Journal of Official Statistics, Vol. 14, No. 3, 1998, pp. 297–314

Maximizing and Minimizing Overlap When Selecting a Large Number of Units per Stratum Simultaneously for Two Designs

Lawrence R. Ernst[1]

A number of procedures have been developed, beginning with the work of Keyfitz, for maximizing or minimizing the overlap of sampling units for two stratified designs. Most of these procedures are not applicable at all, or are not feasible to implement, unless the number of units selected per stratum is very small. The previous procedures that the author is aware of for increasing or decreasing overlap when a large number of units per stratum are selected either do not generally yield an optimal overlap or do not guarantee fixed sample sizes. Furthermore, these overlap procedures have typically been developed for use when the two designs must be selected sequentially, as is the case when the second design is a redesign of the first design. In the current article a very different, large sample per stratum procedure is presented for maximizing or minimizing overlap when the units can be selected for the two designs simultaneously, as may be the case for two different surveys. The procedure guarantees fixed sample sizes and also an optimal overlap if the two designs have identical stratifications, but can still be used, with loss of optimality, if the stratifications differ. An application of this procedure to the joint selection of samples for two U.S. Bureau of Labor Statistics compensation surveys is discussed.

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 is 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, $j = 1, 2$. That is, the sample size for each stratum and design combination is fixed. (1.1)

[1] U.S. Bureau of Labor Statistics, Office of Compensation and Working Conditions, Statistical Methods Group, Washington, D.C. 20212, U.S.A.
The views expressed in this article are attributable to the author and do not necessarily reflect those of the U.S. Bureau of Labor Statistics.
Acknowledgements: The programming support of Patrick Wells, who worked in part on this project under the leadership of Chester Ponikowski, is gratefully acknowledged. The author would also like to thank the reviewer and the Associate Editor for their constructive comments.

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

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

In this article 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)

The algorithm to be described imposes no theoretical limits on the number of units, n_j, selected from S for the D_j sample. Operational limits are discussed in Section 5.

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 (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 cases be fairly large and for which, consequently, none of the above procedures are usable. Brewer, Early, and Joyce (1972), Brick, Morganstein, and Wolters (1987),

Gilliland (1984), and Ernst (1995b) present overlap procedures that are usable under these conditions. The 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 U.S. 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 $\pi_{i2} \leq \pi_{i1}$ for all units in S, in which case, as we will show, (1.3) and (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)

(Note that, as explained in Section 7, the ECI selection probabilities were not proportional to the OCSP selection probabilities. If they had been, it would not have been necessary to use an overlap procedure. We could simply have selected the OCSP sample first and then subsampled the OCSP sample units with equal probability to obtain the ECI sample.)

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 article 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 article 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 article. 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, with the aid of an example, 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.

In Section 5 we consider three separate issues. First we compare the current procedure

with the procedure in Ernst (1996), noting the similarities and differences. Next, with regard to the ''large number of units'' referred to in the title of the article, we explain why there are operational upper limits on the size of S even though there are no theoretical limits on the number of units that can be selected using the procedure. Finally, we discuss the issue of joint inclusion probabilities for pairs of units in the D_j sample.

In Section 6 we describe how our procedure can be modified, although with loss of optimality, for use when the strata definitions are not identical in the two designs.

Finally, in Section 7 we present the results of the application of our procedure to the selection of the ECI and OCSP governments samples. Although the controlled selection procedure was carried far enough to report results, it was not actually used in production. This is because the ECI and OCSP are currently in the process of being integrated into a single compensation program, the National Compensation Survey (NCS). At the time the decision was made to use controlled selection, it was anticipated that complete sample integration of these surveys might still be several years away. However, the integration subsequently was moved forward dramatically in time. Under the NCS design, all sample units, including government units, to be used in ECI estimates will be selected as a subsample of the parent NCS sample, obviating the need for controlled selection.

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

The procedure to be described is applied separately to each stratum S. As a result, the sampling for each design is independent from stratum to stratum. As explained in Ernst (1986), this independence typically does not hold when an overlap procedure is applied to designs that do not have identical stratifications.

