Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Domain sample allocation within primary sampling units in designing domain-level equal probability selection methods

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

by Avinash C. Singh and Rachel M. Harter

Release date: December 17, 2015

Statistics Canada
Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Domain sample allocation within primary sampling units in designing domain-level equal probability selection methods

Avinash C. Singh and Rachel M. Harter[1]

Abstract

Self-weighting estimation through equal probability selection methods (*epsem*) is desirable for variance efficiency. Traditionally, the *epsem* property for (one phase) two stage designs for estimating population-level parameters is realized by using each primary sampling unit (PSU) population count as the measure of size for PSU selection along with equal sample size allocation per PSU under simple random sampling (SRS) of elementary units. However, when self-weighting estimates are desired for parameters corresponding to multiple domains under a pre-specified sample allocation to domains, Folsom, Potter and Williams (1987) showed that a *composite measure of size* can be used to select PSUs to obtain *epsem* designs when besides domain-level PSU counts (i.e., distribution of domain population over PSUs), frame-level domain identifiers for elementary units are also assumed to be available. The term *depsem*-A will be used to denote such (one phase) two stage designs to obtain domain-level *epsem* estimation. Folsom et al. also considered two phase two stage designs when domain-level PSU counts are unknown, but whole PSU counts are known. For these designs (to be termed *depsem*-B) with PSUs selected proportional to the usual size measure (i.e., the total PSU count) at the first stage, all elementary units within each selected PSU are first screened for classification into domains in the first phase of data collection before SRS selection at the second stage. Domain-stratified samples are then selected within PSUs with suitably chosen domain sampling rates such that the desired domain sample sizes are achieved and the resulting design is self-weighting. In this paper, we first present a simple justification of composite measures of size for the *depsem*-A design and of the domain sampling rates for the *depsem*-B design. Then, for *depsem*-A and -B designs, we propose generalizations, first to cases where frame-level domain identifiers for elementary units are not available and domain-level PSU counts are only approximately known from alternative sources, and second to cases where PSU size measures are pre-specified based on other practical and desirable considerations of over- and under-sampling of certain domains. We also present a further generalization in the presence of subsampling of elementary units and nonresponse within selected PSUs at the first phase before selecting phase two elementary units from domains within each selected PSU. This final generalization of *depsem*-B is illustrated for an area sample of housing units.

Key Words: *Epsem* and *depsem* designs; Multiple domain estimation; Self-weighting estimation; Two phase two stage designs.

1 Introduction

For multi-stage design of surveys, an equal probability selection method (or *epsem*, Kish 1965, page 21) is typically desired toward the goal of variance reduction or variance efficiency. In practice, for two or more stage designs, selection probabilities for primary (or first stage) sampling units (PSUs) are often driven by considerations of over- (under-) sampling to obtain adequate domain sample sizes, and operational efficiency such as equal interviewer workload per PSU. The simplest type of an *epsem* design is a single stage simple random sampling (SRS) design without replacement of elementary units with selection probabilities n/N where n, N denote respectively the sample and population sizes. Another example is single stage stratified SRS with proportional allocation; i.e., $n_h / N_h \propto 1$, or $n_h = fN_h$ where $f = n/N$, and n_h, N_h are sample and population sizes, respectively, for the h^{th} stratum. These and other *epsem* designs are described in fundamental sampling texts such as those by Cochran (1977) and Lohr (2010).

[1] Avinash C. Singh, NORC at the University of Chicago, 55 East Monroe St., 20th Floor, Chicago, IL 60603. E-mail: singh-avi@norc.org; Rachel M. Harter, RTI International, P.O. Box 12194, Research Triangle Park, NC 27709.

Yet another example of an *epsem* design is single stage SRS of whole clusters. For area sampling in field surveys, clusters are useful for operational efficiency due to reduced travel cost in interviewing neighboring housing units although there are some drawbacks. Cluster sizes could vary considerably making the logistics difficult for equalizing interviewer workloads. Moreover, a complete enumeration of each cluster may not be desirable due to cost, and inefficient estimation due to reduced effective sample size as a result of intra-cluster correlations. In general, the probability proportional to size (pps) sampling of clusters followed by equal sample allocation of elementary units per cluster to equalize interviewer assignments is a reasonable and practical compromise for area cluster sampling.

Above considerations lead to two stage designs with first stage selection probabilities to be denoted by π_i for the i^{th} PSU, and second stage conditional selection probabilities to be denoted by $\pi_{j|i}$ for the j^{th} elementary unit within the i^{th} selected PSU. For example, in a survey of teachers, PSUs could be schools, while ultimate sampling or elementary units could be teachers within schools. For SRS of size n_i^* within each PSU i with population count N_i, the probabilities $\pi_{j|i}$ and π_i can be defined as follows to obtain an *epsem* design; see Kish (1965, page 222). Here n_i^* are common and equal to n/m where m is the desired number of selected PSUs out of a total of M PSUs in the population. We have

$$\pi_i = m\frac{N_i}{N},\ \pi_{j|i} = \frac{n_i^*}{N_i} = \frac{n}{mN_i}. \tag{1.1}$$

It is easily seen, as expected, that the sum of π_i's over all M PSUs i is the fixed sample size m at the first stage, and the sum of $\pi_{j|i}$'s over all N_i elementary units j within the i^{th} PSU is the fixed sample size n_i^* at the second stage. Moreover, the unconditional (same as joint because of nesting of units within PSUs) selection probability for the j^{th} unit in the i^{th} PSU is the product $\pi_i\pi_{j|i}$; i.e., n/N or f, which is equal for all units, as desired. For generalizations of self-weighting estimation considered in this paper, it is useful to express the implied sample allocation n_i to the i^{th} PSU from (1.1) as

$$n_i = \left(\frac{f}{\pi_i}\right)N_i, \tag{1.2}$$

based on the observation $\pi_i \times n_i/N_i$ equals f where f is the desired sampling fraction n/N. Here, the value of n_i is obtained as $n_i^*\ (= n/m)$. Note that if all PSUs are selected with certainty; i.e., $\pi_i = 1$, the above PSU$-$level allocation reduces to proportional allocation in stratified designs with the number of strata being the total number M of PSUs.

The basic idea for making any design *epsem* is to work backwards; that is, before specifying selection probabilities π_i for PSUs, it is ensured that the sampling rate within any given PSU i is inversely proportional to π_i so that π_i cancels out in the unconditional selection probability $\pi_i\pi_{j|i}$ within the PSU. In this way the unconditional selection probabilities for elementary units can be made common for all sampled units from different PSUs. We will use this strategy throughout the paper.

From (1.1), observe that in order for $n_i \le N_i$, we must have $f \le \pi_i$ for all $i = 1, \ldots, M$. This condition can be satisfied at the design stage by collapsing neighboring PSUs in order to increase N_i (and hence π_i) or by reducing f if necessary. In other words, the sample allocated to the i^{th} PSU must be a fraction of the PSU population size N_i where the fraction is given by the ratio of the desired sampling rate f and the PSU selection probability π_i.

So far we considered *epsem* designs for a single domain; i.e., estimation at the population level only. However, often survey designs are intended to support analytical goals for multiple domains within the target population. For example, in the case of a teacher survey, domains could be male and female teachers. For domain-level *epsem* designs (to be termed *depsem* in this paper), Folsom, Potter and Williams (1987) presented a method for allocating a sample of units to PSUs under two separate designs - $depsem-\mathrm{A1}$ and $depsem-\mathrm{B1}$ defined as follows; the numeric extension in the notation is used to differentiate them from other variations presented later.

