MAXIMIZING AND MINIMIZING OVERLAP WHEN SELECTING A LARGE NUMBER OF UNITS PER STRATUM SIMULTANEOUSLY FOR TWO DESIGNS

Lawrence R. Ernst, Bureau of Labor Statistics
Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Room 3160, Washington DC 20212

Key Words: Controlled selection, Stratified design, Compensation surveys

1. Introduction

Consider the following sampling problem. Sample units are to be selected for two designs, denoted as D_1 and D_2, with identical universes and stratifications, with S denoting one of the strata. The selection probabilities for each unit in S are generally different for the two designs, as are the number of units to be selected from S for each of the designs. The sample units are to be selected simultaneously for the two designs. We wish to maximize the overlap of the sample units, that is to select the sample units so that:

There are a predetermined number of units, n_j, selected from S for the D_j sample, j=1,2. (1.1)

The i-th unit in S is selected for the D_j sample with its assigned probability, denoted π_{ij}. (1.2)

The expected value for the number of sample units common to the two designs is maximized. (1.3)

In this paper we demonstrate how the two-dimensional controlled selection procedure of Causey, Cox and Ernst (1985) can be used to satisfy these conditions and the additional condition that:

The number of sample units in common to any D_1 and D_2 samples is always within one of the maximum expected value. (1.4)

Overlap maximization has generally been used as a technique to reduce data collection costs, such as the costs associated with the hiring of new interviewers when the units being overlapped are primary sampling units (PSUs), that is geographic areas, or the additional costs of an initial interview when the units being overlapped are ultimate sampling units (USUs). Most of the previous work on maximizing the overlap of sample units considered the case when the two sets of sample units are PSUs that must be chosen sequentially, as is the case when the second design is a redesign of the first design. The number of sample PSUs chosen from each stratum is generally small. This problem was first studied by Keyfitz (1951), who presented an overlap procedure for one unit per stratum designs in the special case when the initial and new strata are identical, with only the selection probabilities changing. Keyfitz's procedure is optimal in the sense of actually producing the maximal expected overlap. (Although we refer to all the overlap procedures as procedures for maximizing the overlap, many of these procedures do not actually produce the maximal expected overlap, but instead merely increase the expected overlap to varying degrees in comparison with independent selection of the two samples.) For the more general one unit per stratum problem, Perkins (1970), and Kish and Scott (1971) presented procedures that are not optimal. Causey, Cox and Ernst (1985), Ernst (1986), and Ernst and Ikeda (1995) presented linear programming procedures for overlap maximization under very general conditions. The Causey, Cox and Ernst procedure always yields an optimal overlap, while the other two linear programming procedures generally produce a high, although not necessarily optimal, overlap. These linear programming procedures impose no theoretical restrictions on changes in strata definitions or number of units per stratum, but the size of the linear programming problem increases so rapidly as the number of sample PSUs per stratum increases that these procedures are generally operationally feasible to implement only when the number of sample PSUs per stratum is very small. This operational problem is most severe for the Causey, Cox and Ernst procedure, which is one reason that the other two linear programming procedures have been used even though they do not guarantee an optimal overlap.

Overlap procedures have also been used for sequential selection at the ultimate sampling unit (USU) level, where the number of the sample units per stratum can in some case be fairly large and for which, consequently, none of the above procedures are usable. Brewer, Early and Joyce (1972), Brick, Morganstein and Wolter (1987), Gilliland (1984), and Ernst (1995b) present overlap procedures that are usable under these conditions. These first two of these procedures are optimal but do not guarantee a fixed sample size, while the opposite is true for the other two procedures.

