

**Modified Sieve Sampling: A Method for Single- and Multi-Stage
Probability-Proportional-to-Size Sampling**

Lucas A. Hoogduin*
Senior Manager, Statistician
KPMG Global Services Centre
300 Tice Boulevard
Woodcliff Lake, New Jersey 07677
201-505-2227
lhoogduin@kpmg.com

Thomas W. Hall
Public Accounting Professor
Department of Accounting
P. O. Box 19468
University of Texas at Arlington
Arlington, Texas 76019
817-272-3087
tom.hall@uta.edu

* - Corresponding author.

Running title: Modified Sieve Sampling

Acknowledgement: The authors express their appreciation for the helpful comments of Jane Horgan.

Modified Sieve Sampling: A Method for Single- and Multi-Stage Probability-Proportional-to-Size Sampling

Abstract

Widely used probability-proportional-to-size (PPS) selection methods exhibit two important limitations: (1) a line item sample cannot generally be augmented via PPS selection while maintaining PPS properties and (2) the line item sample size exhibits randomness. Each method also exhibits one or more important method-specific limitations. This paper presents a new method of PPS selection, a modified version of sieve sampling which overcomes the limitations of methods in current use. Simulations indicate that modified sieve sampling maintains PPS properties in single- and multi-stage samples and yields stable line item sample sizes. In single-stage applications the method provides reliable control of sampling risk regardless of how errors are clustered in the population. Also, the method yields tightness and efficiency measures which are comparable to systematic sampling.

Key Words: Monetary-unit sampling, Probability-proportional-to-size, Substantive testing, Accounting populations, Audit sampling

1. INTRODUCTION

Monetary-unit sampling (MUS) is the principal statistical method used by external auditors when collecting evidence regarding errors in financial account balances (Jones, 1999, p. 51; Hall et. al, 2002, p. 129). MUS requires that population line items be chosen using probability-proportional-to-size (PPS) selection. To accomplish this sampling units are defined as individual monetary units (e.g. dollars, euros, etc.), and these units are selected using some form of random dollar selection. Once a monetary unit is selected, its parent line item is identified. The parent line item might be an individual transaction or an outstanding invoice or balance with a debtor. The selected line items are tested, the error amount in each line item is ascertained, and an upper bound on total population error is projected. If this bound is less than, or equal to, a preset level of maximum acceptable error the financial account balance is accepted as materially correct.

In practice, a variety of PPS selection methods have found acceptance and are available in commercial audit software (ACL Services Ltd 2006; CaseWare IDEA Inc 2004). These methods include unrestricted random dollar selection, cell selection, and systematic (fixed-interval) selection (for descriptions see Leslie et al. 1979). Despite their acceptance, these methods exhibit a number of important limitations. For example, during the conduct of an examination it is not unusual for circumstances to require that an initial (single-stage) MUS sample be augmented with a second-stage sample (AICPA 1999, p. 73; IFAC 2006, §530.55). This may occur when secondary audit procedures do not provide the

planned assurance level, and the auditor is forced to seek a higher level of assurance from the MUS sample. A second-stage sample may also be needed when an initial MUS sample exhibits a higher error rate than planned, and the initial sample by itself is not sufficient to quantify an audit adjustment.

Unfortunately, as demonstrated by Wright (1991), an MUS sample which is augmented using PPS selection does not necessarily yield PPS properties. Rather, a two-stage PPS selection process tends to over-represent small items and under-represent large items. This pattern is contrary to the objectives of auditors who prefer to focus effort on large items for efficiency and risk reasons (see Roberts 1978, p. 95; Arens & Loebbecke 1981, p. 302; AICPA 1999, p. 37; IFAC 2006, §530.39). Depending on the pattern of error in the population this deviation from PPS inclusion probabilities may yield an excessive risk of incorrect acceptance or an excessive risk of incorrect rejection, and lead to an incorrect audit conclusion.

Wright (1991) investigated the problem of maintaining PPS properties when augmenting an MUS sample using systematic selection. Wright showed that if $n + k$ monetary units are selected in two stages of first n followed by a supplemental sample of k (selected from the original population excluding the n line items “hooked” by the initial n monetary units), with each stage using systematic selection of monetary units, the inclusion probabilities are not, in general, proportional to the size of book values. For applications which use systematic selection Wright (1991) presented a solution to this problem, but the proposed solution cannot be employed when the audit population contains individually significant items (e.g., a line item with a recorded amount equal to or

exceeding the population recorded amount divided by the number of line items examined -- see Leslie et al. 1979, p. 185; AICPA 1999, p. 62). This limitation precludes use of Wright's solution in many situations as accounting populations often include individually significant items.

