Comparing Alternative Estimation Methods when Using Multi-hit Approach to PSU Selection

Sadeq R Chowdhury
Agency for Healthcare Research and Quality[1]
5600 Fishers Lane, Rockville, MD 20857

Abstract

In a multi-stage sample design, primary sampling units (PSUs) are generally selected using systematic probability proportional to size (PPS) sampling without replacement. Certainty PSUs are usually identified up front iteratively and include those PSUs whose measure of size (MOS) exceeds the sampling interval at each iteration. Then all identified certainty PSUs and a sample of non-certainty PSUs are selected. Sometimes instead of certainty PSUs being identified up front, a multi-hit approach is used where a systematic sampling skip interval is applied through all PSUs. The large PSUs with MOS greater than the skip interval receive one or more hits and the remaining PSUs receive either one or zero hits. A cluster of ultimate sampling units is selected corresponding to each hit. The selection probability or base weight under the multi-hit approach can be calculated in two alternative ways. One approach is to treat all clusters selected with equal probability irrespective of which PSU the cluster was selected from without differentiating between certainty and noncertainty PSUs. The other approach is where large PSUs that definitely receive at least one hit are identified as a certainty PSUs and treated as explicit strata. In the latter case, cluster selection probability varies from one certainty PSU to another depending on the number of hits received. This paper discusses these two alternative methods of computing selection probabilities under the multi-hit approach to PSU/cluster selection and compares relative efficiencies of corresponding estimates.

Key Words: Multi-stage Sampling, PSU Selection, Certainty PSU, Multi-hit Sampling, Selection Probability

1. Introduction

In multi-stage sampling, ultimate sampling units (USUs) are selected in several stages by dividing the population into exhaustive sampling units at each stage where a sampling unit in each stage contains a cluster of USUs. The first-stage sampling units are called Primary Sampling Units (PSUs) within which one or more stages of secondary sampling units

[1] The views expressed in this paper are those of the authors and no official endorsement by the Department of Health and Human Services (DHHS) or the Agency for Healthcare Research and Quality (AHRQ) are intended or should be inferred.

(SSUs) are selected and then USUs are selected at the final stage. The overall probability of selection of an USU is obtained by multiplying probabilities of selection at all stages. For more information about probability sampling or multi-stage sampling methods, see Lohr (2010), Kish (1987), Kalton (1983), Cochran (1977), and Kish (1965).

In multi-stage sampling, PSUs and SSUs are usually selected systematically without replacement with probability proportional to size (PPS) where the measure of size (MOS) is usually the number of USUs. The large PSUs with MOS greater or equal to the sampling skip interval are selected with certainty and are called certainty or self-representing (SR) PSUs. The remaining PSUs are called noncertainty or non self-representing (NSR) PSUs and a sample of NSR PSUs are selected. Under the usual multi-stage sampling, certainty PSUs are explicitly identified upfront based on an iterative procedure that recalculates the sampling skip interval at each iteration. More details on usual procedure of PPS sampling can be found in (Williams and Chromy, 1980; Cochran, 1977; and Madow, 1949). Sometimes instead of identifying certainty PSUs up front, a multi-hit approach is used where a systematic sampling skip interval is applied through all PSUs to identify PSUs that receive one or more hits. The large PSUs with MOS greater than the skip interval receive at least one hit and can be treated as certainty PSUs while the PSUs with MOS smaller than the skip interval receive either zero or one hit and can be treated as noncertainty PSUs. A cluster of ultimate sampling units is selected from each hit location. Therefore, the number of clusters selected from a large PSU is equal to the number of hits it receives while only one cluster is selected from smaller PSUs with only one hit. For computing the selection probability or base weight under the multi-hit approach, two alternative methods can be used. Method A does not differentiate between certainty and non-certainty PSUs and all clusters receive equal weight. Method B differentiates between certainty and noncertainty PSUs and certainty PSUs are treated as explicit strata. Therefore, the cluster selection probability or the base weight varies across certainty PSUs. This paper compares relative efficiencies of estimates corresponding to these two methods used for multi-hit approach to PSU selection.

1.1 Background

The Medical Expenditure Panel Survey (MEPS) is an annual survey that has been conducted since 1996 by the Agency for Healthcare Research and Quality (AHRQ). It provides nationally representative estimates of health care use, expenditures, sources of payment, and health insurance coverage for the U.S. civilian non-institutionalized population. The MEPS Household Component (MEPS-HC) also provides data on respondents' health status, demographic and socio-economic characteristics, employment, access to care, and satisfaction with health care. Estimates can be produced for individuals, families, and selected population subgroups. Each new panel of sample households in the MEPS-HC is selected as a sub-sample of the responding households from the previous year's National Health Interview Survey (NHIS) conducted by the National Center for Health Statistics. Therefore, both surveys are based on the same design which was a multi-stage area sample design until 2016 (Chowdhury, et al. 2018).