The $depsem-\mathrm{A1}$ *Design* is defined as a one phase two stage design where domain-level PSU population counts (N_{id} for the i^{th} PSU and d^{th} domain), desired domain sample size $\left(n_{+d}^{*}\right)$ where '+' denotes sum over m selected PSUs, and equal PSU sample allocation $n_{i+}^{*}\,(= n/m)$ over all domains (i.e., equal interviewer load) are specified. Thus, the desired sampling rate (f_d) for each domain is pre-specified but the PSU selection probabilities (π_i) are not pre-specified and are suitably defined to obtain the *depsem* property. Here it is also assumed that frame-level domain identifiers for elementary units are available. Such a design is applicable to situations where in-person interviews with a list frame are desirable.

The $depsem-\mathrm{B1}$ *Design* is defined as a two phase two stage design where PSU population counts (N_{i+}) and desired domain sample size $\left(n_{+d}^{*}\right)$ over m selected PSUs are specified. Domain-level population counts (N_{+d}) are not specified (which of course implies that domain-level PSU population counts N_{id} are not specified), and $\mathrm{PSU}-$level sample allocations (n_{i+} over all domains) are also not pre-specified. In addition, the desired sampling rates for each domain (f_d) are not pre-specified. However, PSU selection probabilities are specified by using PSU population counts as size measures, and for selected PSUs in the first stage, domain-level population counts N_{id} become available after the first phase census. Here the domain sampling rates f_d are suitably defined to obtain the *depsem* property. The two phase aspect of the design is used to obtain domain membership of selected units in the first phase through screening. Such a design may be applicable more generally than the previous one.

The school/teacher example can be used to make these two *depsem* designs concrete. In $depsem-\mathrm{A1}$, we know in advance how many male and female teachers are in each school from the list frame, and also we know which teachers are male and which are female. The desired sampling rates of male and female teachers, and the equal number of teachers to be selected per school are known. Then school or PSU selection probabilities are obtained to satisfy the *depsem* property. In $depsem-\mathrm{B1}$, we know the probability of selecting each school based on the total number of teachers per school. We do not know how many male and female teachers are in each school, but the desired numbers of male and female teachers in the sample over all selected schools are specified. Then, after screening all teachers in the selected schools for male/female classification, the sampling rates for male and female domains for each pre-selected school are obtained to satisfy the *depsem* property.

For $depsem-\mathrm{A1}$, Folsom et al. (1987) provide a composite measure of size for selecting PSUs such that its inverse appears in the specification of domain sample allocations within each PSU. The sample allocation to domains within PSUs satisfies the desired PSU sample size or interviewer workload exactly. However, the desired domain sample size is achieved only in expectation because the sample size of elementary units within domains is not directly controlled, but the PSU sample size is controlled instead to obtain equal interviewer workload.

For $depsem-\mathrm{B1}$, the same basic method is inverted to produce $depsem$ samples. Here, in the first phase, a census of selected PSUs at the first stage is conducted so that all elementary units within selected PSUs are stratified into domains to obtain domain-level PSU counts and are subsampled such that the desired domain sample size over all PSUs is satisfied. However, any constraint on the PSU sample size is relaxed in the interest of obtaining a $depsem$ sample. $Depsem-\mathrm{B1}$ may be particularly useful for non-face-to-face interview modes such as telephone surveys in the second phase, where the first phase sample of elementary units is used to obtain contact information and domain classification. The first phase results may be based on a self-administered screening questionnaire sent by mail or dropped off after an in-person contact effort to all or a large sample of units in each selected PSU. If the main interview is conducted by phone in the second phase, having equal interviewer workload per PSU is of no practical consequence. Folsom et al. (1987) also considered natural generalizations of both $depsem$ designs to the case of stratified population of PSUs in the first phase.

In this paper, we introduce a systematic general framework for defining $depsem$ designs which provides a simple justification for the $depsem$ property of the above two designs. We then propose generalizations of the two designs under the above framework to obtain new useful variations of $depsem$ designs encountered in practice; see Singh and Harter (2011) for an earlier development. See also Fahimi and Judkins (1991) for an interesting simulation study comparing traditional and nontraditional measures of size with respect to between PSU variance contributions. The organization of this paper is as follows. Section 2 reviews the original composite measure of size method for selecting PSUs as proposed by Folsom et al. (1987) for the $depsem-\mathrm{A1}$ design including its stratified version. Section 3 presents the inverted method of Folsom et al. (1987) for $depsem-\mathrm{B1}$ to obtain domain-level sampling rates over all pre-selected PSUs. Section 4 presents a generalization to a hybrid $depsem-\mathrm{AB}$ design where the domain-level PSU counts for all PSUs are assumed to be only approximately known, and are used first to specify PSU selection probabilities obtained as composite measures of size as in $depsem-\mathrm{A1}$, and then sampling rates from selected PSUs are specified as in $depsem-\mathrm{B1}$ by obtaining true domain-level PSU counts for selected PSUs through first phase screening. Another generalization considered in Section 4 is when PSUs in the first phase are selected with arbitrarily pre-specified selection probabilities. Section 5 further generalizes $depsem-\mathrm{B1}$ to designs where the second phase sample within each selected PSU is not a census (i.e., there is subsampling within PSUs) or when it is a census but is subject to nonresponse, or both. Generalizations to stratified designs are also considered in Section 5. Section 6 presents a hypothetical but realistic example based on a study for which the proposed $depsem$ designs were developed under a two-phase two stage design to establish nationally representative norms for an English and Spanish instrument toolbox for assessing behavioral and cognitive functions. We conclude with remarks in Section 7.

2 Review of $depsem-\mathrm{A}$ designs with a simple justification

Consider a one phase two-stage design where the first stage units are schools, for example, and the second stage units are individual teachers. Two domains of interest may be male and female teachers. Under $depsem-\mathrm{A1}$, it is assumed that the $\mathrm{PSU}-$domain population counts (N_{id}) are known, where the

PSU index i varies from 1 to M, the total number of schools; and the domain index d varies from 1 to D, where D in this example is 2 for male and female teachers. In addition, it is assumed that the frame-level domain identifiers (male/female) are available for each teacher in the list. Now, suppose the desired number of sampled teachers for each domain d is n_{+d}^{*} based on precision requirements, where the subscript '+' in n_{+d}^{*} denotes aggregation over selected PSUs i varying from 1 to m. The sum of n_{+d}^{*} over all domains is the total sample size n. Then we know the desired sampling rate for domain d teachers is $f_d = n_{+d}^{*} / N_{+d}$ where N_{+d} is the sum of N_{id} for domain d across all M schools. In addition, it is desired to have equal sample sizes in all m selected schools; i.e., $n_{i+}^{*} = n/m$ for $i = 1, \ldots, m.$

Folsom et al. (1987) proposed a composite measure of size for selecting schools which can be used to allocate the desired number of sampled teachers within schools in such a way that the selected teachers provide *epsem* designs for both male and female teacher domains. The design satisfies exactly the specified equal sample size $n_{i+}^{*}\,(= n/m)$ for all selected schools but only in expectation the specified domain sample size n_{+d}^{*}. Clearly, it is practical to control directly the sample size within each selected school and not the domain sample size overall selected schools.

We provide a different but simpler derivation of the results given in Folsom et al. (1987). To this end, we observe that the key result (1.2) for *epsem* designs implies that the sampling rate n_{id} / N_{id} in domain d within PSU i should be proportional to f_d / π_i. This is true regardless of how the PSU selection probabilities π_i or the domain sampling rates f_d are specified. For *depsem*−A1, although frame-level domain identifiers for elementary units are assumed to be known, it may not be cost efficient to directly draw samples from domains after stratifying the frame. It may be preferable to select PSUs in the first stage which are then stratified by domains using frame-level information before the second stage sample selection using SRS. So in the interest of equal interviewer workload per PSU, we consider the allocation of the desired sample size n_{i+}^{*} for a given PSU i to domains so that the PSU i sample size is controlled at the desired value. However, the realized domain sample size then becomes random and can be made to satisfy the desired goal in expectation.