In addition to the above sample augmentation problem, all of the methods used in practice exhibit randomness in the line item sample size. This is especially severe for unrestricted random dollar sampling and the original version of sieve sampling (Horgan 1997, p. 45). Auditors prefer to avoid this randomness because it may undermine the credibility of the audit if the final line item sample is less than planned. Horgan (1997) developed a variant of sieve sampling (stabilized sieve sampling) which yields smaller variations in sample size than unrestricted random sampling and original sieve sampling. However, the variation exhibited by this method tends to exceed that of systematic sampling. This fact may explain why the method has not been adopted in practice and is currently not available in commercial audit software.

In addition to these two limitations, each of the PPS sampling methods commonly used in practice exhibit other important disadvantages. When unrestricted random dollar sampling is used, the auditor is subject to a risk that some individually significant items will be omitted from the final sample (Leslie et al. 1979, p. 102; Wurst et al. 1989a, p. 200). Cell sampling exhibits the same behavior (Leslie et al. 1979, p. 103). A consequence of this property is that audit effort may be inefficiently directed to smaller line items rather than the larger significant line items. Additionally, Wurst et al. (1991, p. 342) found that both

unrestricted random dollar sampling and cell sampling are less effective than systematic sampling in bringing errors into the sample, and this weakness was found to increase in severity as sample size increased. This property is undesirable since one important objective of an audit is to identify material errors, individually or in the aggregate, in the population (AICPA 1999, p. 35). Finally, when systematic selection is employed reliable control of sampling risk is contingent on an *unverifiable* assumption that unusual error clustering does not exist in the population errors (Leslie et al. 1979, p. 108). If such clustering does exist the result may be excessive sampling risk and an incorrect audit decision.

This paper introduces and tests a new method of PPS selection, a modified version of sieve sampling which overcomes limitations of selection methods currently used in practice. Although the method is marginally biased in favor of larger line items, extensive simulations using realistic audit scenarios indicate that the method yields single- and multi-stage inclusion probabilities which are indistinguishable from PPS probabilities. Simulation results also indicate the method yields stable sample sizes, provides reliable control of single-stage beta risk regardless of how errors are clustered in the population, and exhibits tightness and efficiency measures that are comparable to those of systematic selection. While use of a properly designed multi-stage sample is recognized as a valid audit technique (AICPA 1999, p. 73; IFAC 2006, §530.55), measuring the sampling risk of such designs is more complex (see Arkin 1982, p. 26) and beyond the scope of this study.

The remaining sections of the paper are organized as follows. Section 2 discusses procedures for the original sieve method and explains how these procedures are altered to yield modified sieve sampling. Section 3 provides details of the study methodology while section 4 provides simulation results. Section 5 concludes with a discussion of the new method.

2. MODIFIED SIEVE SELECTION (MSS)

2.1 Original Sieve Selection

Sieve sampling in its original form was developed by Rietveld (1978, 1979a, 1979b). With sieve sampling, a random determination is made for each line item in the population to indicate whether it is included in the sample and, if so, which monetary unit is the one selected for the sample. Specifically, let Y_i denote the book value of line item i , in a population with N line items. In addition, let n_l represent the target sample size (initial sample) of monetary units, and let n^* denote the achieved line item sample size, a random variable in advance of sampling. Sampling is performed by selecting a random number (with replacement) between 0 and $(\sum Y)/n_l$ for each line item in the population. If the random number for line item i is Y_i or less then item i is selected, and if the random number is greater than Y_i , then line item i is not selected (see Horgan 1997, p. 42).

Although sieve sampling has existed for over 25 years, it has not been widely used in practice. The reasons for this may be that sieve line item sample sizes can be highly variable (Horgan 1997, p. 45) and the bounds produced by this method are more variable than those produced by cell selection (Wurst et al.

1989b, p. 245). Horgan (1997, 1998) proposed a variant of sieve sampling that yields smaller variations in sample size, but the level of variation still tends to exceed that of systematic sampling. Also, practical problems arise if a population item is selected more than once and appears to be in error.

2.2 Modified Sieve Selection Procedures

To develop modified sieve sampling the procedures for original sieve sampling are restated to a different, but functionally similar, formulation. These procedures are then modified to yield a stable sample size. For circumstances where individually significant items are present, a second modification is employed to ensure that, when a sample is augmented, the threshold at which line items become individually significant is revised downward and all items meeting this revised threshold are included in the augmented sample.