In certain overlap applications it is possible to choose the samples for the two designs simultaneously. For example, the Bureau of Labor Statistics recently planned to select new sample establishments from industry × size class strata for the governments samples for two compensation surveys, the Economic Cost Index (ECI) and the Occupational Compensation

Surveys Program (OCSP). To reduce interviewing expenses we wanted the two surveys to have as many sample establishments in common as possible. Since ECI has a much smaller sample than OCSP we actually wanted an ultimate form of overlap, that is for the ECI governments sample to be a subsample of the OCSP governments sample. In fact, a special case of (1.1)-(1.4), which generally applies in this application, occurs when $\pi_{i2} \le \pi_{i1}$ for all units in S, in which case, as we will show, (1.3), (1.4) can be replaced with the more stringent requirement that:

Each D_2 sample unit in S is a D_1 sample unit. (1.5)

Previously, Ernst (1996) presented an optimal solution to the overlap problem in the context of simultaneous selection under different conditions than considered here. That solution is limited to one unit per stratum designs, in contrast to the procedure in this paper which has no restriction on the number of sample units in a stratum. On the other hand, the procedure in Ernst (1996) applies when the two designs have different stratifications, while the procedure in the current paper requires that the stratifications be identical to insure that the optimal overlap is attained. The procedure of Ernst (1996) also uses the controlled selection algorithm of Causey, Cox and Ernst (1985), although in a different way than in the current paper. Pruhs (1989) had earlier developed a solution to the overlap problem considered in Ernst (1996) using a much more complex graph theory approach.

In Section 2 we describe the basic idea of the current procedure and list a set of requirements that are sufficient to satisfy (1.1-1.4). In Section 3 the controlled selection procedure of Causey, Cox, and Ernst (1985) is presented and a solution to our overlap problem is obtained which satisfies the set of requirements listed in Section 2.

In Section 4 it is shown how the procedure of Sections 2 and 3 can be easily modified to solve the problem of minimizing the expected overlap of sample units under the same assumptions. Overlap minimization has typically been used to reduce respondent burden. Most, but not all, of the overlap maximization procedures previously mentioned can also be used to minimize overlap. In addition, Perry, Burt and Iwig (1993) presented a different approach than presented here to the minimization of overlap when the samples are selected simultaneously. Their approach has the advantage of not being restricted to two designs. However, their method is not optimal and assumes equal probability of selection within a stratum.

Due to space limitations, an illustrative example is omitted here as have the final two sections of the paper, one which describes how our procedure can be modified, although with loss of optimality, for use when the strata definitions are not identical in the two designs, and the other which presents the results of the application of our procedure to the selection of the ECI and OCSP government samples. The complete paper is available from the author.

2. Outline of Overlap Procedure and List of Set of Conditions to Be Satisfied

The basic idea of controlled selection is as follows. First, a two-dimensional, real valued, tabular array, $\mathbf{S} = (s_{ij})$, is constructed which specifies the probability and expected value conditions that must be satisfied for the particular problem. (A tabular array is one in which the final row and final column are marginal values, that is each entry in a particular column in the last row is the sum of the other entries in that column and each entry in a particular row of the last column is the sum of the other entries in that row.) The array **S** is known as the controlled selection problem. Next, a sequence of integer valued, tabular arrays, $\mathbf{M}_1 = (m_{ij1})$, $\mathbf{M}_2 = (m_{ij2}), \ldots, \mathbf{M}_l = (m_{ijl})$, with the same number of rows and columns as **S**, and associated probabilities, $p_1,\ldots,p_l$, are constructed which satisfy certain conditions. This set of integer valued arrays and probabilities constitute a solution to the controlled selection problem **S**. Finally, a random array, $\mathbf{M} = (m_{ij})$, is then chosen from among these l arrays using the indicated probabilities. The selected array determines the sample allocation. The set of integer valued arrays and their associated probabilities guarantee the expected value conditions specified by **S** are satisfied.

We proceed to describe **S** and $\mathbf{M}_1,\ldots,\mathbf{M}_l$ for the procedure of this paper in greater detail. In our application of controlled selection, each stratum corresponds to a separate controlled selection problem and **S** is a $(N+1)\times 5$ array, where N is the number of units in the stratum universe. Thus, there are N internal rows and 4 internal columns in **S.** Each internal row of the selected array corresponds to a unit in the stratum universe. In the i-th internal row, the first element is the probability that the i-th unit is in the D_1 sample only; the second element is the probability that it is in the D_2 sample only; the third element is the probability that it is in both samples; and the fourth element is the probability that it is in neither sample. The marginals in the final column of the N internal rows are all 1 since each unit must fall in exactly one of the four categories. The marginals in the first 4 columns of the final row are the expected number of units in the corresponding category, and the grand total is N.