***Depsem*−A1 Design:** For each given PSU i, the sample allocations n_{id}^{A1} over domains are obtained as $n_{id}^{A1} \propto f_d N_{id} / \pi_i$, which implies that

$$n_{id}^{A1} = \left(n_{i+}^{*}\right) \frac{f_d N_{id} / \pi_i}{\sum_{d'=1}^{D} f_{d'} N_{id'} / \pi_i} = \left(\frac{n}{m}\right) \frac{f_d N_{id}}{S_i}, \tag{2.1}$$

where S_i denotes $\sum_{d'=1}^{D} f_{d'} N_{id'}$ as the unspecified π_i cancels out. However, we can set π_i^{A1}, the selection probability for PSU i, as mS_i / S_+, where $S_+ = \sum_{i=1}^{M} \sum_{d=1}^{D} f_d N_{id}$. By exchanging summations, S_+ reduces to $\sum_{d=1}^{D} f_d \sum_{i=1}^{M} N_{id}$ or $\sum_{d=1}^{D} f_d N_{+d}\,(= n)$. Then the allocated sample size n_{id}^{A1} over domains can be expressed analogous to (1.2) as

$$n_{id}^{A1} = \left(\frac{f_d}{\pi_i^{A1}}\right) N_{id} \ . \tag{2.2}$$

Observe that if $\pi_i^{A1} = 1;$ i.e., if PSUs are selected with certainty, the above allocation behaves like the proportional allocation in stratified designs for domains within PSUs acting as strata. It is easy to show from equation (2.2) that the probability of an individual teacher j being selected is equal for all sampled teachers in domain d where teachers are selected by a stratified SRS from each selected PSU stratified by domains. The probability depends only on d because

$$\text{Pr (teacher } j \text{ in domain } d \,|\, \text{school } i\text{) Pr (school } i\text{)} = \frac{n_{id}^{A1}}{N_{id}} \pi_i^{A1} = f_d . \tag{2.3}$$

Thus $S_i,$ a composite measure of size, provides the appropriate size measure for PSU i to obtain a $depsem - \text{A1}$ design. Unlike the traditional size measure given by the PSU population count N_i used in population level *epsem* designs, the new size measure S_i depends on the desired domain sample size n_{+d}^* as well as the domain-level PSU population size N_{id} because of domain-level *epsem* requirements. The measure S_i can be interpreted as the approximate total desired sample size over all domains within each PSU i.

It is also observed that for PSU i while the sample allocations $\{n_{id}^{A1}\}_{1 \le d \le D}$ over domains satisfy the desired sample size n_{i+}^* exactly by construction $\left(\text{i.e.,} \sum_{d=1}^{D} n_{id}^{A1} = n_{i+}^*\right),$ the resulting allocations $\{n_{id}^{A1}\}_{1 \le i \le m}$ for any given domain d over selected PSUs satisfy the desired sample size n_{+d}^* only in expectation; i.e.,

$$E\left(n_{+d}^{A1}\right) = f_d E\left(\sum_{i=1}^{m} \frac{N_{id}}{\pi_i^{A1}}\right) = \frac{n_{+d}^*}{N_{+d}} E\left(\hat{N}_{+d}^{A1}\right) = n_{+d}^* , \tag{2.4}$$

where $\hat{N}_{+d}^{A1}$ denotes $\sum_{i=1}^{m} N_{id} \big/ \pi_i^A$ and estimates N_{+d} unbiasedly, and E is the expectation operator for the first stage randomization.

It should be remarked that in practice the allocations $\{n_{id}^{A1}\}_{1 \le d \le D}$ need not be integers, and may require random rounding. To do this, consider the fractional parts $\{n_{id}^{A1} - [n_{id}^{A1}]\}_{1 \le d \le D}$ where [.] denotes the greatest integer contained in the quantity in brackets. These fractional parts in a PSU can be treated as selection probabilities for selecting without replacement a sample of size defined by the sum of the fractional parts for that PSU, which is necessarily an integer. Then allocations for domains so selected are rounded up, while for others they are rounded down. Thus the randomly rounded domain allocations continue to satisfy the condition of fixed sample size $n_{i+}^*,$ but the desired domain allocation n_{+d}^* is now satisfied under the joint expectation of random design and rounding mechanisms.

The above derivation of sample allocation assumed implicitly that $n_{id}^{A1} \le N_{id};$ i.e., the allocated sample size does not exceed the corresponding population size. This assumption requires that the factor $\left(f_d \big/ \pi_i^{A1}\right)$ must be less than or equal to 1 for all d and i in view of (2.2). In other words, we must have

$$\max\{f_d\}_{1 \le d \le D} \le \min\{\pi_i^{A1}\}_{1 \le i \le M} . \tag{2.5}$$

By reducing values of $f_d,$ or by collapsing neighboring PSUs to increase $\pi_i^{A1},$ it is generally not difficult in practice to satisfy the above condition. Incidentally, randomly rounded n_{id}^{A1}'s continue to satisfy (2.5) if the original n_{id}^{A1}'s do.

***Depsem*−A2 Design:** All the above results easily generalize to stratified two stage designs denoted by $depsem-\text{A2}$; e.g., schools may be stratified by school districts in our simple example. Specifically, the key result (2.2) is generalized to obtain the domain sample allocations $\left\{ n_{hid}^{A2} \right\}_{1 \le d \le D}$ of n_{hi+}^{*} within PSU i of the h^{th} stratum, $h = 1, \ldots, H$, as follows:

$$n_{hid}^{A2} = \left(\frac{f_d}{\pi_{hi}^{A2}} \right) N_{hid}, \tag{2.6}$$

where notations with subscript h signify that the terms are stratum-specific. Other results mentioned above for the unstratified case can be easily extended in an analogous manner to the stratified case.

3 Review of *depsem*−B designs with a simple justification

Now suppose that the schools have already been sampled with usual PSU population counts N_i as size measures, and therefore their probabilities of selection π_i are known as given in (1.1). Under $depsem-\text{B1}$ involving two phase designs, f_d's are not specified, but the desired values of n_{+d}^{*} are pre-specified for all domains. For example, the schools are pre-selected and the desired numbers of sampled male and female teachers are pre-specified, but the sampling rates for male and female teachers are not specified. It is still possible to select *epsem* samples of male and female teachers using suitable values of f_d in (2.2), but not with equal sample size per school, as shown in Folsom et al. (1987) and described below.