The original sieve selection procedures for an initial sample can be reformulated as follows. Item i is included in the sample if and only if:

$$RN_i \cdot (\sum Y) / n_l \leq Y_i,$$

where RN_i is selected under the uniform distribution on (0, 1).

Note that this formulation can be restated as:

$$(\sum Y) / n_l \leq (Y_i / RN_i).$$

Given this reformulation of the inclusion condition, a stable sieve sample of n^* line items can be selected using the following procedures. For each line item:

1. select a random number (RN_i) between 0 and 1,
2. compute the ratio $R_i = Y_i / RN_i$,
3. sort all R_i in descending order, and
4. include in the sample the n^* items with the largest R_i ratios.

When the population contains individually significant items these procedures must be further modified to ensure that these items are included in the sample. For a single-stage sample these are items where: $Y_i \geq (\Sigma Y)/n_1$; and for a two-stage sample these are items where: $Y_i \geq (\Sigma Y)/(n_1 + n_2)$. Ensuring selection of all significant items is easily achieved for a single-stage application by identifying all such items prior to sample selection and including them in an upper stratum (n_1^{upper}) which is audited. The remaining lower stratum sample items ($n_1^{lower} = n_1 - n_1^{upper}$) are selected using the modified procedures set forth previously. However, if a sample is augmented there is no guarantee that all significant items, as redefined after sample augmentation, are among the $(n_1 + n_2)$ items with the largest R_i . This objective can be achieved as follows.

The first step is to calculate the ratios (R_i) as defined above and sort the population line items in descending order based on their ratios (R_i). Based on this sort procedure each line item is assigned an order number g_i from 1 to N , with $g_i = 1$ for the line item with the largest R_i ratio, and $g_i = N$ for the item with the smallest ratio. Absent the presence of individually significant items this order number (g_i) indicates the sequence in which line items enter the sample.

To ensure that individually significant items are always included in multi-stage samples the following procedures are employed. First, sort the population line items in descending order based on their book values (Y_i). For each line item calculate the sample size (f_i) at which the item becomes individually significant, taking into account that if an item is individually significant, all items with a larger book value are also individually significant. To compute f_i let Z_1, Z_2, \dots, Z_N

be the book values of the line items in descending order from large to small, so that Z_1 is the largest item and Z_N is the smallest. Then, calculate f_i as:

$$f_i = \left[\sum Y - \sum_{j=1}^{i-1} Z_j \right] / Z_i \quad [1]$$

Since f_i must range between 1 and N it indicates the order in which population line items enter the sample owing to their status as individually significant.

For each line item a comparison is made of the point at which it would enter a sample based on its: (1) ratio order number (g_i) and (2) individually significant order number (f_i). From this comparison the earliest sample entry point for each line item: $h_i = \min(g_i, f_i)$ is determined. Then, the population is sorted in ascending order on h_i to establish the actual selection order (s_i) for each line item. Here $s_i=1$ indicates the first line item selected while $s_i=N$ is the last item selected. The result of this process is a listing of all population line items in descending order of selection (applicable to single- and multi-stage samples). This list-sequential property of the method ensures that a single-stage sample of m line items ($m = n_1 + n_2$) will consist of exactly the same line items selected in two stages of first n_1 followed by n_2 . This property permits sample augmentation to proceed in small, and therefore efficient, increments.

To determine the total line item sample size, divided between the upper census stratum and the lower PPS stratum, the following procedures are employed. For every possible line item total sample size ($n = 1, 2, 3, \dots, N$):

- (1) calculate the total number of individually significant items representing the upper census stratum,
- (2) calculate the difference between the total sample size and the number of individually significant items to determine the

number of line items selected from the lower PPS stratum,

- (3) calculate the lower PPS stratum selection parameter S_i as the book value of the entire population less the book value of upper stratum line items at that stage, divided by the number of sample items selected at that stage. This value will decline as the lower stratum sample size increases.

The required total sample size is found at the stage where S_i first reaches, or falls below, the threshold for defining individually significant items. For a single-stage sample of size n_1 this is: $(\sum Y)/n_1$, and for a two-stage sample of n_1 followed by n_2 this is: $(\sum Y)/(n_1 + n_2)$.