In contrast to prior multi-stage designs, the 2016 redesign of NHIS (Moriarity and Parsons, 2015; Parsons, 2014) utilized USPS listings of addresses[2] in a PSU instead of traditional listings within selected segments. PSUs within a sampling stratum were selected using a multi-hit approach. First, the cluster locations in terms of PSUs were identified systematically starting with a random number and a sampling skip interval for each stratum. Then clusters of addresses from each identified PSU were selected directly from the PSU-wide listing of addresses. The number of clusters selected from a PSU was based on the number of hits a PSU received. A cluster within a PSU included a number of sub-clusters of addresses which were selected systematically from the PSU-wide list of addresses.

Selection probabilities for clusters of households were calculated under the multi-hit approach to PSU/cluster selection as if in a single-stage cluster sample design. The addresses (households) within a stratum were assigned equal selection probability irrespective of whether it is a certainty or a noncertainty PSU. Unlike in a traditional multi-stage design a certainty PSU was not treated as explicit stratum. An alternative argument was that since the sample was selected in two steps, first identifying PSUs and then selecting households within PSU, the design should be treated as two-stage and selection probabilities should be calculated accordingly. Basically the certainty PSUs should be treated as explicit stratum under that approach.

The two alternative methods of estimation under the multi-hit approach discussed and compared in this paper correspond to the two possible alternative methods discussed above for the 2016 NHIS redesign.

2. Usual Procedure of PSU Selection (Method 0) in Multi-stage sampling

PSUs and SSUs are usually selected systematically without replacement with probability proportional to size (PPS) with MOS generally the number of USUs or households or persons in a household survey. All USUs selected within a sampled PSU are treated as a cluster, also called an ultimate cluster. A cluster may include consecutive units or randomly selected sample of units selected either in a single stage or in multiple stages. In multi-stage area sampling, within a selected area PSU, segments or Census blocks are selected as second-stage units, households as the third stage and persons within selected households are selected as USUs. In multi-stage sampling, all units within each PSU i.e., each ultimate cluster is treated as a single unit for variance estimation. The probability of selection of an USU is obtained by multiplying probabilities of selection at all stages. A design usually targets to achieve equal probability of selection (EPS) of all units in the target population for efficiency of estimation.

[2] Except in some rural areas where traditional listing was needed

A large PSU i whose MOS (M_i) is greater than or equal to the sampling interval (SI) i.e., $M_i \geq SI$ is selected with certainty, where $SI = M_0/n$ and $M_0 = \sum_{i=1}^{N} M_i$ with M_i the MOS of PSU i and n the number of PSUs to be selected. The PSUs with MOS < SI are sampled with probability < 1.0 and are called non-certainty PSUs. Certainty PSUs are usually identified iteratively where at each iteration previously identified certainty PSUs are removed and additional certainty PSUs (if any) are identified with recalculated sampling interval. At iteration 1, PSUs with $M_i \geq (SI_1 = M_0/n)$ are identified as certainty then at iteration 2, sampling skip interval is recalculated as $SI_2 = (M_0 - \sum_{i \in c1} M_i)/(n - n_{c1})$, where c1 is the set of PSUs and n_{c1} is the number of PSUs identified as certainty at iteration 1, and PSUs with $M_i \geq SI_2$ are identified as certainty. The iteration process is continued until no other PSU has MOS greater than the recalculated sampling skip interval. Once all certainty PSUs are identified through this iteration process, a sample of noncertainty PSUs are selected from the remaining PSUs to represent the rest of the population. The skip interval used to select the NSR PSUs is

$$SI_{nc} = \frac{(M_0 - \sum_{i \in c} M_i) = M_{nc}}{(n - n_c) = n_{nc}},$$

where c is the set of n_c PSUs identified as certainty.

The first-stage selection probability of a certainty PSU is 1.0, while the selection probability of a noncertainty PSU i is

$$P_i = \frac{(n - n_c)M_i}{(M_0 - \sum_{i \in c} M_i)}.$$

Each certainty PSU is treated as a separate stratum and a sampling rate within the PSU is applied to ensure equal selection probability of selecting USUs across all PSUs.