We next explain how the values for the internal elements of **S** are computed. The key value is s_{i3}, the probability that the i-th unit is in both samples. Let

$$s_{i3} = \min\{\pi_{i1,} \pi_{i2}\}, \quad (2.1)$$

$$s_{ij} = \pi_{ij} - s_{i3}, \ j = 1,2, \quad (2.2)$$

$$s_{i4} = 1 - \sum_{j=1}^{3} s_{ij} \quad (2.3)$$

Now (2.1) is motivated by (1.2) and (1.3). That is, if (1.2) held then the probability that the i-th unit is in both samples clearly could not exceed either π_{i1} or π_{i2}, and therefore (1.3) would be satisfied if the probability that unit i is in both samples equals s_{i3} for each i. Also (2.2) is required by (1.2), that is the probability that the i-th unit is in the D_j sample only is simply the probability that it is in the D_j sample minus the probability that is in both samples. Finally, (2.3) is required by the fact that for each sample, every unit must be in exactly one of the four categories determined by the four internal columns.

Note that by (2.1), (2.2), all the entries in the second column of **S** are 0 in the special case when $\pi_{i2} \le \pi_{i1}$ for all units in S, and hence each D_2 sample unit in S will be a D_1 sample unit, as required by (1.5), provided the sampling procedure preserves all the probability and expected value conditions specified in **S**.

We next describe the conditions that must be satisfied by the sequence of integer valued arrays, $\mathbf{M}_1, \ldots, \mathbf{M}_l$, and associated probabilities, $p_1, \ldots, p_l$, which determine the sample allocation. In each internal row of each of these arrays, one of the four internal columns has the value 1 and the other three have the value 0. A 1 in the first column indicates that the unit is only in the D_1 sample; a 1 in the second column indicates that the unit is only in the D_2 sample; a 1 in the third column indicates that the unit is in both samples; and a 1 in the fourth column indicates that the unit is in neither sample. The probability mechanism for selecting the integer valued array guarantees, as will be shown in the next section, that for each unit a 1 appears in each column with the correct probability, that is the probability determined by **S**.

We next list a set of requirements which, if met by the random array **M**, are sufficient to satisfy (1.1)-(1.4). Note that (1.2) will be satisfied if

$$P(m_{ij} = 1) + P(m_{i3} = 1) = s_{ij} + s_{i3} = \pi_{ij},$$
$$i = 1, \ldots, N, \ j = 1,2. \quad (2.4)$$

In addition, (1.3) will be satisfied if we also have

$$P(m_{i3} = 1) = s_{i3}, \ i = 1, \ldots, N. \quad (2.5)$$

Consequently, if we can establish that

$$E(m_{ij}) = \sum_{k=1}^{l} p_k m_{ijk} = s_{ij},$$
$$i = 1, \ldots, N+1, \ j = 1, \ldots, 5, \quad (2.6)$$

then (1.2) and (1.3) hold, since (2.6) implies (2.4), (2.5).

To additionally establish (1.1) we need only show that

$$m_{(N+1)jk} + m_{(N+1)3k} = n_j, \ j = 1,2, \ k = 1, \ldots, l. \quad (2.7)$$

Finally, to establish (1.4) it suffices to show that

$$\left| m_{ijk} - s_{ij} \right| < 1, \ i = 1, \ldots, N+1,$$
$$j = 1, \ldots, 5, \ k = 1, \ldots, l, \quad (2.8)$$

since then, in particular,

$$\left| m_{(N+1)3k} - s_{(N+1)3} \right| < 1, \ k = 1, \ldots, l,$$

where $s_{(N+1)3}$ is the maximum expected number of units in common to the two samples and $m_{(N+1)3k}$ is the number of units in common to the k-th possible sample.

Also observe that in the special case when $\pi_{i2} \le \pi_{i1}$ for all units in S, then $s_{i2} = 0, \ j = 1, \ldots, N$. Consequently, by (2.6). (2.8), we would have $m_{i2k} = 0$, $i = 1, \ldots, N$, $k = 1, \ldots, l$, and hence (1.5) would follow.