As we proceed to explain controlled selection and its application to the overlap procedure of this article, we will illustrate certain aspects of the procedure by use of the same example, much of which is presented in Table 1 and Figures 1 and 2. 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. Each of the arrays in Figures 1 and 2 is a tabular array.) The array $\mathbf{S}$ 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 $\mathbf{S}$, 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

Table 1. Selection probabilities for example

	i				
	1	2	3	4	5
π_{i1}	.6	.4	.8	.6	.6
π_{i2}	.8	.4	.2	.4	.2

$$\mathbf{S} = \begin{array}{cccc|c} 0 & .2 & .6 & .2 & 1 \\ 0 & 0 & .4 & .6 & 1 \\ .6 & 0 & .2 & .2 & 1 \\ .2 & 0 & .4 & .4 & 1 \\ .4 & 0 & .2 & .4 & 1 \\ \hline 1.2 & .2 & 1.8 & 1.8 & 5 \end{array}$$

Fig. 1. 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 conditions specified by **S** are satisfied.

$$\mathbf{A}_1 = \mathbf{S} = \begin{array}{cccc|c} 0 & .2 & .6 & .2 & 1 \\ 0 & 0 & .4 & .6 & 1 \\ .6 & 0 & .2 & .2 & 1 \\ .2 & 0 & .4 & .4 & 1 \\ .4 & 0 & .2 & .4 & 1 \\ \hline 1.2 & .2 & 1.8 & 1.8 & 5 \end{array} \qquad \mathbf{M}_1 = \begin{array}{cccc|c} 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ \hline 1 & 0 & 2 & 2 & 5 \end{array} \qquad \mathbf{A}_2 = \begin{array}{cccc|c} 0 & .33 & .33 & .33 & 1 \\ 0 & 0 & .67 & .33 & 1 \\ .33 & 0 & .33 & .33 & 1 \\ .33 & 0 & 0 & .67 & 1 \\ .67 & 0 & .33 & 0 & 1 \\ \hline 1.33 & .33 & 1.67 & 1.67 & 5 \end{array}$$

$d_1 = .6, \; p_1 = .4$

$$\mathbf{M}_2 = \begin{array}{cccc|c} 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 2 & 2 & 5 \end{array} \qquad \mathbf{A}_3 = \begin{array}{cccc|c} 0 & .5 & 0 & .5 & 1 \\ 0 & 0 & .5 & .5 & 1 \\ .5 & 0 & .5 & 0 & 1 \\ .5 & 0 & 0 & .5 & 1 \\ .5 & 0 & .5 & 0 & 1 \\ \hline 1.5 & .5 & 1.5 & 1.5 & 5 \end{array} \qquad \mathbf{M}_3 = \begin{array}{cccc|c} 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ \hline 2 & 1 & 1 & 1 & 5 \end{array}$$

$d_2 = .67, \; p_2 = .2$ $\qquad$ $d_3 = .5, \; p_3 = .2$

$$\mathbf{M}_4 = \mathbf{A}_4 = \begin{array}{cccc|c} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 2 & 2 & 5 \end{array}$$

$d_4 = 0, \; p_4 = .2$

Fig. 2. An example of the controlled selection algorithm

We proceed to describe $\mathbf{S}$ and $\mathbf{M}_1, \ldots, \mathbf{M}_l$ for the procedure of this article in greater detail. In our application of controlled selection, each stratum corresponds to a separate controlled selection problem and $\mathbf{S}$ is an $(N+1)\times 5$ array, where N is the number of units in the stratum universe. Thus, there are N internal rows and 4 internal columns in $\mathbf{S}$. Each internal row of the selected array corresponds to a unit in the stratum universe. In the ith internal row, the first element is the probability that the ith 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 four 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 $\mathbf{S}$ are computed. The key value is s_{i3}, the probability that the ith unit is in both samples. Let

$$s_{i3} = \min\{\pi_{i1}, \pi_{i2}\} \tag{2.1}$$

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

$$s_{i4} = 1 - \sum_{j=1}^{3} s_{ij} \tag{2.3}$$

Now (2.1) is motivated by (1.2) and (1.3). That is, if (1.2) held, then the probability that the ith 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 ith unit is in the D_j sample only is simply the probability that it is in the D_j sample minus the probability that it 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.