***Depsem*−B1 Design:** Here π_i is set by the usual size measure as mN_i / N for $i = 1, \ldots, M$. Denote it by π_i^{B1}. As in $depsem-\text{A1}$, the sampling rate n_{id} / N_{id} in domain d within PSU i should be set proportional to f_d / π_i^{B1}, although here f_d is not known. Under $depsem-\text{B1}$, each selected PSU is stratified by domains for the selection of elementary units using the first phase domain-screening information while under $depsem-\text{A1}$, domain memberships of elementary units are assumed to be available in the frame itself. In this case, the condition of equal interviewer workload per PSU is relaxed and the desired PSU sample sizes n_{i+}^{*} over all domains are not pre-specified. Instead, it is the desired domain sample size n_{+d}^{*} that is directly controlled by allocating it to PSUs within each domain d. Thus, n_{+d}^{*} is rendered nonrandom which is clearly preferable for control on resulting precision of domain level estimates. It follows that, analogous to (2.1), the sample allocations $\left\{ n_{id}^{B1} \right\}_{1 \le i \le m}$ of the domain total n_{+d}^{*} for each domain d to selected PSUs are given by

$$n_{id}^{B1} = \left(n_{+d}^{*} \right) \frac{f_d N_{id} / \pi_i^{B1}}{\sum_{i'=1}^{m} f_d N_{i'd} / \pi_{i'}^{B1}} = \left(\frac{n_{+d}^{*}}{\hat{N}_{+d}^{B1}} \right) \frac{N_{id}}{\pi_i^{B1}} = \left(\frac{\hat{f}_d^{B1}}{\pi_i^{B1}} \right) N_{id}, \tag{3.1}$$

where $\hat{N}_{+d}^{B1} = \sum_{i'=1}^{m} N_{i'd} / \pi_{i'}^{B1}$ and the unspecified f_d cancels out. However, we can set $\hat{f}_d^{B1}$, the sampling rate for domain d, as $n_{+d}^{*} / \hat{N}_{+d}^{B1}$. Clearly, $\left\{ n_{id}^{B1} \right\}_{1 \le i \le m}$ satisfies n_{+d}^{*} exactly by construction. However, the allocations do not satisfy n_{i+}^{*} (or n/m), even in expectation, because in general

$$E\left(n_{i+}^{B1}\right) = E\left(\frac{\sum_{d=1}^{D} \hat{f}_d^{B1} N_{id}}{\pi_i^{B1}}\right) = \left(\frac{n}{m}\right) \frac{\sum_{d=1}^{D} E\left(\hat{f}_d^{B1}\right) N_{id}}{\sum_{d=1}^{D} f N_{id}} \neq \frac{n}{m}, \tag{3.2}$$

unless $\hat{f}_d^{B1}$ is constant and equals $f\,(= n/N)$, which is in conflict with the desired disproportionate domain allocations.

Other considerations such as random rounding of n_{id}^{B1} to obtain integer allocations carry over in a manner analogous to $depsem-\mathrm{A1}$. However, if the requirement of $n_{id}^{B1} \leq N_{id}$ for all domains within each $i = 1, \ldots, m,$ is not satisfied, one option is to reduce n_{+d}^{*}, while the other option is to collapse neighboring PSUs. For example, collapsing i and i', and letting $\tilde{i}$ denote the collapsed PSU, we have $N_{\tilde{i}d} = N_{id} + N_{i'd}$. Then $\pi_{\tilde{i}}^{B1}$ required for calculating sample allocations in the second phase from (3.2) is now given by $\pi_{\tilde{i}}^{B1} = \pi_i^{B1} + \pi_{i'}^{B1} - \pi_{ii'}^{B1}$ which, incidentally, also requires knowledge of the second order inclusion probability $\pi_{ii'}^{B1}$.

***Depsem*–B2 Design:** We next consider a generalization of the above case to stratified designs. In our example of the teacher survey, this case corresponds to schools stratified by school districts. This extension carries over in a manner analogous to $depsem$ – A2. That is, suppose for the first phase sample, m_h PSUs are to be selected from the h^{th} stratum, $h = 1, \ldots, H$ with the usual pre-specified selection probabilities $m_h N_{hi} / N_h$ to be denoted by π_{hi}^{B2}. The sample allocations $\left\{n_{hid}^{B2}\right\}_{1\leq i\leq m_h, 1\leq h\leq H}$ of the domain total n_{++d}^{*} to selected PSUs within each stratum h, analogous to formula (3.1) of $depsem-\mathrm{B1}$, are given by

$$n_{hid}^{B2} = \left(\frac{\hat{f}_d^{B2}}{\pi_{hi}^{B2}}\right) N_{hid}, \tag{3.3}$$

where N_{hid} is the domain d population count within PSU i and stratum h,

$$\hat{f}_d^{B2} = n_{++d}^{*} \big/ \hat{N}_{++d}^{B2}, \text{ and } \hat{N}_{++d}^{B2} = \sum_{h=1}^{H} \sum_{i=1}^{m_h} N_{hid} \Big/ \pi_{hi}^{B2}.$$

4 Proposed generalizations of *depsem* – A/B designs

$Depsem-\mathrm{A1}$ and $-B1$ designs require relatively stringent assumptions regarding the provision of frame-level domain membership information of elementary units for $depsem-\mathrm{A}$ designs and domain-screening of all elementary units from selected PSUs in the first phase for $depsem-\mathrm{B}$ designs. In practice, the assumptions may not be true exactly, yet the goal of $depsem$ sample designs may still be desirable. In this section we loosen the requirement for $depsem-\mathrm{A1}$ that domain-level PSU counts are known exactly which leads to a new hybrid design $depsem-\mathrm{AB1}$ where PSU – domain counts are initially assumed to be only approximately known in order to specify PSU selection probabilities as in $depsem-\mathrm{A1}$. Later, true domain-level PSU counts for selected PSUs at the first stage are obtained as in $depsem-\mathrm{B1}$ by conducting a census of elementary units within PSUs in the first phase. Another design termed $depsem-\mathrm{C}$ generalizes $depsem-\mathrm{B}$ by employing a general pre-specification of PSU selection probabilities. Both cases use the same strategy of making the domain-level PSU sampling rates inversely proportional to the PSU selection probabilities. Table 4.1 provides a quick summary of old and new designs considered in this paper.

Table 4.1
Summary of different *depsem* designs (old and new)

***Depsem* design description unstratified (or stratified)**	**PSU selection probability π_i (or π_{hi})**	**Domain-level PSU population count N_{id} (or N_{hid})**	**Domain sampling rate f_d**	**Domain sample size n_{+d} (or n_{++d})**	**PSU sample size n_{i+} (or n_{hi+})**
A1 (or A2): One phase two stage (Old)	Find	Specified (also frame-level domain identifiers)	Specified	Specified (in expectation)	Specified
B1(or B2): Two phase two stage (Old)	Specified	Obtain from phase one census of selected PSUs	*Find*	Specified	Not specified
AB1 (or AB2): Hybrid one/two phase two stage (New)	Specified using A1 and initial values $\tilde{N}_{id}$	Specified approximate initial values $\tilde{N}_{id}$ for all PSUs; and exact values N_{id} for selected PSUs from phase one census	*Find*	Specified	Not specified
C1 (or C2): Two phase two stage (New)	Specified	Specified from first phase census of selected PSUs	*Find*	Specified	Not specified
$C1'$ (or $C2'$): Two phase two stage with subsampling and nonresponse at phase one (New)	Specified (also response and subsampling rates within PSUs)	Specified for selected PSUs from phase one respondents	*Find*	Specified	Not specified

***Depsem*$-$AB1 Design:** Consider a new case for *depsem* designs using a variation of $depsem-\mathrm{A1}$ in which domain-level PSU population counts N_{id} are only approximately known and given by $\tilde{N}_{id}$. The approximations may be available from alternative sources such as the most recent census or a suitable administrative database. In our teacher example, the number of male and female teachers in each school may be known for the prior year, which serves as an approximation for the current year domain-level PSU counts. The m PSUs are selected using π_i^{AB1} probabilities which, similar to π_i^{A1} under $depsem-\mathrm{A1}$, are defined as

$$\pi_i^{AB1} = m\sum_{d=1}^{D} \tilde{f}_d \tilde{N}_{id} \Big/ n,\ \tilde{f}_d = n_{+d}^{*} \big/ \tilde{N}_{+d}\,. \tag{4.1}$$