We demonstrate in the next section how the controlled selection procedure of Causey, Cox and Ernst can be used to establish (2.6)-(2.8) in general, which will complete the development of the overlap procedure.

3. Completion of the Overlap Algorithm

The concept of controlled selection was first developed by Goodman and Kish (1950), but they did not present a general algorithm for solving such problems. In Causey, Cox and Ernst (1985), an algorithm for obtaining a solution to the controlled selection problem was obtained. We demonstrate here how their solution can be used to complete the algorithm of this paper, that is to construct a finite set of $(N+1)\times 5$ nonnegative, integer valued, tabular arrays,

$\mathbf{M}_1,\ldots,\mathbf{M}_l$, and associated probabilities, $p_1,\ldots,p_l$, satisfying (2.6)-(2.8).

The discussion of controlled selection will be limited to the two-dimensional problem. Although the concept can be generalized to higher dimensions, Causey, Cox and Ernst (1985) proved that solutions to controlled selection problems do not always exist for dimensions greater than two.

The controlled selection procedure of Causey, Cox and Ernst is built upon the theory of controlled rounding developed by Cox and Ernst (1982). In general, a controlled rounding of an $(N+1)\times(M+1)$ tabular array $\mathbf{S}=(s_{ij})$ to a positive integer base b is an $(N+1)\times(M+1)$ tabular array $\mathbf{M}=(m_{ij})$ for which $m_{ij}=\lfloor s_{ij}/b\rfloor b$ or $(\lfloor s_{ij}/b\rfloor+1)b$ for all i,j, where $\lfloor x\rfloor$ denotes the greatest integer not exceeding x. A zero-restricted controlled rounding to a base b is a controlled rounding that satisfies the additional condition that $m_{ij}=s_{ij}$ whenever s_{ij} is an integral multiple of b. If no base is specified, then base 1 is understood.

By modeling the controlled rounding problem as a transportation problem, Cox and Ernst (1982) obtained a constructive proof that a zero-restricted controlled rounding exists for every two-dimensional array. Thus, while conventional rounding of a tabular array commonly results in an array that is no longer additive, this result shows that is possible to always preserve additivity if the original values are allowed to be rounded either up or down.

With $\mathbf{S}$ as above, a solution to the controlled selection problem for this array is a finite sequence of $(N+1)\times(M+1)$ tabular arrays, $\mathbf{M}_1=(m_{ij1})$, $\mathbf{M}_2=(m_{ij2}),\ldots,\mathbf{M}_l=(m_{ijl})$, and associated probabilities, $p_1,\ldots,p_l$, satisfying:

$\mathbf{M}_k$ is a zero-restricted controlled rounding of $\mathbf{S}$ for all $k=1,\ldots,l$, (3.1)

$$\sum_{k=1}^{l} p_k = 1, \tag{3.2}$$

$$\sum_{k=1}^{l} m_{ijk}p_k = s_{ij}, \quad i=1,\ldots,N+1, \quad j=1,\ldots,M+1. \tag{3.3}$$

If $\mathbf{S}$ arises from a sampling problem for which s_{ij} is the expected number of sample units selected in cell (i,j), and the actual number selected in each cell is determined by choosing one of the $\mathbf{M}_k$'s with its associated probability, then by (3.1) the deviation of s_{ij} from the number of sample units actually selected from cell (i,j) is less than 1 in absolute value, whether (i,j) is an internal cell or a total cell. By (3.2), (3.3) the expected number of sample units selected is s_{ij}. Consequently, with $\mathbf{S}$ as defined in Section 2, a solution to the controlled selection problem satisfies (2.6), (2.8).

Although, as noted, any solution to a controlled selection problem satisfies (2.6), (2.8), it requires a great deal more work to establish (2.7), including an understanding of how solutions to controlled selection problems are obtained using the Causey Cox and Ernst (1985) algorithm, which we proceed to present.