To illustrate, we consider an example for which $N = 5$, $n_1 = 3$, $n_2 = 2$, and the π_{ij}'s are given in Table 1. Then by (2.1)–(2.3) the array $\mathbf{S}$ is in Figure 1.

Note that by (2.1) and (2.2), all the entries in the second column of $\mathbf{S}$ 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 $\mathbf{S}$.

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. In our example, the arrays $\mathbf{M}_1, \ldots, \mathbf{M}_4$ in Figure 2 are the controlled selection arrays that can be selected. If, to illustrate, $\mathbf{M}_1$ is selected, then units 1 and 4 are in both samples; unit 3 is in the D_1 sample only; and units 2 and 5 are 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**. The probabilities $p_1, \ldots, p_4$ for selecting the arrays $\mathbf{M}_1, \ldots, \mathbf{M}_4$ of our example are listed in Figure 2. (The $\mathbf{A}_k$'s and d_k's, also listed in Figure 2, are obtained as part of the controlled selection algorithm, as will be described in Section 3.)

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

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

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

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

Consequently, if we can establish that

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

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

To additionally establish (1.1) we need only show that

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

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

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

since then, in particular,

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

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 kth possible sample.

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

It is readily verified that the set of arrays $\mathbf{M}_1, \ldots, \mathbf{M}_4$ and associated probabilities $p_1, \ldots, p_4$ in Figure 2 satisfy (2.6)–(2.8) for the array **S** in Figure 1. 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 article, 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 $m_{ij}=\lfloor s_{ij}/b\rfloor$ or $(\lfloor s_{ij}/b\rfloor+1)b$ for all ij, 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. As an example, each of the arrays $\mathbf{M}_1,\ldots,\mathbf{M}_4$ in Figure 2 is a zero-restricted controlled rounding of the array $\mathbf{S}$ in Figure 1, that is each $\mathbf{M}_k$ rounds every element s_{ij} of $\mathbf{S}$ that is not an integer to either the next integer above or the next integer below s_{ij} and leaves integer elements of $\mathbf{S}$ unchanged. In addition, for the arrays in Figure 2, $\mathbf{M}_k$ is a zero-restricted controlled rounding of $\mathbf{A}_k$, $k=1,2,3,4$.

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, $\mathbf{M}_1=(m_{ij1})$, $\mathbf{M}_2=(m_{ij2})$, …, $\mathbf{M}_l=(m_{ijl})$, and associated probabilities, $p_1,\ldots,p_l$, satisfying:

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

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

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

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) and (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) and (2.8).

To illustrate controlled selection, consider the example presented in Section 2. The controlled selection problem $\mathbf{S}$ for this example is in Figure 1. A solution to this problem is the set of arrays presented in Figure 2, together with their associated probabilities.

Although, as noted, any solution to a controlled selection problem satisfies (2.6) and (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 **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

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

and then let

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

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 $d_t = 0$ 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 article that for all i, j, k,

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

Figure 2 illustrates this algorithm for the controlled selection problem of Figure 1.

Now to obtain (2.7), first note that for the array **S** defined by (2.1)–(2.3) we have by (2.2) that

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

Observe that (3.8) is not sufficient to guarantee that all solutions to the controlled selection problem **S** obtained by the algorithm just described satisfy (2.7). To illustrate, for **S** in Figure 1 we have $n_1 = 3$, $m_{(N+1)jk} = 1$ or 2, $j = 1, 3$ and hence $m_{(N+1)1k} + m_{(N+1)3k}$ can equal 2, 3 or 4.