Now we consider two phases in addition to two stages, as in $depsem-\mathrm{B1}$, because first stage units within selected PSUs need to be classified into domains, and corresponding true counts N_{id}'s are to be determined. In this case, all elementary units in the PSU are selected in the first phase sample. Now, analogous to formula (3.1) of $depsem-\mathrm{B1}$, the sample allocations $\{n_{id}^{AB1}\}_{1\le i\le m}$ of the domain sample size n_{+d}^{*} to selected PSUs are given by

$$n_{id}^{AB1} = \left(\frac{\hat{f}_d^{AB1}}{\pi_i^{AB1}} \right) N_{id} \tag{4.2}$$

where $\hat{f}_d^{AB1} = n_{+d}^* \big/ \hat{N}_{+d}^{AB1}$, and $\hat{N}_{+d}^{AB1} = \sum_{i=1}^m N_{id} \big/ \pi_i^{AB1}$. Clearly, $\{n_{id}^{AB1}\}_{1 \le i \le m}$ satisfies n_{+d}^* but does not satisfy n_{i+}^* (or n/m), even in expectation as in $depsem-\text{B1}$, because, in general,

$$E\left(n_{i+}^{AB1}\right) = E\left(\frac{\sum_{d=1}^D \hat{f}_d^{AB1} N_{id}}{\pi_i^{AB1}} \right) = \left(\frac{n}{m} \right) \frac{\sum_{d=1}^D E\left(\hat{f}_d^{AB1}\right) N_{id}}{\sum_{d=1}^D \tilde{f}_d \tilde{N}_{id}} \neq \frac{n}{m}. \tag{4.3}$$

In fact, using Jensen's inequality, it follows that

$$E(n_{i+}^{AB1}) \ge \left(\frac{n}{m} \right) \frac{\sum_{d=1}^D f_d N_{id}}{\sum_{d=1}^D \tilde{f}_d \tilde{N}_{id}} \tag{4.4}$$

where f_d is the domain sampling rate corresponding to the true unknown N_{+d}. Other considerations such as random rounding of n_{id}^{AB1} to obtain integer allocations, the requirement of $n_{id}^{AB1} \le N_{id}$ for all domains within each $i = 1, \ldots, m,$ and the extension to stratified designs (denote by $depsem-\text{AB2}$) carry over in a manner analogous to formula (3.3) for $depsem-\text{B2}$.

***Depsem*–C1 Design:** We propose a $depsem$ design more general than $depsem-\text{AB1}$ for pre-specified π_i (or π_i^{C1}) when PSU–domain population counts are not known even approximately, so $depsem-\text{AB1}$ is not applicable. As in $depsem-\text{AB1}$, true counts of the PSU–domain sizes N_{id} are obtained through the use of a phase one census of elementary units within selected PSUs. For example, suppose no information about the number of male and female teachers is available for the selected schools. After the schools are selected, we obtain the sex of every teacher in the selected schools for stratification and selection in phase two.

The phase two sample allocations of the desired domain sample sizes to selected PSUs and their properties for $depsem-\text{C1}$ follow easily from those for $depsem-\text{AB1}$. The sample allocations $\{n_{id}^{C1}\}_{1 \le i \le m}$ of the domain total n_{+d}^* for each domain d to selected PSUs are given by

$$n_{id}^{C1} = \left(n_{+d}^*\right) \frac{f_d N_{id} \big/ \pi_i^{C1}}{\sum_{i'=1}^m f_d N_{i'd} \big/ \pi_{i'}^{C1}} = \left(\frac{n_{+d}^*}{\hat{N}_{+d}^{C1}} \right) \frac{N_{id}}{\pi_i^{C1}} = \left(\frac{\hat{f}_d^{C1}}{\pi_i^{C1}} \right) N_{id}, \tag{4.5}$$

where $\hat{N}_{+d}^{C1} = \sum_{i=1}^m N_{i'd} \big/ \pi_{i'}^{C1}$ as the unspecified f_d cancels out. Here, we can set $\hat{f}_d^{C1}$, the domain-level sampling rate, as $n_{+d}^* \big/ \hat{N}_{+d}^{C1}$. As before, an extension to stratified designs (denote by $depsem - \text{C2}$) carries over in a manner analogous to formula (3.3) for $depsem-\text{B2}$.

5 Generalizations of $depsem-\mathrm{C}$ designs in the presence of subsampling within PSUs and nonresponse at the first phase

Often in practice there is subsampling of elementary units within selected PSUs in the first phase because conducting a census of each selected PSU for domain classification may be too costly. In this section, we will consider further generalizations of $depsem-\mathrm{C1}$ where the within$-$PSU domain totals are estimated through a sub-sample of elementary units in the first phase rather than determining the PSU$-$domain totals exactly (i.e., by census) after the first stage selection. The allocation formulas will differ because we need to take the first phase sampling probabilities of elementary units within selected PSUs into account. Given a selected PSU, let g_i denote the conditional probability of selection for any elementary unit in the first phase sample in PSU $i,$ assuming equal selection probabilities within each PSU. The phase one sample sizes within PSUs are not pre-determined, so the phase one sample should be as large as the schedule and budget allow to maximizing the frames for phase two sampling, especially in PSUs with higher numbers and concentrations of rarer domains.

In addition, up to now the $depsem$ designs have ignored nonresponse at the first phase. If response rates are equal across all PSUs, then the evidence suggests that response probabilities are approximately equal, as well, and can be ignored in specifying suitable sample allocations. However, if response rates vary considerably, then the observed response rate (r_i) in PSU i as an estimated first phase response propensity (assumed to be uniform for all units within PSU i) can be built into the allocation of sample units at the second phase. Building in the response propensity in sample allocation is equivalent to adjusting the phase one selection probabilities for nonresponse, and then a $depsem$ design can be constructed by suitably specifying the domain sampling rates. The design denoted $depsem-\mathrm{C1}'$ includes both sampling of phase one elementary units and phase one nonresponse; this variation was included at the suggestion of Eltinge (2011). Finally, we expand $depsem-\mathrm{C1}'$ to include stratification of PSUs resulting in $depsem-\mathrm{C2}'$ design.

$\boldsymbol{Depsem-\mathrm{C1}'}$ **Design:** The sample allocations $\{n_{id}^{C1'}\}_{1\le i\le m}$ of the domain total n_{+d}^{*} to selected PSUs, analogous to formula (3.1) of $depsem-\mathrm{B1},$ are given by

$$n_{id}^{C1'} = \left(\frac{\hat{f}_d^{\,C1'}}{\pi_i^{C1'}}\right) N'_{id}, \tag{5.1}$$

where N'_{id} is the size of domain d among phase one sample respondents within PSU $i,$ $\pi_i^{C1'} = \pi_i^{C1} g_i r_i,$ where the unconditional probability of selection for a unit to be in the phase one sample is now $\pi_i^{C1} g_i,$ modified from $depsem-\mathrm{C1}$ due to subsampling in the first phase, $\hat{f}_d^{\,C1'} = n_{+d}^{*} \big/ \hat{N}_{+d}^{C1'},$ and $\hat{N}_{+d}^{C1'} = \sum_{i=1}^{m} N'_{id} \big/ \pi_i^{C1'}.$ Notice that if $r_i = r;$ i.e., equal response rates across all PSUs, then it cancels out in equation (5.1) and has no impact on the sample allocation. Clearly, $\{n_{id}^{C1'}\}_{1\le i\le m}$ satisfies n_{+d}^{*} but does not satisfy n_{i+}^{*} (or n/m) even in expectation as in $depsem-\mathrm{B1}.$

$\boldsymbol{Depsem-\mathrm{C2}'}$ **Design:** The formula (5.1) can be generalized in a natural way to the stratified case similar to formula (3.1). This case is used in the application considered in the next section. In particular, the sample allocations $\left\{n_{hid}^{C2'}\right\}_{1\le i\le m_h, 1\le h\le H}$ of the domain total n_{++d}^{*} to selected PSUs within each stratum h, analogous to formula (3.3) of $depsem-\mathrm{B2}$, are given by

$$n_{hid}^{C2'} = \left(\frac{\hat{f}_d^{C2'}}{\pi_{hi}^{C2'}}\right) N'_{hid}, \qquad (5.2)$$

where N'_{hid} is the first phase respondent sample size for domain d within PSU i and stratum $h, \pi_{hi}^{C2'} = \pi_{hi}^{C2} g_{hi} r_{hi}$, where terms are defined in a natural way for stratified designs, $\hat{f}_d^{C2'} = n_{++d}^{*} \big/ \hat{N}_{++d}^{C2'}$, and $\hat{N}_{++d}^{C2'} = \sum_{h=1}^{H} \sum_{i=1}^{m_h} N'_{hid} \big/ \pi_{hi}^{C2'}$.