The second stage selection probability of jth unit in the ith PSU is

$$P_{ij} = \frac{k\,\bar{m}}{M_i} \qquad \text{if SR PSU and}$$

$$P_{ij} = \frac{\bar{m}}{M_i} \qquad \text{if NSR PSU,}$$

where $\bar{m}$=cluster size and k can be non-integer too. Generally, cluster sizes ($\bar{m}$) are equal in NSR PSUs and cluster sizes vary ($k\bar{m}$) depending on the size of a SR PSU. The overall selection probability is,

$$P_{ij} = P_i P_{i/j} = \frac{(n-n_c)\,M_i}{(M_0 - \sum_{i \in c} M_i)} \frac{\bar{m}}{M_i} = \frac{(n-n_c)\bar{m}}{(M_0 - \sum_{i \in c} M_i)} \qquad \text{in NSR PSUs}$$

and

$$P_{ij} = 1\text{x}\frac{k\,\bar{m}}{M_i} \qquad \text{in SR PSUs}$$

In this paper, the usual procedure of identifying certainty PSUs and selecting non certainties as described above is referred to as Method 0.

3. Multi-Hit Approach to PSU Selection

A multi-hit approach is an alternative method of selecting PSUs or clusters. Under this approach, certainty PSUs are not identified up front. A systematic sampling skip interval (SI=M_0/n) is calculated only once. This skip interval is applied through all PSUs and the PSUs with MOS ≥ SI receive at least one hit. The PSUs with MOS <SI receive either zero

or one hit based on the random process. The number of clusters selected from a PSU is equal to the number of hits a PSU receives. Under the multi-hit approach, the cluster size is usually equal in all PSUs.

3.1 Multi-Hit Selection Probability – Method A

Under this approach, the size of a certainty PSU or whether it is a certainty or noncertainty has no impact on the selection probability. All clusters represent the same number of population units and hence all sampling units have the same selection probability irrespective of which PSU a unit belongs to. PSUs are used like implicit strata with implicit boundaries and there is no designation of any certainty or noncertainty PSU. All clusters are of equal size and selected with equal probability as if in a single stage selection of equal size clusters. A cluster or hit represents the population covered by the skip interval, i.e., sampling weight = $SI/\bar{m}$ with probability of selection, $P_{ij} = \bar{m}/SI = n\bar{m}/M_0$ and $SI = M_0/n$. A cluster can represent a whole or part of a PSU or more than one PSU depending on SI and M_i. For example, if $M_i = 1.25\ SI$ for PSU *i* then it can have 1 or 2 hits or clusters. If 2 clusters are selected, these clusters will represent the whole current PSU ($1.25\ SI$ units) plus $.75\ SI$ units from the next PSU. If only 1 cluster is selected, it will represent 1.0 SI units from the current PSU and the cluster selected from the next PSU will represent the remaining $.25\ SI$ units from this PSU and $.75\ SI$ units from the next PSU.

3.2 Multi-Hit Selection Probability – Method B

Under this approach, certainty and noncertainty PSUs are designated. PSUs with $MOS \geq SI$ receive at least one hit and are treated as certainty with selection prob=1.0. A certainty PSU is treated like an explicit separate stratum and the selection probability of an ultimate sampling unit depends on the size of the PSU and the number of clusters selected from a certainty PSU. The selection probability is calculated separately in each certainty PSU and can vary from PSU to PSU while in all noncertainty PSUs the selection probability is the same. The selection probability in a certainty PSU with k hits is

$$P_{ij} = 1\text{x}\frac{k\bar{m}}{M_i},$$

while in all non-certainty PSUs, the selection probability is the same as

$$P_{ij} = \frac{nM_i}{M_0}\frac{\bar{m}}{M_i} = \frac{\bar{m}}{SI}.$$

For example, if the size of a PSU is 1.25 times of SI (i.e., $M_i = 1.25\ SI$) then the PSU can have either one or two hits. Then the selection probability will depend on the number of hits the PSU receives as follows.

$$P_{ij} = \bar{m}/M_i = \bar{m}/1.25\,M_i \quad \text{if one hit or}$$
$$P_{ij} = 2\bar{m}/M_i = 2\,\bar{m}/1.25\,M_i \quad \text{if two hits}$$