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

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

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

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

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

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

or

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

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

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

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

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

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

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

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

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

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

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 **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) and (3.7) that $0 < a_{i^*1k} < 1$ for some i^* with $1 \le i^* \le N$. In addition, we have $0 < a_{i^*j^*k} < 1$ for some $j^* \in \{2,3,4\}$, since $a_{i^*5k} = 1$ by (3.7). Furthermore, $j^* \neq 2$ 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 $a'_{i^*1k} = a_{i1k} - \epsilon$, $a'_{i^*j^*k} = a_{i^*j^*k} + \epsilon$, $a'_{ijk} = a_{ijk}$ for all other i,j, where $\epsilon > 0$ 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 ith unit is in sample for both designs, the minimum possible value for this probability is $\max\{\pi_{i1} + \pi_{i2} - 1.0\}$.

5. Miscellaneous Issues

We consider here the three separate issues noted in the Introduction.

5.1. Comparison of the current procedure with the procedure in Ernst (1996)

The overlap procedure just described and the overlap procedure in Ernst (1996) share the following common features. Both procedures have been developed for use when the samples for the two designs can be selected simultaneously. Also, both procedures yield optimal solutions to the maximization and also the minimization problem under specified conditions. In fact, both procedures take advantage of the extra flexibility in sample selection offered by simultaneous selection to produce an overlap that is generally higher for the maximization problem and lower for the minimization than can be produced by any overlap procedure that selects the two samples sequentially. This issue is discussed in Ernst

(1996, Section 5). Finally, both procedures use the controlled selection procedure of Causey, Cox, and Ernst (1985).

However, the two procedures use controlled selection in very different ways. The procedure in Ernst (1996) allows for the D_1 and D_2 designs to have different stratifications, but requires the two designs to be 1 unit per stratum designs. The selection of the entire sample for both designs requires the solution of a single controlled selection problem. For this controlled selection problem, each internal row except the final internal row corresponds to a D_1 stratum and each internal column except the final internal column corresponds to a D_2 stratum. Each of the row and column marginals has the value 1, and consequently the selected array $\mathbf{M}_k$ in the solution has a single 1 in each internal row and column, with the remaining internal cells having the value 0. If there is a 1 in cell (i,j) of $\mathbf{M}_k$ where neither i is the final internal row nor j is the final internal column, then a unit that is in both the ith D_1 stratum and jth D_2 stratum is selected to be in sample for both designs from among all such units, using the conditional probabilities in Ernst (1996, Eq. 18). If there is a 1 in the final internal column of row i, then a unit is selected from among all the units in the ith D_1 stratum to be in sample for D_1 only, using the conditional probabilities in Ernst (1996, Eq. 26). Analogously, if there is a 1 in the final row of column j, then a unit is selected from among all the units in the jth D_2 stratum to be in sample for D_2 only, using the conditional probabilities in Ernst (1996, Eq. 27). For the maximization problem the controlled selection array is constructed to maximize the sum of the values of internal cells that are not in the final internal row or column, and hence maximize the expected number of units selected that are in sample for both designs, while for the minimization problem this array is constructed to minimize the sum of the values in these cells.

Unfortunately, this author does not know how to generalize the procedure in Ernst (1996) to designs with more than 1 unit per stratum. The difficulties in developing a generalization are explained in Section 7 of that article.

The current procedure makes use of the identical stratifications assumption for the two designs to construct a separate controlled selection problem for each stratum. The controlled selection array is different here, with each internal row corresponding to a unit and each internal column to the sampling status for the unit.

Note that in the particular case 1 unit per stratum designs with identical stratifications, both procedures are applicable and, since they are both optimal, yield the same expected overlap. Furthermore, the expected overlap for the maximization problem under these conditions is the same as produced by the original procedure of Keyfitz (1951).