6 Application of $depsem-\mathrm{C2}'$ design to toolbox development

A team of university researchers developed a set of tests for behavioral and cognitive functions. They desired to "norm" the tests, establishing typical ranges of results for the general population, by measuring the results on children recruited to take the tests. Because the test results vary by age and gender, the goal was to recruit male and female children by year of age. Furthermore, the researchers wanted Spanish-speaking children as well as English-speaking children. The desired domain sample sizes n_{+d}^{*} of completes for twelve age/gender/language cells or domains are shown in Table 6.1.

Table 6.1
Desired completed tests $\left(n_{++d}^{*}\right)$ by demographic domain

	English-speaking		Spanish-speaking	
Age	**male**	**female**	**male**	**female**
3	200	200	200	200
4	200	200	200	200
5	200	200	200	200

Originally the researchers desired a probability sample representative of the U.S. population for each of these domains (as well as many additional age groups, which we omit here for simplicity). Once recruited, the sample children were required to be brought to a test site to take the tests in person. Therefore, an area probability design with a limited number of test sites was an efficient design of choice. NORC proposed to select a subsample of the PSU geographies in NORC's National Frame (Harter, Eckman, English and O'Muircheartaigh 2010). The National Frame is a multi-stage cluster sample of geographies, with housing unit addresses compiled for the smallest level of geography in the sample. The geographies are sampled and the address lists are compiled following the decennial census to support face-to-face interviews throughout the decade.

For norming the tests, 16 of the National Frame's 79 highest level geographies were selected as PSUs. The population of PSUs was stratified in the same way as the National Frame had been stratified, basically by metropolitan statistical area (MSA) status and size. The strata and sample sizes of PSUs are shown in Table 6.2. For the National Frame, stratum 1 MSAs had been selected with certainty. For the proposed design, the National Frame PSUs within strata were subsampled systematically with pps, where the measure of size was the number of Spanish-speaking households, because the cells for Spanish-speaking children would be the hardest to fill. Probabilities of selection for the PSUs were the product of the original National Frame probabilities and the subsampling probabilities. Some of the Stratum 1 PSUs were subsampled with certainty.

Table 6.2
Subsampling of PSUs from NORC's National Frame

Stratum h	Population #PSUs	National Frame #PSUs	Sample #PSUs (m_h)
1. Largest MSAs	24	24	12
2. Other MSAs	607	17	2
3. Non-MSA Counties	1,852	38	2
Total	2,483	79	16

Each PSU was to be divided into smaller geographical 'site areas'. Each site area would contain a testing site, and the site areas were to be approximately 10×10 miles in urban areas and 30×30 miles in rural areas to provide reasonable driving distances for children to be brought to a test site. Figures 6.1 and 6.2 illustrate the process of defining site areas. In Figure 6.1, a 10×10 mile grid is placed over the Chicago MSA. Then each census tract in the Chicago MSA is assigned to a grid cell based on the geographic location of the tract centroid. The resulting site areas are shown in Figure 6.2. One site area was to be selected per PSU, using systematic pps sampling where the measure of size was the number of Spanish-speaking households. Therefore, in subsequent notation, subscript i denotes both the PSU and the site area.

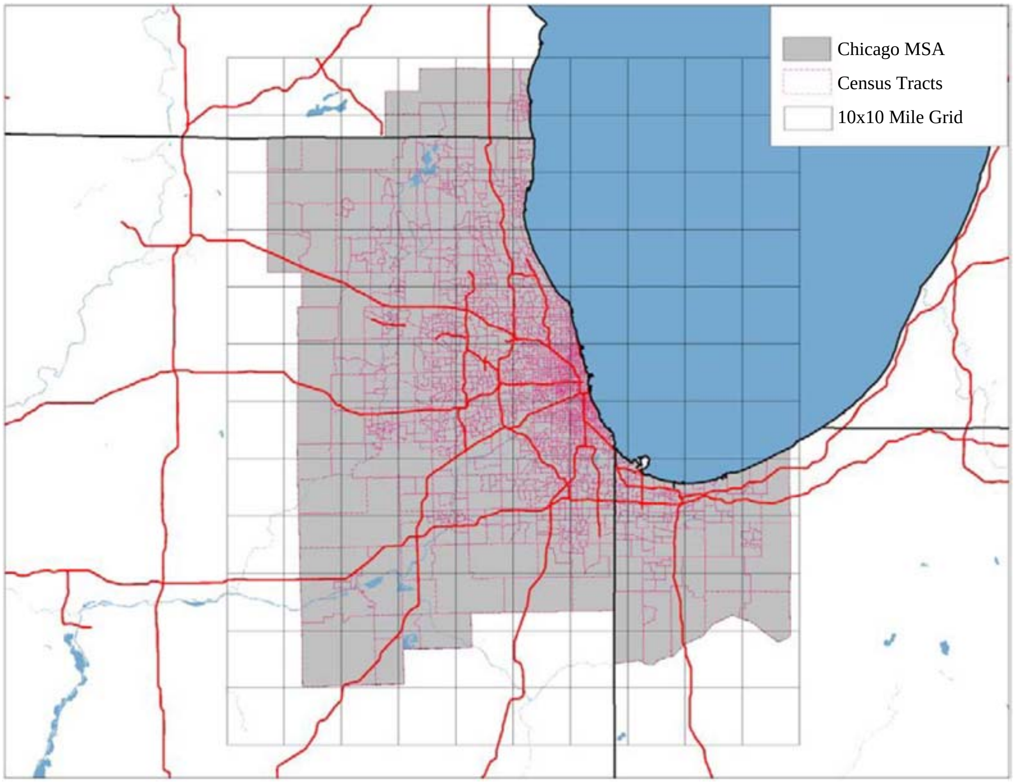

Figure 6.1 10×10 Mile grid over Chicago MSA.

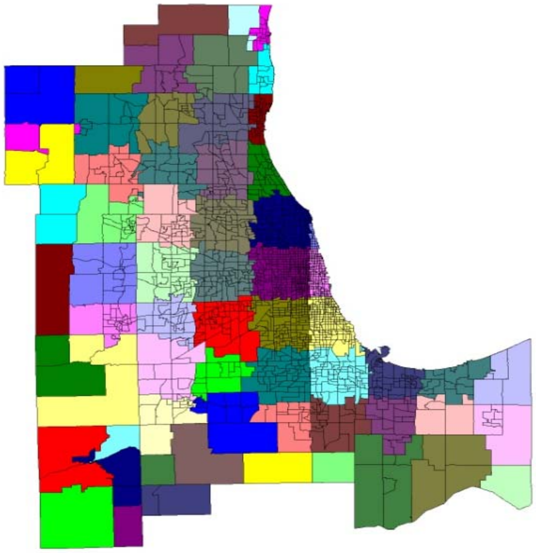

Figure 6.2 Site areas in Chicago MSA with census tracts assigned to grid cells.