Causey Cox and Ernst obtained a solution to the controlled problem $\mathbf{S}$ by means of the following recursive computation of the sequences $\mathbf{M}_1,\ldots,\mathbf{M}_l$ and $p_1,\ldots,p_l$, along with a recursive computation of a sequence of real valued $(N+1)\times(M+1)$ tabular arrays $\mathbf{A}_k=(a_{ijk})$, $k=1,\ldots,l$. Let $\mathbf{A}_1=\mathbf{S}$, while for $k\geq 1$ we define $\mathbf{M}_k, p_k, \mathbf{A}_{k+1}$ in terms of $\mathbf{A}_k$ as follows. $\mathbf{M}_k$ is any zero-restricted controlled rounding of $\mathbf{A}_k$. To define p_k, first let

$$d_k=\max\{|m_{ijk}-a_{ijk}|:\ i=1,\ldots,N+1,\ j=1,\ldots,M+1\}, \tag{3.4}$$

and then let

$$\begin{aligned} p_k &= (1-d_k) && \text{if } k=1, \\ &= (1-\sum_{i=1}^{k-1} p_i)(1-d_k) && \text{if } k>1. \end{aligned} \tag{3.5}$$

If $d_k>0$ then define $\mathbf{A}_{k+1}$ by letting for all i,j,

$$a_{ij(k+1)}=m_{ijk}+(a_{ijk}-m_{ijk})/d_k. \tag{3.6}$$

It is established in Causey, Cox and Ernst (1985) that eventually there is an integer l for which $d_l=0$ and that this terminates the algorithm; that is, $\mathbf{M}_1,\ldots,\mathbf{M}_l$ and $p_1,\ldots,p_l$ constitute a solution to the controlled selection problem satisfying (3.1)-(3.3). It is also established in their paper that for all i,j,k,

$$\lfloor s_{ij}\rfloor \leq a_{ijk} \leq \lfloor s_{ij}\rfloor+1, \text{ and } a_{ijk}=s_{ij} \text{ if } s_{ij} \text{ is an integer.} \tag{3.7}$$

Now to obtain (2.7), first note that for the array $\mathbf{S}$ defined by (2.1)-(2.3) we have by (2.2) that

$$s_{(N+1)j}+s_{(N+1)3}=n_j,\ j=1,2. \tag{3.8}$$

Unfortunately, (3.8) is not sufficient to guarantee that all solutions to the controlled selection problem $\mathbf{S}$

obtained by the algorithm just described satisfy (2.7). A particular solution to the controlled selection problem that does satisfy (2.7) can be obtained, however, using the following approach. We first demonstrate that it is sufficient to show that if

$$a_{(N+1)jk} + a_{(N+1)3k} = n_j, \; j = 1,2, \tag{3.9}$$

for a particular k, then there exists a zero-restricted controlled rounding $\mathbf{M}_k$ of $\mathbf{A}_k$ for which

$$m_{(N+1)jk} + m_{(N+1)3k} = n_j, \; j = 1,2. \tag{3.10}$$

This is sufficient because (3.9) holds for $k = 1$ by (3.8), while if (3.9) holds for any positive integer k and $\mathbf{M}_k$ satisfies (3.10) for that value of k, then (3.9) holds for $k+1$ by (3.6); consequently by recursion we could obtain a zero-restricted controlled rounding $\mathbf{M}_k$ of $\mathbf{A}_k$ satisfying (3.10) for each k, and thus (2.7) would hold for this set of arrays.

To establish that (3.9) implies (3.10), we observe that by (3.9) and the fact that

$$a_{(N+1)5k} = s_{(N+1)5} = N, \text{ which is an integer;} \tag{3.11}$$

it follows that the fractional parts of $a_{(N+1)jk}$, $j = 1,2$, are the same, as are the fractional parts of $a_{(N+1)jk}$, $j = 3,4$. Furthermore, one of two possible sets of additional conditions must hold. The first possibility is that $a_{(N+1)jk}$ is an integer for all $j = 1,2,3,4$. In this case (3.10) holds for any zero-restricted controlled rounding of $\mathbf{A}_k$.