5.2. Limitations on stratum size

The title of the article refers to selecting a large number of units per stratum. The procedure that has been presented imposes no theoretical limits on the number of units selected. The only limitation is operational, that is, there is an upper limit to the size of the controlled selection problem that can be solved in practice on a given computer. Furthermore, the size of the controlled problem to solve does not depend directly on the number of units, n_j, selected from S for the D_j sample, but instead on the stratum size, N. The solution of a controlled selection problem requires the solution of a sequence of controlled

rounding problems, each of which requires the solution of a transportation problem. The number of variables in the transportation problem is of the order of the number of internal cells in the controlled selection array, that is $4N$. If N is large enough, the number of variables would be too large for the memory capacity of the computer. However, an N this large is unlikely to occur in practice. For in the application discussed in Ernst and Ikeda (1995), the authors were able to successfully solve transportation problems with as many as 5 million variables, corresponding to an N larger than 1 million.

Of more practical concern, because it can lead to excessively long CPU times, is the number of controlled rounding problems, l, that must be solved in the solution of a controlled selection problem. It can be shown that $3N$ is an upper bound on l. (This is obtained by combining the following three facts: Each $\mathbf{A}_k$ must have at least one more integer valued cell than the preceding member of the sequence. When three internal cells in a row in $\mathbf{A}_k$ are integers, so is the fourth. When k is reached for which $\mathbf{A}_k$ is an integer valued array the algorithm stops.) Furthermore, not only is this upper bound on the number of transportation problems to be solved proportional to N, but from data presented in Ernst and Ikeda (1995, Table 6) it can be surmised that the CPU time required to solve a transportation problem is roughly proportional to the number of variables in the problem. Consequently, the CPU time required for the solution of a controlled selection problem is roughly proportional to N^2. As a result, the procedure may not be practical to run if N exceeds a few thousand.

In the application presented in Section 7, the largest value of N was 214. This did not require a noticeable amount of CPU time.

5.3. Joint selection probabilities

For variance estimation purposes it would be desirable if the procedure presented in Sections 2 and 3 was able not only to satisfy (1.2) for individual units, but also for the inclusion probability, $\pi_{i_1 i_2 j}$, in the D_j sample for each pair of units i_1, i_2 in S to satisfy a predetermined value. Unfortunately, the procedure does not have this property. The value of $\pi_{i_1 i_2 j}$ is readily computable, however. It may be 0 though for some pairs of units, which would preclude the computation of unbiased variance estimates.

To illustrate the computation of $\pi_{i_1 i_2 j}$, consider π_{341} for the solution to the controlled selection problem in Figure 2. This pair of units is in the D_1 sample if either $\mathbf{M}_1$ or $\mathbf{M}_4$ is the selected array, and hence $\pi_{341} = p_1 + p_4 = .6$. However, the same pair of units is not in the D_2 sample no matter which of $\mathbf{M}_1, \ldots, \mathbf{M}_4$ is selected, and hence $\pi_{342} = 0$.

6. Maximization and Minimization of Overlap with Different Stratifications

In the previous sections we have assumed that the two surveys to be overlapped have identical stratifications. We now consider a relatively simple generalization of the work in the previous sections that is applicable when this condition does not hold. Unlike the identical stratification case, this generalization will not guarantee that the optimal overlap is attained.

We introduce the following additional notation. For $k = 1, 2$, let M_k denote the number of D_k strata; let S_{ik}, $i = 1, \ldots, M_k$, denote the set of D_k strata, and let n_{ik} denote the number of sample units to be selected for D_k from S_{ik}. For $i = 1, \ldots, M_1$, $j = 1, \ldots, M_2$, let

$S^*_{ij} = S_{i1} \cap S_{j2}$ and let N_{ij} denote the number of units in S^*_{ij}; let r_{ijk} denote the sum of the D_k selection probabilities for all units in S^*_{ij} and let n_{ijk} denote the number of sample units to be chosen from S^*_{ij} for D_k, $k = 1, 2$. Now n_{ik} is a constant. However, since r_{ijk} is in general not an integer, n_{ijk} must be a random variable, which is chosen to satisfy the following conditions:

$$E(n_{ijk}) = r_{ijk} \text{ and } |n_{ijk} - r_{ijk}| < 1 \text{ for all samples for all } i, j, k$$