If the size of a PSU is 2.5 times of SI (i.e., $M_i = 2.5\ SI$) then the PSU can have either two or three hits. Then the selection probability depends on whether the PSU receives two or three hits as follows:

$$P_{ij} = 2\,\bar{m}/M_i = 2\,\bar{m}/2.5\,M_i \text{ if two hits or}$$
$$P_{ij} = 3\bar{m}/M_i = 3\,\bar{m}/2.5\,M_i \text{ if three hits.}$$

In the above examples, selection probabilities are random because whether the PSU receives (one or two) or (two or three) hits will be random. In contrast, there will be only one hit in a selected noncertainty PSU and the selection probability will be the same in all noncertainty PSUs as follows,

$$P_{ij} = \frac{\bar{m}}{SI} \text{ or } P_{ij} = \frac{nM_i}{M_0}\frac{\bar{m}}{M_i} = \frac{\bar{m}}{SI}$$

Under this approach, selection probabilities in noncertainty PSUs can also be calculated by deriving the skip interval after deducting the MOS of all certainty PSUs as follows

$$P_{ij} = \frac{(n-n_c)M_i}{(M_0 - \sum_{i\in c} M_i)}\frac{\bar{m}}{M_i} = \frac{\bar{m}}{(M_0 - \sum_{i\in c} M_i)/(n-n_c)} = \frac{\bar{m}}{SI}$$

Therefore, the sampling skip interval for noncertainty PSUs can be calculated with or without deducting the MOS of the certainty PSUs. In both cases, the cluster selection probabilities will be the same as shown above.

3.3 Similarities and Differences - Method 0, Method A and Method B

While selection probabilities are the same ($\bar{m}/SI$) in all non-certainty PSUs under both Methods A and B, they differ for certainty PSUs. In the certainty PSUs, the selection probability is random under Method B but if an expectation over these randomness is taken then it becomes equal to the selection probability under Method A. For example, if there is a PSU with size, $M_i = 1.25\,SI$ then under Method B the PSU can have either 1 or 2 hits and the selection probability will be one of the following:

$P_{ij} = \bar{m}/M_i = \bar{m}/1.25\,SI$ if one hit, probability of occurrence =0.75

$P_{ij} = 2\bar{m}/M_i = 2\,\bar{m}/1.25\,SI$ if two hits, probability of occurrence =0.25

Selection probabilities are random here because the number of hits the PSU receives is random with probabilities shown above. If an expectation over this randomness is taken then the selection probability will be

$$P_{ij} = .75\frac{\bar{m}}{1.25SI} + .25\frac{2\bar{m}}{1.25SI} = \frac{1.25\bar{m}}{1.25SI} = \frac{\bar{m}}{SI}$$

which is equal to the selection probability under Method A.

Table 1 summarizes the similarities and differences among the three methods of PSU/cluster selection (Method 0, Method A and Method B). Method 0 and Method B select certainty PSUs with probability 1 and treat each of them as an explicit separate stratum. Certainty PSUs are identified iteratively under Method 0 but at a single round of selection under Method B. Method A does not create explicit strata and treats certainty PSUs in the same manner as noncertainty PSUs. Under Method 0, noncertainty PSUs are selected by using a different skip interval after removing the certainty PSUs but under Method B, noncertainty PSUs are also selected concurrently with the certainty PSUs using the same skip interval. Method A uses the same skip interval as Method B. Therefore, the probability of selection of a PSU/cluster is the same under both Method A and B.

Table 1. Summary of Method 0, Method A and Method B

PSU Type	Method	PSU Selection Probability	Comment
Certainty (SR)	Method 0	1	Explicit Stratum, Identifies SR PSUs iteratively
	Method A	N/A	Implicit Stratum, No Identification of SR/NSR
	Method B	1	Explicit Stratum, Identifies SR PSUs at a single pass
Noncertainty (NSR)	Method 0	$\frac{n_{nc}M_i}{M_{nc}=\sum_{i\in nc}M_i}$	n_{nc} NSR PSUs
	Method A	$\frac{nM_i}{M_0=\sum_i M_i}$	n PSUs (no SR or NSR PSUs)
	Method B	$\frac{nM_i}{M_0=\sum_i M_i}$	n_{nc} NSR PSUs