In the second case, which is assumed throughout the remainder of this section, none of $a_{(N+1)jk}$, $j = 1,2,3,4$, are integers, but the fractional part of $a_{(N+1)jk}$, $j = 1,2$ plus the fractional part of $a_{(N+1)jk}$, $j = 3,4$ is 1. In this case $m_{(N+1)jk} = \lfloor a_{(N+1)jk} \rfloor + 1$ for exactly two j's among $j = 1,2,3,4$ for every zero-restricted controlled rounding $\mathbf{M}_k$ of $\mathbf{A}_k$, since

$$N = m_{(N+1)5k} = \sum_{j=1}^{4} m_{(N+1)jk} = \sum_{j=1}^{4} a_{(N+1)jk};$$

and that for $\mathbf{M}_k$ to satisfy (3.10) it is sufficient that additionally either

$$m_{(N+1)jk} = \lfloor a_{(N+1)jk} \rfloor, \; j = 1,2, \tag{3.12}$$

or

$$m_{(N+1)jk} = \lfloor a_{(N+1)jk} \rfloor + 1, \; j = 1,2. \tag{3.13}$$

To show that we can obtain a zero-restricted controlled rounding $\mathbf{M}_k$ of $\mathbf{A}_k$ satisfying (3.12) or (3.13) we proceed as follows. It is established in Cox and Ernst (1982) that a linear programming problem which minimizes an objective function of the form

$$\sum_{i=i}^{N+1} \sum_{j=1}^{5} c_{ij} x_{ij}, \tag{3.14}$$

where the x_{ij}'s are variables and the c_{ij}'s are constants, subject to the constraints

$$\sum_{i=1}^{N} x_{ij} = x_{(N+1)j}, \; j = 1,\ldots,5, \tag{3.15}$$

$$\sum_{j=1}^{4} x_{ij} = x_{i5}, \; i = 1,\ldots,N+1, \tag{3.16}$$

$$\lfloor a_{ijk} \rfloor \le x_{ij} \le \lfloor a_{ijk} \rfloor + 1, \; i = 1,\ldots,N+1, \; j = 1,\ldots,5, \tag{3.17}$$

$$x_{ij} = a_{ijk} \text{ if } a_{ijk} \text{ is an integer, } i = 1,\ldots,N+1, \; j = 1,\ldots,5, \tag{3.18}$$

can be transformed into a transportation problem for which there is an integer valued solution $\mathbf{M}_k$, that is $\mathbf{M}_k$ is a zero-restricted controlled rounding of $\mathbf{A}_k$. In particular, since $\mathbf{A}_k$ also satisfies (3.15)-(3.18) we have

$$\sum_{i=1}^{N+1} \sum_{j=1}^{5} c_{ij} m_{ijk} \le \sum_{i=1}^{N+1} \sum_{j=1}^{5} c_{ij} a_{ijk}. \tag{3.19}$$

We will show that with the appropriate choice of objective function (3.14), a zero-restricted controlled rounding $\mathbf{M}_k$ of $\mathbf{A}_k$ which is a solution to the linear programming problem (3.14)-(3.18) will satisfy (3.12) or (3.13) and hence a solution to the controlled selection problem $\mathbf{S}$ that satisfies (2.7) can be obtained.

There are three cases to consider. First if

$$\sum_{j=1}^{2} a_{(N+1)jk} < \sum_{j=1}^{2} \lfloor a_{(N+1)jk} \rfloor + 1, \tag{3.20}$$