For each $i = 1, \ldots, M_1$, an allocation $n_{ij1}, j = 1, \ldots, M_2$, satisfying these conditions, of the n_{i1} units to be selected in S_{i1} among the $S^*_{ij}, j = 1, \ldots, M_2$, can be obtained by systematic probability proportional to size sampling; similarly, for each $j = 1, \ldots, M_2$ an allocation n_{ij2}, $i = 1, \ldots, M_1$, can be obtained of the n_{j2} units to be selected in S_{j2} among the S^*_{ij}, $i = 1, \ldots, M_1$.

Once the allocations n_{ij1}, n_{ij2} are determined for each i, j, the selection of specific units in S^*_{ij} can be determined using the method of the previous sections, with each S^*_{ij} corresponding to a separate controlled selection problem. In applying the procedure to S^*_{ij} we do not use the unconditional D_k selection probabilities, since these probabilities sum to r_{ijk} not n_{ijk}. Instead we use selection probabilities conditioned on n_{ijk}, which are obtained by multiplying the unconditional probabilities by n_{ijk}/r_{ijk}. This approach preserves the unconditional selection probabilities since $E(n_{ijk}) = r_{ijk}$. (This method assumes that none of the conditional selection probabilities exceed 1. Otherwise, the conditional selection probabilities must be computed in a more complex fashion that will not be discussed here.)

The amount of deviation from the optimal overlap when the two designs are not identical and this approach is used, depends on the stratifications and the number of units selected from each stratum for the two designs. The deviation arises from the use of conditional selection probabilities instead of unconditional selection probabilities in choosing the samples for S^*_{ij}. If r_{ij1}, r_{ij2} are both large for all nonempty S^*_{ij}, then n_{ijk}/r_{ijk} will be near 1, and the deviation from optimality will be small. At the opposite extreme, if none of the S^*_{ij} contain more than 1 unit then this approach does no better than independent selection, since if S^*_{ij} consists of 1 unit then that unit will be in the sample for D_k if $n_{ijk} = 1$ and will not be in sample if $n_{ijk} = 0$; hence there is no overlap procedure to apply to S^*_{ij} in that situation.

7. An Application to the Selection of the OCSP and ECI Governments Samples

As noted in the Introduction, our controlled selection procedure was carried out for the selection of the new governments samples for OCSP and ECI, although it was never actually used in production. We detail this application.

The OCSP sample selection process, has traditionally chosen sample establishments with equal probability from industry × employment size class strata within sample geographic PSUs. ECI has in the past chosen sample establishments with probability proportional to size from industry strata without geographic clustering. The two surveys generally have selected their samples independently of each other. However, as part of an effort to integrate the two surveys, all newly selected ECI samples are now selected from OCSP sample PSUs. Furthermore, to reduce data collection expenses it was decided to

have the ECI governments sample selected, if possible, as a subsample of the much larger OCSP sample. To assist this effort, identical industry stratifications were used in both surveys, which had not been the case in the past. Now at the time that the government samples were originally to be drawn we had not had sufficient time to integrate the computer systems for the two surveys, and it was therefore necessary that OCSP and ECI maintain their separate sampling approaches within industry strata in each sample PSU. That is, OCSP was to select its governments sample from industry × employment size class strata, with all establishments within a size class chosen with equal probability, while ECI was to select a single sample for the industry with establishments selected with probability proportional to size. To further complicate matters, at the time the OCSP sample needed to be selected, the ECI sample sizes had not yet been determined.