4. Empirical Example

Table 2 presents an example that shows how a multi-hit selection procedure works and the corresponding selection probabilities under Method A and Method B discussed in Section 3. The table includes a population of four PSUs with total MOS of 60,000 households. If the target is to select three clusters systematically then the skip interval to select cluster locations or hits would be 20,000. At each hit location a cluster of 100 households will be selected, which can be consecutively located or dispersed around the hit location. The first column shows the relative size of each PSU in terms of MOS, the second column shows the coverage of each skip interval, the third column shows the probability of having hits in each PSU, the fourth column shows how the selection probability will be calculated under Method A and the remaining columns show how selection probabilities will be calculated under Method B. Since the MOS of PSU 1 is greater than the skip interval, it will receive at least one hit with probability 1.0 and the probability of receiving the second hit is 0.25. Since the size of PSU 2 is half the skip interval, it will have a probability of 0.5 to receive the second hit and PSU 3 will have a probability of 0.25 of receiving the second hit. Therefore, the second hit can be in one of the three sub-cells: A in PSU 1, B in PSU 2 and C in PSU 3 as shown in column 2. The probabilities of receiving the third hit in PSUs 3 and 4 are also shown in column 3. Under Method A, since no PSU is treated as a certainty PSU, irrespective of which PSU a household belongs to, it will have a probability of .005 since a cluster of 100 households are selected per 20,000 households. For example, the second cluster may come from any of PSUs 1 or 2 or 3 (sub-cells A, B, C as shown in columns 2 and 4) but the probability of selecting a household from sub-cells A or B or C is .005 as shown in Table 2a. Under Method B, since the first PSU will be treated as a certainty PSU, the combined probability of selecting a household will be different (.004 or .008) depending on whether the PSU receives one or two hits. The selection probability in all noncertainty PSUs will be the same as under Method A.

Table 2. Example of a multi-hit selection of three clusters from four PSUs

[1]	[2]	[3]	[4]	[5]	[6]	[7]
PSU	Skip= 20,000	Hits or Clusters	Method A	Method B		
				1st Stage	2nd Stage	Combined
PSU 1 (MOS =25,000)	Skip 1	1st Hit Prob 1.0	100/20K =.005	1.0	100/25K=.004 or 200/25K=.008	.004 or .008
	A	2nd Hit Prob .25	A			
PSU 2 (MOS =10,000)	B Skip 2	2nd Hit Prob .50	B .005	(3x10K/ 60K) = 0.5	(100/10K) =.01	.005
PSU 3 (MOS =15,000)	C	2nd Hit Prob .25	C			
	Skip 3	3rd Hit Prob .50	.005	(3x15K/ 60K) = 0.75	100/15K =.0067	.005
PSU 4 MOS =10,000		3rd Hit Prob .50	.005	(3x10K/ 60K) = .5	(100/10K) =.01	.005

Table 2a. Probability of selection in different PSUs under Skip 2

Sub-cell A	Sub-cell B	Sub-cell C
$\frac{5,000}{20,000} \times \frac{100}{5,000} = .005$	$\frac{10,000}{20,000} \times \frac{100}{10,000} = .005$	$\frac{5,000}{20,000} \times \frac{100}{5,000} = .005$

Using the same example, Table 3 shows how the PSUs will be selected and the selection probabilities will be calculated under the usual method of PSU selection (Method 0) in multi-stage sampling. Since the certainty PSU will be determined upfront, PSU 1 will always be a certainty and it will also be predetermined how many clusters will be selected from PSU 1 and how many from the remaining PSUs (noncertainty). In practice, the cluster size will be adjusted for a certainty PSU so that the overall selection probability remains equal across all PSUs. Here for the comparison with the multi-hit approach, since the cluster size will not be adjusted, either one or two clusters will be selected from PSU 1 (the certainty PSU). So there will be two scenarios and Table 3 shows the selection probabilities under both. Under Scenario 1, one cluster will be selected from PSU 1 and two clusters will be selected by selecting two PSUs at the first stage from the remaining PSUs. Under Scenario 2, two clusters will be selected from PSU 1 and one cluster will be selected by selecting a PSU at the first stage from among the remaining PSUs. Sampling skip intervals for selecting 1 or 2 PSUs from noncertainty PSUs will be calculated separately after setting aside the certainty PSU. At the second stage, a random sample of 100 households can be selected and treated as a cluster. Under this usual method, one of these two scenarios will be predetermined and the variance of an estimate will be calculated for repeated drawing of sample under that scenario. Under the multi-hit approach, one of these scenarios will be used in each draw randomly depending on how many hits the certainty PSU receives.