When the population consists entirely of lower stratum items these procedures will always yield a fixed sample size. If the population includes upper stratum items these procedures will, in almost all cases, yield a fixed sample size. There are unique circumstances where these procedures yield a small variance in line item sample size when upper stratum items exist. But, based on simulation results, these circumstances are rare and occur when the last required sample item

added has a book value that is smaller than $\sum Y - \sum_{j=1}^{i-1} Z_j$ and larger than

$\sum Y - \sum_{j=1}^i Z_j$ and the population contains multiple items with book values

between these values.

Depending on the population characteristics, a sample selected in this manner may exhibit selection probabilities that are approximately PPS. This is an artifact of the procedures used to stabilize the sample size only, as the division of the book value by a random number and comparing the result to the sampling

interval is mathematically equivalent to the comparison between the book value and the product of sampling interval and the same random number.

To illustrate this property consider the following simple case. Assume that a sample of one line item is to be selected from a population of two line items with values Y_1 and Y_2 . If the selection process is truly PPS the selection probability π_1 of the first item is $Y_1/(Y_1+Y_2)$ and for the second item $\pi_2 = Y_2/(Y_1+Y_2)$. Using modified sieve procedures, with a sample of one from a population of two items, Y_1 will be selected if $R_1 > R_2$, and Y_2 alternatively. If Y_1 is larger than Y_2 , the selection probability of Y_2 is now $\pi_2/2\pi_1$, and it is easily seen that this probability equals π_2 if and only if $\pi_1 = \pi_2 = 0.5$. Also, since $\pi_2/2\pi_1 < \pi_2$, it follows that Y_1 has a greater than proportional to size selection probability, and Y_2 has a less than proportional to size selection probability. The effect is to slightly favor population elements with larger book values. However, as demonstrated in the simulation results which are presented later, in realistic audit situations this selection bias is statistically undetectable. In all cases the selection bias has no adverse effect on observed sampling risk. The next section describes the procedures used to test modified sieve sampling.

3. STUDY METHODOLOGY

3.1 Overview

To test the performance of modified sieve sampling 256 individual simulations were performed. These simulations involved the use of 2 actual accounting populations seeded with 4 different rates of total error with 2 alternative error bunching patterns (a total of 16 distinct populations). These

populations were fully crossed with 2 sampling plans and 8 different levels of planned beta risk (risk of incorrect acceptance). In each simulation modified sieve sampling was used to select a total of 100,000 single-stage samples. These samples were audited, upper bounds computed, and an audit decision (accept or reject) reached. To provide a control comparison all simulations were repeated using systematic n th dollar selection on a randomized order population (hereinafter referred to as randomized systematic sample).

For each simulation, lower stratum line item selection counts were tested for deviation from PPS selection properties. Upper stratum selection counts were not tested because a census is conducted in the upper stratum. Analyses were also conducted to determine the observed beta risk (non-coverage) and distributional characteristics of the projected bounds. Results for modified sieve and randomized systematic sampling were compared to investigate the relative performance of these two selection methods.

3.2 Study Populations

Audit values for the study populations were created by seeding overstatement errors into the book values of two actual accounting populations (1M and 4) similar to those used in previous studies (Neter & Loebbecke 1975; Wurst et al. 1989a). Population 1M is characterized by a large number of relatively small accounts. Population 4 has fewer accounts, but the individual amounts are much larger than those in population 1M. The major characteristics of the book values of the two accounting populations are presented in Table 1.

These populations were used because they exhibit the characteristics of realistic accounting populations and the variation that appears in accounting populations.

[Insert Table 1 here]

The total overstatement error seeded into these populations was manipulated on four levels. The total error amounts, specified in terms of fractions of the populations' total book values, were: 0.0125, 0.025, 0.05 and 0.10. These values are believed to provide a realistic range of tolerable error rates used in practice, and are similar to those used in several other studies (Wurst et al. 1989a; Plante et al. 1985). Because one objective of the study was to determine the reliability of modified sieve in controlling beta risk, error taints were set to 100 percent. For a given seeded material error amount, setting all taints to 100 percent yielded the smallest possible number of line items in error. This provides a worst-case test of beta risk (non-coverage) since the probability of including an erroneous line item in the sample declines as the number of erroneous line items declines.

The second characteristic manipulated in creating population audit values was the pattern of error bunching. Two bunching patterns are used: low and uniform bunching. In low bunching the smallest population line items were seeded to contain the total error. This bunching pattern was of particular interest since modified sieve selection is known to exhibit a minor bias against smaller population line items. As such, the low bunching pattern represents a worst-case test for modified sieve selection. For the uniform bunching pattern the total error amount was divided equally between small and large line items. To achieve this,

the population was split into two groups by book value such that the total book amounts for the two groups were approximately the same. Then errors were allocated randomly to items within each group.