We had no information on the number of English-speaking and Spanish-speaking male and female children by year of age for the selected site areas. This problem is best represented by the $depsem-\mathrm{C}2'$ design. Using an address-based sampling frame based on the U.S. Postal Service's Delivery Sequence File, we planned to select a large phase one sample of housing units for a mail screener to roster the households' children by gender, age, and language. The screener also would solicit telephone numbers for contacting parents to gain cooperation for testing the children. In this way we planned to obtain the phase one response rates r_{hi} and the phase two domain frame totals N'_{hid} for each site area i in stratum $h.$

With the phase one response rates, selection probabilities, and N'_{hid} frame totals in hand, and the specified sample sizes by domain, we were prepared to allocate the desired samples by domain and site area for the second phase of the study to conduct the behavioral and cognitive tests. We would recruit by telephone, with incentives for the sample participants to be brought to the test site.

Ultimately the sample design was never implemented, although we had subsampled the PSUs from the National Frame. Limitations in grant funding led the researchers to revert to convenience sampling near their network of cooperating universities. Nevertheless, the original plan for a probability sample allowed the original Folsom et al. (1987) result for $depsem$ samples to be generalized in a concrete way. For the sake of illustration, we continue the stratified two-phase two stage example with somewhat realistic but hypothetical probabilities and results.

Table 6.3 shows illustrative probabilities of selection for 16 test sites. These hypothetical unconditional site area (PSU) probabilities $\pi_{hi}^{C2'}$ reflect the initial National frame probabilities, the subsampling probabilities for PSUs, and the selection of one test site per PSU. The g_{hi} values are the conditional probabilities of selecting a phase one sample address in stratum $h.$ The product, then, is the unconditional probability of selecting a housing unit (HU) for phase one. Table 6.3 also shows hypothetical site-level response rates $r_{hi},$ leading to the probabilities $\pi_{hi}^{C2'}$ of an address being selected in phase one and the corresponding household responding and being available for phase two, if eligible.

Suppose that we mailed questionnaires to selected addresses in the site areas to collect household rosters and telephone numbers. Table 6.4 illustrates hypothetical expected counts N'_{hid} (shown as top entries in each cell) by stratum/site area by domain across all 16 test sites. These counts are not true population counts, but they define second phase frame or population counts for our phase two sampling and are based on first phase screener responses.

Table 6.3
Probabilities of phase 1 completion incorporating subsampling and nonresponse

Stratum/PSU (hi)	Unconditional Site Area Probability $\pi^{C2}_{hi} \times 10^6$	Conditional Phase One Sampling rate g_{hi}	Unconditional Phase One Probability $\pi^{C2}_{hi} g_{hi} \times 10^6$	Household Response Rate Per Site r_{hi}	Probability for Phase One Completion $\pi^{C2'}_{hi} \times 10^6$
(1,1)	1,239	0.60	743	0.40	297
(1,2)	972	1.00	972	0.50	486
(1,3)	3,408	0.60	2,045	0.30	613
(1,4)	3,561	0.60	2,137	0.50	1,068
(1,5)	1,985	0.60	1,191	0.40	476
(1,6)	2,083	0.60	1,250	0.40	500
(1,7)	3,142	0.60	1,885	0.60	1,131
(1,8)	5,058	0.60	3,035	0.50	1,517
(1,9)	3,001	0.60	1,801	0.60	1,080
(1,10)	1,621	0.60	973	0.40	389
(1,11)	1,081	0.60	648	0.30	194
(1,12)	533	1.00	533	0.50	266
(2,1)	686	1.00	686	0.40	274
(2,2)	77	1.00	77	0.40	31
(3,1)	328	1.00	328	0.60	197
(3,2)	2,555	0.60	1,533	0.50	766

Table 6.4
Eligible children (N'_{hid}) and sample allocation $(n^{C2'}_{hid})$ by stratum, site area, and domain

(Phase two sampling frame counts (top entry) with sample size (bottom entry)
E=English-speaking HU, S=Spanish-speaking HU, M=Male, F=Female, A3=Age 3, A4=Age 4, A5=Age 5)

	Stratum by Site Area (PSU); i.e., (h,i)							
Domain (d)	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)	(1,8)
(E,M,A3)	311	18	254	140	47	187	113	221
	18.8	0.7	7.4	2.4	1.8	6.7	1.8	2.6
(E,M,A4)	297	20	281	151	34	182	149	180
	18.4	0.8	8.4	2.6	1.3	6.7	2.4	2.2
(E,M,A5)	338	27	329	164	56	230	178	234
	19.0	0.9	9.0	2.6	2.0	7.7	2.6	2.6
(E,F,A3)	299	20	248	135	41	158	65	218
	19.3	0.8	7.8	2.4	1.7	6.1	1.1	2.8
(E,F,A4)	317	16	252	155	38	153	45	212
	19.8	0.6	7.6	2.7	1.5	5.7	0.7	2.6
(E,F,A5)	335	11	338	173	72	180	106	232
	19.7	0.4	9.7	2.8	2.6	6.3	1.6	2.7
(S,M, A3)	56	70	63	54	52	29	90	11
	35.0	26.8	19.1	9.4	20.3	10.8	14.8	1.3
(S,M,A4)	56	83	54	43	41	23	72	7
	34.9	31.6	16.3	7.4	15.9	8.5	11.8	0.9
(S,M,A5)	61	68	50	47	41	32	86	11
	37.9	25.8	15.0	8.1	15.9	11.8	14.0	1.3
(S,F,A3)	52	74	61	41	29	23	95	14
	33	28.7	18.7	7.2	11.5	8.7	15.8	1.7
(S,F,A4)	63	79	56	52	29	25	79	11
	41.2	31.6	17.7	9.5	11.8	9.7	13.6	1.4
(S,F,A5)	54	86	61	50	16	32	81	11
	33.9	33.0	18.6	8.7	6.3	12.0	13.4	1.4
Column Margins	2,239	572	2,047	1,205	496	1,254	1,159	1,362
	330.9	181.7	155.3	65.8	92.6	100.7	93.6	23.5

Table 6.4 (cont.)
Eligible children $\left(N'_{hid}\right)$ and sample allocation $\left(n_{hid}^{C2'}\right)$ by stratum, site area, and domain
(Phase two sampling frame counts with sample size underneath;
E=English-speaking HU, S=Spanish-speaking HU, M=Male, F=Female, A3=Age 3, A4=Age 4, A5=Age 5)

	Stratum by Site Area (PSU; i.e., (h,i))								**Row**
Domain (d)	**(1,9)**	**(1,10)**	**(1,11)**	**(1,12)**	**(2,1)**	**(2,2)**	**(3,1)**	**(3,2)**	**Margins**
(E,M,A3)	209	252	189	99	52	191	36	7	2,326
	3.5	11.6	17.5	6.7	3.4	111.7	3.3	0.2	200.1
(E,M,A4)	191	221	198	113	63	182	38	5	2,305
	3.3	10.5	18.8	7.8	4.2	109.0	3.6	0.1	200.1
(E,M,A5)	252	234	205	117	65	198	32	9	2,668
	3.9	10.1	17.6	7.4	4.0	107.8	2.7	0.2	200.1
(E,F,A3)	198	236	191	110	63	173	34	5	2,194
	3.5	11.7	18.9	7.9	4.4	108.2	3.3	0.1	200
(E,F,A4)	205	234	187	97	68	185	29	2	2,195
	3.5	11.2	17.9	6.8	4.6	111.9	2.7	0	199.8
(E,F,A5)	245	261	223	124	61	180	41	2	2,584
	4.0	11.8	20.1	8.2	3.9	102.5	3.6	0	199.9
(S,M, A3)	36	45	18	5	11	0	0	27	567
	6.2	21.5	17.2	3.5	7.5	0	0	6.6	200
(S,M,A4)	27	43	25	7	16	0	0	34	531
	4.6	20.5	23.8	4.9	10.8	0	0	8.2	200.1
(S,M,A5)	41	38	27	2	18	0	0	25	547
	7.0	18.0	25.6	1.4	12.1	0	0	6	199.9
(S,F,A3)	34	54	23	7	14	0	0	23	544
	5.9	26.2	22.3	5.0	9.6	0	0	5.7	200
(S,F,A4)	36	50	18	0	11	0	0	25	534
	6.5	25.0	18.0	0	7.8	0	0	6.3	200.1
(S,F,A5)	38	43	29	5	11	0	2	20	539
	6.6	20.6	27.8	3.5	7.5	0	1.9	4.9	200.1
Column Margins	1,512	1,711	1,333	686	453	1,109	212	184	17,534
	58.5	198.7	245.5	63.1	79.8	651.1	21.1	38.3	2,400.2