Table 3. PSU selection procedure and probabilities under Method O (Non Multi-hit)

PSU	MOS	Selection Probabilities					
		Scenario 1: 1 SR and 2 NSR			Scenario 2: 2 SR and 1 NSR		
		1st Stage	2nd Stage	1^{st}-2^{nd} Combined	1st Stage	2nd Stage	1^{st}-2^{nd} Combined
1	25,000	1	0.0040	0.0040	1	0.0080	0.0080
2	10,000	0.571	0.0100	0.0057	0.2857	0.0100	0.0029
3	15,000	0.857	0.0067	0.0057	0.4286	0.0067	0.0029
4	10,000	0.571	0.0100	0.0057	0.2857	0.0100	0.0029
Total	60,000	(60,000-25,000) = 35,000, SI=35,000/2=17,500			(60,000-25,000) = 35,000, SI=35,000/1=35,000		

Table 4 shows how the probabilities will be calculated under these two scenarios for multi-hit Method B which treats certainty and noncertainty PSUs differently. Since the sampling skip interval is calculated only once under the multi-hit approach, the first stage selection probability of selecting PSUs will remain the same for the scenarios. However, the second stage selection probability will vary across scenarios depending on how many hits the certainty PSU receives. Since a scenario will be selected randomly from these two scenarios, Scenario 1 will be realized 75% of the time and Scenario 2 will be realized 25% of the time. The last column in the table shows the expectation of the combined selection probability over the two scenarios, which is equal to the selection probability under Method A where selection probability is equal across all PSUs. Under the usual method (Method 0), the cluster size will be varied in a manner that ensures equal probability of selection across all PSUs.

Table 4. PSU selection procedure and selection probabilities under multi-hit Method B

PSU	MOS	Selection Probabilities						
		Method B						
		Scenario 1: 1 SR and 2 NSR			Scenario 2: 2 SR and 1 NSR			Expectation over two Scenarios (.75S1+.25S2)
		1st stage	2nd stage	1^{st}-2^{nd} Combined	1st stage	2nd stage	1^{st}-2^{nd} Combined	
1	25,000	1	0.0040	0.0040	1	0.0080	0.0080	0.005
2	10,000	0.500	0.0100	0.0050	0.5000	0.0100	0.0050	0.005
3	15,000	0.750	0.0067	0.0050	0.7500	0.0067	0.0050	0.005
4	10,000	0.500	0.0100	0.0050	0.5000	0.0100	0.0050	0.005
Total	60,000	SI = 60,000/3=20,00			SI = 60,000/3=20,000			

To summarize, both Method 0 and Method B identifies certainty and noncertainty PSUs. However, a scenario is predetermined under Method 0 and the selection probability is fixed, while which scenario is used under Method B depends on the random draw i.e., the random staring point of the skip interval. Therefore, the selection probability under Method

B is random and if an expectation over this randomness is taken then the selection probability becomes equal to that under Method A.

5. Comparison of Multi-hit Methods A & B

Using the same example above, properties of an estimate produced under Method A and Method B will be compared in this section. If the target is to produce the estimate of the total number of households from each sample then the properties of the estimators can be compared since the total number of households in these four PSUs is known. Under Method A, the base weight for producing estimates is the same for all records because selection probabilities are equal. The form of the estimator is $\widehat{N} = \sum_{i=1}^{n} \sum_{j=1}^{m} W_{ij}$ and the estimate from a PSU is $\widehat{N}_i = \sum_{j}^{m} W_{ij}$. If the cluster size is 100 and the weight is 1/.005 then the estimate from the PSU will be 20,000 and if two clusters are selected from a PSU then the estimate will be 40,000.

However, the base weight and selection probability will be different in certainty PSUs than noncertainty PSUs. Tables 5 and 6 show the estimates obtained from each PSU (rows) under Method A and Method B and all possible combinations of PSUs that can be selected if three clusters are selected from these four PSUs (columns). Table 5 shows that under Method A the estimate of the total number of households is 60,000 from each of all possible combinations of PSU samples. That means the estimate is unbiased and the standard error (SE) or relative standard error (RSE) of the estimate is zero.

Table 5. All possible samples and distribution of estimates under Method A

	PSU Combinations			
PSU	1, 2, 3	1, 2, 4	1, 3, 4	1, 1, 3
1	20,000	20,000	20,000	40,000
2	20,000	20,000	-	-
3	20,000	-	20,000	20,000
4	-	20,000	20,000	-
Estimate	60,000	60,000	60,000	60,000
Probability	0.25	0.25	0.25	0.25
Distribution of Estimates	Mean of estimates = 60,000 SE of estimates = 0, RSE of Estimates = 0%			

Table 6 shows the same for Method B where the estimate of the total number of households is 65,000 from three of the four possible combinations of PSU samples and 45,000 from the last combination. If we take the expected value of these four estimates then it becomes 60,000 implying that the estimate is unbiased. The expected value is calculated as $E(x) = \sum x\, p(x)$ where $p(x)$ is the probability of realizing each sample as shown in the second to last row of the table. However, since $p(x) = 0.25$ is the same for each sample, the expected value is equal to the simple mean. While the expected value of the estimate is the same under both methods, the SE under Method B is 7,500 and RSE=13%.

