

$$\Omega_0 = 2\mu_2 + \mu_1 = 2 - [nc(1-c)^{n-1} + 2(1-c)^n].$$

Then (4.2) follows from this last relation with the aid of L'Hospital's rule.

Proof of (4.3). Because p_i, π_i, p_{ij}, π_{ij} are independent of i,j for this example, and each PSU in *S*

was in a different initial stratum, any ordering of the pairs would yield Ω_R. Therefore, we let

$f(k) = k,\ k=1,...,n,\quad g_k(\ell) = k+\ell,\ k=1,...,n-1,\ \ell=1,...,n-k.$

Next, observe that for j=1,...,n, if $j \in I^* \cap N$, then $j \in I \cap N$ with certainty, while if $j \in \mathrm{N} \sim I^*$, then $j \in I \cap N$ with probability at most c since each PSU in S was in a different initial stratum. Consequently,

$$\Omega_R = \sum_{j=1}^{n} P(j \in I \cap N) \le c \sum_{j=1}^{n} P(j \in (N \sim I^*) \cap I) + \sum_{j=1}^{n} P(j \in I^* \cap N) \qquad \text{(E.8)}$$

Now

$$\sum_{j=1}^{n} P(j \in (N \sim I^*) \cap I) \le \sum_{j=1}^{n} P(j \in N) = 2. \qquad \text{(E.9)}$$

Therefore, if we can prove that

$$\lim_{n\to\infty} \sum_{j=1}^{n} P(j \in I^* \cap N) = 0, \qquad \text{(E.10)}$$

then it will follow from (E.8) - (E.10) that $\lim_{n\to\infty} \Omega_R \le 2c$. Since we also have $2c = \Omega_I \le \Omega_R$, (4.3) will then follow.

To establish (E.10), observe that for the specified ordering, $j \in I^*$ if and only if $j \in I$ and at most one of 1,..,j-1 was in I. Consequently,

$$P(j \in I^*) = c(1-c)^{j-1} + c^2(j-1)(1-c)^{j-2} \le j(1-c)^{j-2}.$$

Since we also have that $P(j \in \mathrm{N}) = \pi_j = 2/n,$ it follows that

$$\sum_{j=1}^{n} P(j \in I^* \cap N) \leq \sum_{j=1}^{n} \min\{j(1-c)^{j-2}, \quad 2/n\},$$

and thus we need only show that

$$\lim_{n\to\infty} \sum_{j=1}^{n} \min\{j(1-c)^{j-2}, \quad 2/n\} = 0. \tag{E.11}$$

Let $\tau(n)$ denote the largest integer j in $\{1,...,n\}$ for which $j(1-c)^{j-2} \geq 2/n.$ If no such j exists, let $\tau(n) = 0.$ Then

$$\lim_{n\to\infty} \sum_{j=1}^{n} \min\{j(1-c)^{j-2}, \ 2/n\} \leq 2 \lim_{n\to\infty} \frac{\tau(n)}{n} + \lim_{n\to\infty} \sum_{j=\tau(n)+1}^{n} j(1-c)^{j-2}, \tag{E.12}$$

and hence it suffices to show that the two limits on the right hand side of (E.12) are 0.

To compute the first limit, note that if $\tau(n) > 0$ then

$$n(1-c)^{\tau(n)-2} \geq \tau(n)(1-c)^{\tau(n)-2} \geq 2/n, \tag{E.13}$$

and hence

$$\tau(n) \leq \frac{\log 2 - 2\log n}{\log(1-c)} + 2.$$

Consequently, by L'Hospital's rule,

$$\lim_{n\to\infty}\frac{\tau(n)}{n}\le\lim_{n\to\infty}\left[\frac{\log 2-2\log n}{n\log(1-c)}+\frac{2}{n}\right]=0.$$

To demonstrate that the second limit in (E.12) is 0, observe that since

$$\lim_{n\to\infty}\sum_{j=\tau(n)+1}^{n} j(1-c)^{j-2}\le\lim_{n\to\infty}\sum_{j=\tau(n)+1}^{\infty} j(1-c)^{j-2},$$

it suffices to prove that $\sum_{j=1}^{\infty} j(1-c)^{j-2}$ is a convergent series and $\lim_{n\to\infty}\tau(n)=\infty$. The series is convergent since it is an arithmetic-geometric series which is known to converge because $|1-c|<1$. To prove that $\lim_{n\to\infty}\tau(n)=\infty$, note that either $\tau(n)=n$ or

$(1-c)^{\tau(n)}<(\tau(n)+1)(1-c)^{\tau(n)-1}<2/n$. Conseqently, $\lim_{n\to\infty}\tau(n)\ge\lim_{n\to\infty}\min\left\{n,\ \frac{\log 2-\log n}{\log(1-c)}\right\}=\infty.$

Thus, (4.3) is proven.

ACKNOWLEDGEMENT

The programming assistance of Todd Williams is gratefully acknowledged.

REFERENCES

Aragon, J., and Pathak, P.K. (1990), "An Algorithm for Optimal Integration of Two Surveys," *Sankyā*: *The Indian Journal of Statistics*, 52, Series B, Pt. 2, 198-203.

Arthanari, T.S., and Dodge, Y. (1981), *Mathematical Programming in Statistics*, New York: John Wiley and Sons.

Causey, B.D., Cox, L.H., and Ernst, L.R. (1985), "Applications of Transportation Theory to Statistical Problems," *Journal of the American Statistical Association*, 80, 903-909

Ernst, L.R. (1986), "Maximizing the Overlap Between Surveys When Information Is Incomplete," *European Journal of Operational Research*, 27, 192-200.

_____ (1989), "Further Applications of Linear Programming to Sampling Problems," Bureau of the Census, Statistical Research Division, Research Report Series, No. RR-89/05.

Ernst, L.R., and Ikeda, M. (1992a), "Modification of the Reduced-Size Transportation Problem for Maximizing Overlap When Primary Sampling Units Are Redefined in the New Design," Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-91/01.

_____ (1992b), "Summary of the Performance of the Maximum Overlap Algorithms for the 1990's Redesign of the Demographic Surveys," Bureau of the Census, Statistical Research Division, Technical Note Series, No. TN-92/01.

Glover, F., Karney, D., Klingman, D., and Napier, A. (1974), "A Computation Study on Start Procedures, Basic Change Criteria and Solution Algorithms for Transportation Problems," *Management Sciences*, 20, 793-813.

Keyfitz, N. (1951), "Sampling With Probabilities Proportional to Size: Adjustment for Changes in Probabilities," *Journal of the American Statistical Association*, 46, 105-109.

Kish, L., and Scott, A. (1971), "Retaining Units After Changing Strata and Probabilities," *Journal of the American Statistical Association*, 66, 461-470.

Perkins, W.M. (1970), "1970 CPS Redesign: Proposed Method for Deriving Sample PSU Selection Probabilities Within 1970 NSR Strata," memorandum to Joseph Waksberg, Bureau of the Census.

Raj, D. (1968), *Sampling Theory*, New York: McGraw Hill.