Consequently, controlled selection was used in the following way. The procedure was applied separately to each OCSP industry × employment size class strata S within each sample PSU. We designate the ECI and OCSP designs as D_1 and D_2, respectively, and let N, n_1, n_2 be as defined in Section 2. In addition, we let T_i denote the measure of size for the ith establishment in S and let $T = \sum_{i=1}^{N} T_i$. T_i is the total employment for the ith establishment, obtained from unemployment insurance records. For each S, the value of N was known and n_1 was determined from the OCSP sample allocation program. Furthermore, $\pi_{i1} = n_1/N$ for all i, while $\pi_{i2} = n_2 T_i/T$. As for n_2, since the number of ECI units to be selected from S was unknown, we selected the maximum number for which $\pi_{i2} \leq \pi_{i1}$, $i = 1, \ldots, N$; that is

$$n_2 = \lfloor n_1 T/(N \max\{T_i : i = 1, \ldots, N\}) \rfloor \tag{7.1}$$

To illustrate, consider an artificial example for which $N = 5$, $n_1 = 4$, and $T_1, \ldots, T_5$ are, respectively, 110, 110, 165, 220, 220. Then $\pi_1 = .8$, $T = 825$, $n_2 = \lfloor .8T/T_4 \rfloor = 3$, and the π_{i2}'s are, respectively, .4, .4, .6, .8, .8. The controlled selection array **S** for this problem is given in Figure 3. **S** is computed as described in Section 2, except that the second column, the ECI sample units only column, is omitted, since it consists solely of 0's.

The solution to the controlled selection problem proceeded as described in Sections 2 and 3 with one major exception. Instead of computing each of the zero-restricted controlled roundings using the transportation problem approach of Cox and Ernst (1982), we were able to use a simplified algorithm, described in Ernst and Ponikowski, because of the special structure of **S** in this application, including the presence of only three internal columns.

The n_2 ECI sample units selected from each stratum S were denoted as the ECI controlled selection sample for S. Later, when the final allocation, denoted n_3, of ECI

$$\mathbf{S} = \begin{array}{|cccc|} \hline .4 & .4 & .2 & 1 \\ .4 & .4 & .2 & 1 \\ .2 & .6 & .2 & 1 \\ 0 & .8 & .2 & 1 \\ 0 & .8 & .2 & 1 \\ \hline 1 & 3 & 1 & 5 \\ \hline \end{array}$$

Fig. 3. Controlled section array for OCSP-ECI example

sample units for each S was determined, using a systematic sampling procedure described in Ernst (1995a), the n_2 ECI units in the controlled selection sample were to be subsampled with equal probability to obtain the final ECI sample from S provided $n_3 \leq n_2$. If $n_3 > n_2$ then all n_2 units in the ECI controlled selection sample were to be part of the final ECI sample from S. The remaining $n_3 - n_2$ units in the final ECI sample were to be selected independently of the controlled selection sample from among all the N units in S using systematic, probability proportional to size, without replacement sampling. (Note that with this approach the same unit can be selected twice for the ECI sample, once if it is among the n_2 ECI units in the controlled selection sample and then a second time if it is among the $n_3 - n_2$ units in the final ECI sample selected in the supplemental sample.) The procedure was actually carried out as far as determining n_3 for each S, but a final ECI sample was never selected.

We computed the expected overlap for the controlled selection procedure as follows. For those strata for which $n_3 \leq n_2$, all ECI sample units would also have been OCSP sample units and hence the expected overlap is n_3. For those strata for which $n_3 > n_2$, the probability of each of the $n_3 - n_2$ units selected into the ECI sample as part of the supplemental sample also being in the OCSP sample is n_1/N and, consequently, the expected number of units in the ECI supplemental sample that would also have been sample units in OCSP is $n_1(n_3 - n_2)/N$. Since all n_2 units in the ECI controlled selection sample are also OCSP sample units, the total expected overlap is $n_2 + n_1(n_3 - n_2)/N$ for those strata for which $n_3 > n_2$. (Note that in this calculation, a unit that is in both the ECI controlled selection sample and the ECI supplemental sample is counted as two overlapped units.)

There were a total of 397 ECI sample establishments that would have been selected using the controlled selection procedure just described. The expected number of these units that would also have been in the OCSP sample was 276.4. Without use of controlled selection, that is if all ECI sample units had been selected independently of the OCSP sample units, the expected number of sample units that would have been in the sample for both surveys would have been 256.4. Thus, the increase in expected overlap by using the controlled selection procedure is relatively small.