Table 6. All possible samples and distribution of estimates under Method B

	PSU Combinations			
PSU	1, 2, 3	1, 2, 4	1, 3, 4	1, 1, 3
1	25,000	25,000	25,000	25,000
2	20,000	20,000	-	-
3	20,000	-	20,000	20,000
4	-	20,000	20,000	-
Estimate	65,000	65,000	65,000	45,000
Probability	0.25	0.25	0.25	0.25
Distribution of Estimates	Mean of estimates = 60,000 SE of estimates = 7,500, RSE of estimates = 13%			

Comparing Tables 5 and 6 shows that the estimates are unbiased under both methods but Method B is less efficient than Method A. The relative properties of these two estimators also holds in similar examples. The reason for the inefficiency of Method B compared to Method A is that Method B treats the certainty PSU as a separate stratum and varies the selection probabilities, and hence base weights across PSUs, while the selection probability is the same across all PSUs under Method A.

5.1 Comparison when Actual MOS Different than Sampling MOS

Sometimes the MOS used in PSU selection or identifying the hit location is different than the actual MOS found in the field. For example, the census counts of households used in selecting the PSU may be different than counts found when the listing of households is done in the field. When such a discrepancy exists between the design and actual MOS, the selection probability should be adjusted. Otherwise, estimates will be biased and the extent of bias will inversely depend on the correlation between the two MOSs within a sampling stratum. If the two MOSs are perfectly correlated then there will not be any impact on the selection probability and there will be no bias.

In this section, we compare the methods using the same example above but introducing some discrepancies between the design and actual MOSs. In such a situation, the first stage selection is already done and no adjustment to selection probability is required. Only the second-stage selection probability is adjusted based on the actual counts. Table 7 shows the recalculated second-stage selection probabilities and the corresponding recalculated combined selection probabilities for both scenarios under Method B and the expected selection probabilities over these two scenarios, which is Method A.

Corresponding to probabilities recalculated in Table 7, Table 8 presents the estimates under Method A for all possible samples (i.e., all possible PSU combinations). The estimates vary from 52,000 to 66,000 but the expected value is 60,000 implying the estimate is unbiased. The SE of the estimates is 5,099 i.e., RSE=8.5%. Table 9 presents the same distribution of estimates under Method B. The estimates vary from 40,000 to 70,000 but the expected

value is 60,000 and the SE of the estimates is 10,000 i.e., RSE=16.7%. Therefore, this comparison shows that, even with discrepancy between design and actual MOSs, estimates under both methods are unbiased but Method B is less efficient than Method A.

Table 7. Adjusted second probabilities for discrepancy between design and actual MOSs

PSU	Design MOS	Actual MOS	Selection Probabilities: Method B: Scenario 1: 1 SR and 2 NSR: 1st Stage	2nd Stage	1st-2nd Combined	Scenario 2: 2 SR and 1 NSR: 1st Stage	2nd Stage	1st-2nd Combined	Method A: Expectation over two Scenarios (.75S1+.25S2)
1	25,000	20,000	1.000	0.0050	0.0050	1.000	0.0100	0.0100	0.006
2	10,000	12,000	0.500	0.0083	0.0042	0.5000	0.0100	0.0050	0.004
3	15,000	15,000	0.750	0.0067	0.0050	0.7500	0.0083	0.0063	0.005
4	10,000	13,000	0.500	0.0077	0.0038	0.5000	0.0100	0.0050	0.004
Total	60,000	60,000	Probability of Scenario 1 =75%			Probability of Scenario 2 =25%			

Table 8. All possible samples and distribution of estimates under Method A

PSU	PSU Combinations: 1, 2, 3	1, 2, 4	1, 3, 4	1, 1, 3
1	16,000	16,000	16,000	32,000
2	24,000	24,000	-	-
3	20,000	-	20,000	20,000
4	-	26,000	26,000	-
Estimate	60,000	66,000	62,000	52,000
Probability	0.25	0.25	0.25	0.25
Distribution of Estimates	Mean of estimates = 60,000 SE of estimates = 5,099, RSE of estimates = 8.5%			