then by (3.19) a controlled rounding obtained by minimizing $\sum_{j=1}^{2} x_{(N+1)j}$ subject to (3.15-3.18) will satisfy (3.12). Similarly, if the inequality in (3.20) is reversed, a controlled rounding satisfying (3.13) can be obtained by minimizing $-\sum_{j=1}^{2} x_{(N+1)j}$, which is equivalent to maximizing $\sum_{j=1}^{2} x_{(N+1)j}$. Finally, if the inequality in (3.20) is an equality instead then, since $a_{(N+1)1k}$ is not an integer, we have by (2.2), (3.7) that $0 < a_{i*1k} < 1$ for some $i*$ with $1 \le i* \le N$. In addition, we have that $0 < a_{i*j*k} < 1$ for some $j* \in \{2,3,4\}$, since $a_{i*5k} = 1$ by (3.7). Furthermore, $j* \ne 2$ since $a_{i*2k} = 0$ by (2.2), (3.7). Then consider the $(N+1)\times 5$ tabular array $\mathbf{A}'_k = (a'_{ijk})$ with internal elements $a'_{i*1k} = a_{i1k} - \varepsilon$, $a'_{i*j*k} = a_{i*j*k} + \varepsilon$, $a'_{ijk} = a_{ijk}$ for all other i,j, where $\varepsilon > 0$ is sufficiently small that the tabular arrays $\mathbf{A}'_k$ and $\mathbf{A}_k$ have the same set of zero-restricted controlled roundings. Since $\sum_{j=1}^{2} a'_{(N+1)jk} < \sum_{j=1}^{2} \lfloor a'_{(N+1)jk} \rfloor + 1$, a zero-restricted controlled rounding of $\mathbf{A}'_k$ and hence of $\mathbf{A}_k$ can be obtained which satisfies (3.12).

4. Minimization of Overlap

Sometimes it is considered desirable to minimize the expected number of sample units in S common to two designs rather than maximize it. The procedure described in Sections 2 and 3 can very easily be modified to minimize overlap. Simply redefine $s_{i3} = \max\{\pi_{i1} + \pi_{i2} - 1, 0\}$. The remainder of the procedure is identical to the maximization procedure.

The rationale for the definition of s_{i3} in the minimization case is analogous to the rationale for the definition of s_{i3} in the maximization case presented in Section 2. For while $\min\{\pi_{i1}, \pi_{i2}\}$ is the maximum possible value for the probability that the i-th unit is in sample for both designs, the minimum possible value for this probability is $\max\{\pi_{i1} + \pi_{i2} - 1, 0\}$.

References

Brewer, K. R. W., Early, L. J., and Joyce, S. F. (1972). Selecting Several Samples from a Single Population. Australian Journal of Statistics, 14, 231-239.

Brick, J. M., Morganstein, D. R., and Wolter, C. L. (1987). Additional Uses for Keyfitz Selection. Proceedings of the Section on Survey Research Methods, American Statistical Association, 787-791.

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.

Cox, L. H. and Ernst, L. R. (1982). Controlled Rounding. INFOR, 20, 423-432.

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

Ernst, L. R. (1995a). Allocation of New ECI Governments Sample. Memorandum to The Record, Bureau of Labor Statistics, dated November 20.

Ernst, L. R. (1995b). Maximizing and Minimizing Overlap of Ultimate Sampling Units. Proceedings of the Section on Survey Research Methods, American Statistical Association, 706-711.

Ernst. L. R. (1996). Maximizing the Overlap of Sample Units for Two Designs with Simultaneous Selection. Journal of Official Statistics, 12, 33-45.

Ernst, L. R. and Ikeda, M. (1995). A Reduced-Size Transportation Algorithm for Maximizing the Overlap Between Surveys. Survey Methodology, 21, 147-157.

Ernst, L. R. and Ponikowski, C. H. (1996). Controlled Selection of OCSP and ECI Governments Samples. Memorandum to The Record, Bureau of Labor Statistics, dated January 8.

Gilliland, P. D. (1984). 1985 PATC Measures of Size Specifications. Memorandum to W. Brown, Bureau of Labor Statistics, dated September 7.

Goodman, R. and Kish, L. (1950). Controlled Selection —A Technique in Probability Sampling. Journal of the American Statistical Association, 45, 350-372.

Keyfitz, N. (1951). Sampling With Probabilities Proportionate 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, U. S. Bureau of the Census, dated August 5.

Perry, C. R., Burt, J. C., and Iwig, W. C. (1993). Methods of Selecting Samples in Multiple Surveys to Reduce Respondent Burden. Proceedings of the International Conference on Establishment Surveys, American Statistical Association, 345-351.

Pruhs, K. (1989). The Computational Complexity of Some Survey Overlap Problems. Proceedings of the Section on Survey Research Methods, American Statistical Association, 747-752.

Any opinions expressed in this paper are those of the author and do not constitute policy of the Bureau of Labor Statistics.