There are two reasons why the controlled selection procedure did not produce as large an increase in overlap over independent selection as hoped. First, because the final ECI sample size n_3 to be selected for an OCSP stratum was unknown at the time the OCSP sample was selected, we were forced to use a modified form of controlled selection based on the selection of n_2 units defined by (7.1), which results in a smaller overlap that if the controlled selection procedure was based on the final ECI sample size. In addition, except for 34 units, the ECI sample units were to be selected from two categories of OCSP strata, described below, for which the modified form of controlled selection yielded the same overlap as independent selection. If a larger proportion of the OCSP strata had not been in these two categories the results from using this modified controlled selection procedure would have been better.

Of the 397 ECI sample units to be selected, 205 were to be selected from OCSP certainty strata. The expected number of these units in common to both surveys is 205, whether controlled selection is used or not. The second category of OCSP strata for which the modified form of controlled selection did not increase overlap, from which 158 of the

ECI sample establishments were to be selected, is the set of OCSP noncertainty strata for which $n_1 = 1$. When $n_1 = 1$, it follows from (7.1) that $n_2 = 0$ except in the event, which never occurred, that the measure of size is the same for all units in a stratum. Thus, for the units selected from these 158 strata the expected overlap was the same for both the controlled selection and independent selection procedures, namely 43.2. This is because no units could be selected from these strata by controlled selection, that is all 158 sample units would have to have been selected from the independent supplemental sample. (Although the controlled selection procedure had no effect on the overlap for these 158 strata, we could have used the overlap maximization procedure of Keyfitz (1951) to increase the overlap beyond 43.2.)

For the remaining 34 ECI sample units, that is those that were to be selected from OCSP noncertainty strata with $n_1 > 1$, the expected overlap was 28.2 units for the controlled selection procedure in comparison with 8.2 units for independent selection, and thus the gains from using the controlled selection procedure were noteworthy in this case. For some of the OCSP strata from which these 34 units were to be selected we had $n_3 > n_2$, which is why the controlled selection procedure did not produce a perfect expected overlap for these 34 units.

A natural question to ask regarding this application is if it is really necessary to use the controlled selection process at all. That is, in general, can (1.1), (1.2), and (1.5) be satisfied by first selecting the n_1 units in the D_1 sample from S and then subsampling in some way these n_1 units to obtain the D_2 sample from S? The following example illustrates that this is generally not possible. Let $N = 4$, $n_1 = 2$, $n_2 = 1$, $\pi_{i1} = .5$, $i = 1, \ldots, 4$, $\pi_{12} = \pi_{22} = .45$, $\pi_{32} = \pi_{42} = .05$. Then if the D_1 sample is selected by simple random sampling without replacement, the probability that the D_1 sample would consist of units 3 and 4 would be 1/6; if one of these two units is then selected to be the D_2 sample, either the probability that unit 3 or the probability that unit 4 is in the D_2 sample must be at least 1/12 and hence the requirement $\pi_{32} = \pi_{42} = .05$ cannot be met.

8. 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 Wolters, C.L. (1987). Additional Uses for Keyfitz Selection. Proceedings of the American Statistical Association, Section on Survey Research Methods, 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, U.S. Bureau of Labor Statistics, November 20.

Ernst, L.R. (1995b). Maximizing and Minimizing Overlap of Ultimate Sampling Units. Proceedings of the American Statistical Association, Section on Survey Research Methods, 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, U.S. Bureau of Labor Statistics, January 8.

Gilliland, P.D. (1984). 1985 PATC Measures of Size Specifications. Memorandum to W. Brown, U.S. Bureau of Labor Statistics, 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 NSR Strata. Memorandum to Joseph Waksberg, U.S. Bureau of the Census, 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 American Statistical Association, Section on Survey Research Methods, 747–752.

Received December 1996
Revised October 1997