Table 9. All possible samples and distribution of estimates under Method B

PSU	PSU Combinations: 1, 2, 3	1, 2, 4	1, 3, 4	1, 1, 3
1	20,000	20,000	20,000	20,000
2	24,000	24,000	-	-
3	20,000	-	20,000	20,000
4	-	26,000	26,000	-
Estimate	64,000	70,000	66,000	40,000
Probability	0.25	0.25	0.25	0.25
Distribution of Estimates	Mean of estimates = 60,000 SE of estimates = 10,000, RSE of estimates = 16.7%			

6. Summary and Conclusion

In a multi-stage sample design, a multi-hit approach to PSU selection is occasionally used where a systematic sampling skip interval is applied through all PSUs to identify certainty and noncertainty PSUs in one pass. This approach differs from the usual process where certainty PSUs are identified iteratively. A cluster of ultimate sampling units is selected from each non-certainty PSU that receives a hit while the number of clusters selected from each certainty PSU is equal to the number of hits a certainty PSU receives. Two alternative methods of computing selection probabilities under the multi-hit approach are discussed and the corresponding estimators are compared.

Method A is similar to a single-stage selection of clusters. All clusters have equal probability of selection and there is no separate stratum for a certainty PSU and no distinction between certainty and noncertainty PSUs. The overall selection probability is the same across all PSUs within a sampling stratum. In contrast, Method B distinguishes between certainty and noncertainty PSUs and each certainty PSU is treated as an explicit separate stratum. Consequently, the selection probability in a certainty PSU becomes different than other certainty PSUs or noncertainty PSUs. The overall selection probability is the same across all noncertainty PSUs. Method B is more like a two stage sample design but the number of clusters to be selected from a certainty PSU depends on the random draw, which depends on the starting random number within a stratum. The selection probability under Method A is equal to the expected selection probabilities over all possible random scenarios under Method B.

The comparison between the two multi-hit methods shows that both methods produce unbiased estimates but Method B is less efficient (i.e., higher variance of estimates) than Method A. Method B is less efficient because it ignores the variation in selection probabilities across all selection possibilities. Instead, the selection probabilities are calculated based on a realized sample which is random. In other words, it does not take the randomness into consideration while Method A takes an expectation over all random possibilities. The selection probability in a certainty PSU under Method B depends on the number of random hits the PSU receives, which causes selection probabilities to vary among certainty PSUs and hence increases variation in sampling weights. It essentially converts the design into a two-stage design after selecting the sample where the numbers of cases to be selected from certainty PSUs are not predetermined but depend on random draws. In a usual two-stage design, the number of cases to be selected from a PSU is predetermined and typically proportional to the size of a certainty PSU. The usual two-stage design can also be implemented under the multi-hit approach without losing any efficiency if the cluster sizes are varied proportional to the size of each certainty PSU.

Acknowledgement

The author would like to acknowledge Steve Machlin, Director, Division of Statistical Research and Methods, AHRQ, for his encouragement for giving the talk at JSM and his contributions in reviewing the paper.

References

Chowdhury, S.R., Machlin, S.R., Gwet, K.L. (2019). *Sample Designs of the Medical Expenditure Panel Survey Household Component, 1996–2006 and 2007–2016*. Methodology Report #33. January 2019. Agency for Healthcare Research and Quality, Rockville, MD. https://meps.ahrq.gov/data_files/publications/mr33/mr33.shtml

Cochran, W.G. (1977). *Sampling Techniques*. 3rd Edition, John Wiley & Sons, New York.

Kalton, G. (1983). *Introduction to Survey Sampling. Vol. 07-035 of University Paper Series on Quantitative Applications in the Social Sciences*. Beverly Hills, CA: Sage Publications.

Kish, L. (1987). *Statistical Design for Research*. New York: John Wiley & Sons.

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

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

Madow, W. G. (1949). "On the Theory of Systematic Sampling, II." *Annals of Mathematical Statistics* 20:333–354.

Moriarity, C. and Parsons, V. (2015). 2016 Sample Redesign of the National Health Interview Survey. *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association (CD-ROM).

Parsons, V. (2014). Designing Flexibility for State Samples into the 2016 NHIS. *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association (CD-ROM).

Williams, R. L., and Chromy, J. R. (1980). "SAS Sample Selection Macros." In *Proceedings of the Fifth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc. http://www.sascommunity.org/sugi/SUGI80/Sugi-80-71%20Williams%20Chromy.pdf.