Using terms defined for equation (5.2), we computed estimated domain counts $\left(\hat{N}_{++d}^{C2'}\right)$ from the first phase sample as shown in Table 6.5. The desired initial domain sample sizes $\left(n_{++d}^{*}\right)$ in Table 6.1 divided by the $\hat{N}_{++d}^{C2'}$ values in Table 6.5 give us the estimated overall sampling rate $\hat{f}_{d}^{C2'}$ for each domain, also shown in Table 6.5.

Again using equation (5.2), we determined the allocations for each stratum, each site area, and each domain within each site area. The resulting allocations are also shown in Table 6.4 as bottom entries in each cell. The allocations are not integers, but random rounding can be used to preserve the *epsem* property in expectation while converting the allocations to integers, as discussed in Section 2. Alternatively, simple rounding will lead to an approximately *depsem* sample design.

Table 6.5
Estimated domain counts $\hat{N}_{++d}^{C2'}$ (in 000)
(Sampling rates $(\hat{f}_{d}^{C2'})$ for phase 2 underneath)

	English-Speaking		**Spanish-speaking**	
Age	**male**	**female**	**male**	**female**
3	11,122	10,399	1,075	1,061
	0.0178	0.0192	0.1860	0.1885
4	10,858	10,749	1,081	1,029
	0.0184	0.0186	0.1851	0.1944
5	11,948	11,415	1,084	1,071
	0.0167	0.0175	0.1845	0.1867

7 Summary and concluding remarks

In the design of any survey, there is a need for good representation of analysis domains in the sample. The sample allocation is not that simple because, unlike information about indicators for commonly used strata available in the sampling frame, domain indicators are generally not available or even if available, it may not be practical to stratify by domains due to interviewer travel costs for in-person surveys. What is needed is a method of sample allocation which allows for desired over-(under-) sampling of domains such that the resulting design is self-weighting or *epsem* for domains. Such designs are desirable for variance efficiency in general. In the case of one phase two stage designs, under certain assumptions, it is possible to allocate equal interviewer workload per selected PSU such that the sample size for all selected PSUs is controlled at the desired level, but domain sizes over all selected PSUs satisfy the desired level only in expectation. On the other hand, in the case of two phase two stage designs, it is possible to allocate domain sample sizes within PSUs such that the domain sizes over all PSUs are controlled at desired levels but the sample size per selected PSU is not controlled as the equal interviewer workload per PSU is not deemed important in this case. Although the *epsem* design of Kish at the population level is well known, domain level *epsem* (denoted *depsem* in this paper) designs are not well known among practitioners.

In this paper, we considered two main scenarios for *depsem* designs considered by Folsom et al. (1987). First, for two stage designs with known domain level PSU population counts (as well as known frame-level domain identifiers for elementary units) and pre-specified domain sample sizes, the PSU selection probabilities are defined such that the desired PSU sample size (equal per PSU) is allocated to domains within PSUs to obtain a *depsem* design. Second, for two phase two stage designs with known PSU selection probabilities and pre-specified domain sample sizes, the domain sampling rates are defined such that the desired domain sample size is allocated to PSUs within domains to obtain a *depsem* design. These two designs were referred to as $depsem-\mathrm{A1}$ and $\mathrm{B1}$ respectively. A simple justification of these two designs was provided. It is based on the key idea for obtaining *depsem* designs that the sampling rate (n_{id}/N_{id}) at the PSU by domain level should be made directly proportional to the domain level sampling rate (f_d) but inversely proportional to the PSU selection probability (π_i). For $depsem-\mathrm{A1},$ f_d is known but π_i is suitably defined (it was termed *composite measure of size* by Folsom et al. 1987), while for $depsem-\mathrm{B1},$ π_i is known but f_d is suitably defined. The corresponding stratified versions (denoted by $\mathrm{A2}/\mathrm{B2})$ can also be easily defined.

As a generalization of $depsem-\mathrm{A1}/\mathrm{B1}$ designs, $depsem-\mathrm{AB1}$ was proposed where domain-level PSU population counts are only approximately known for specifying PSU selection probabilities, but a two phase design is used to allocate desired domain sample sizes to PSUs after obtaining the true domain-level population counts for selected PSUs in the first phase. Also generalizations of $depsem-\mathrm{B1}$ were considered to obtain $depsem-\mathrm{C1}$ when PSU selection probabilities are pre-specified from other considerations. The $depsem-\mathrm{C1}'$ extends $depsem-\mathrm{C1}$ to cover certain practical realistic situations: 1) subsampling of elementary units within each selected PSU in the first phase to reduce cost, and 2) nonresponse in screening units for domain classification. The $depsem-\mathrm{C2}'.$ design allows for stratification in addition to practical features of $depsem-\mathrm{C1}'$ mentioned above. For all *depsem* designs except for $\mathrm{A1},$ PSU sample size is not directly controlled, but domain sample size is controlled via stratification of the first phase before the second phase. This is not a limitation in various practical applications where interviews are not conducted face-to-face.

The initial *depsem* design framework of Folsom et al. (1987) to allocate equal probability samples for multiple domains in two-stage designs in conjunction with one/two phase is a useful technique currently available in the SUDAAN software system (http://www.rti.org/page.cfm/SUDAAN) and employed successfully at RTI International for many years for studies such as the National Survey of Child and Adolescent Well-Being. The generalizations presented here extend the technique to the situation of multiple domains where the domain-level population counts need to be estimated for all selected PSUs, and where PSU selection probabilities are pre-specified from other considerations. These techniques are expected to be useful to sampling statisticians in a variety of situations.

Acknowledgements

The first author is grateful to Ralph Folsom of RTI International for several useful discussions. The authors thank Ned English of NORC at the University of Chicago for his help in creating site areas for the toolbox application. Thanks are also due to the associate editor and two referees for their very useful comments which helped improve clarity and presentation of the paper.

References

Cochran, W. (1977). *Sampling Techniques,* 3rd Ed. New York: John Wiley & Sons, Inc.

Eltinge, J. (2011). Personal Communication.

Fahimi, M., and Judkins, D. (1991). PSU Probabilities given differential sampling at second stage. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 538-543.

Folsom, R.E., Potter, F.J. and Williams, S.R. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 792-796.

Harter, R., Eckman, S., English, N. and O’Muircheartaigh, C. (2010). Applied sampling for large-scale multi-stage area probability designs. In *Handbook of Survey Research, Second Edition,* (Eds., P. Marsden and J. Wright), Emerald: United Kingdom.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Lohr, S. (2010). *Sampling: Design and Analysis,* 2nd Ed. Boston: Brooks/Cole.

Singh, A.C., and Harter, R.M. (2011). A generalized epsem two-phase design for domain estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3269-3282.