3.3 Planned Sample Size

In computing the required sample size for each simulation the study employed two alternative sampling plans commonly used in practice. One plan allows for $k = 0$ errors while the other plan allows for $k = 1$ error. The first plan yields a minimum sample size and is often employed when the population error rate is expected to be zero or very near zero. However, this plan is subject to a higher alpha risk (risk of incorrect rejection) if population errors do exist, and a greater likelihood of requiring a second-stage sample. The second plan uses a larger sample but has a lower alpha risk (*ceteris paribus*) and is typically used when the population error rate is thought to be non-trivially above zero.

In planning sample sizes the tolerable error rate was set equal to the rate of error in the population to be sampled (.0125, .025, .05, or .10). Planned beta risk was manipulated on eight levels: .01, .04, .05, .07, .15, .29, .37, and .47. Values on the range of .04 to .37 were selected for study because they are frequently used in practice. The values of .01 and .47 were included to test whether simulation results are valid beyond the range commonly used in practice (AICPA, 1999).

Given the allowable error count (k), total error (M) implied by the tolerable error rate, planned beta risk (β), and population book value (N), the sample size (n) was derived using the hypergeometric based computation of

equation [2]. Here, the sample size is determined by finding the smallest integer n that satisfies the inequality.

$$\sum_k \frac{\binom{N-M}{n-k} \binom{M}{k}}{\binom{N}{n}} \leq \beta, \quad [2]$$

This method of planning the sample size was used rather than the Poisson based method of Leslie et al. (1979, p. 186) because it more accurately describes sampling without replacement and because this method is known to be widely used in practice. Actual sample sizes ranged from 8 to 528 for population 1M and from 8 to 490 for population 4.

3.4 Control Sample Selection Method

To provide a control comparison, all simulations were replicated using randomized systematic sampling. Systematic selection was chosen as the control because it is reported to be the primary method presently used in practice (IFAC 2006, §530.39; Wright 1991, p. 148). For a detailed description of systematic selection see Anderson and Teitlebaum (1973), Leslie et al. (1979, p. 108) or AICPA (1999, p. 64). Because the performance of systematic selection may be adversely affected by unusual error clustering in the population, simulation procedures included a unique randomization of the population before selection of each systematic sample (similar to Plante et al. 1985, p. 44). This randomization procedure eliminated any potential weakness in the performance of systematic sampling and thus provided as strong a control comparison as possible.

3.5 Sample Evaluation Method

The upper bound for the population total error amount was used to conduct the hypothesis test. This was done by comparing the sample upper bound with the materiality amount (the smallest amount of error that would cause the population to be considered unacceptable). If the bound was less than the materiality amount, the population was regarded as acceptable; otherwise, it was concluded that the population may be materially misstated and the population was not accepted. For simulation purposes, the amount of error seeded into each population was considered material.

The sample upper bound was computed as $\Sigma(Y).P_{1-\beta}(n, k)$. In this computation $P_{1-\beta}(n, k)$ is the upper $100(1-\beta)\%$ confidence limit for the population proportion of misstatements when a sample of n is selected and k misstatements are found in the sample. $P_{1-\beta}(n, i)$ is defined as M_U/N with M_U being the smallest M that satisfies equation [3] below.

$$\sum_k \frac{\binom{N-M}{n-k} \binom{M}{k}}{\binom{N}{n}} \leq \beta, \quad [3]$$

It should be noted that M_U is a discrete random variable. Therefore, confidence is not exactly equal to $1-\beta$ but rather $\geq 1-\beta$ with $\gg 1-\beta$ being a not uncommon occurrence.

3.6 Simulation Details

Random numbers were generated using a multiplicative congruential generator (multiplier 950706376) available in the IMSL Libraries (Visual

Numerics 2003, 1311). This generator and multiplier are known to perform well (see Gentle 1998, 170). To ensure independence across samples comprising a particular simulation, the study utilized a unique seed number for each sample h ($h = 1, 2, \dots, 100,000$) within a simulation. To permit inferences about multi-stage selection probabilities for modified sieve samples the same set of seed numbers ($SN_{h=1}, SN_{h=2}, \dots, SN_{h=100,000}$) was used in each modified sieve simulation. Owing to the list-sequential nature of modified sieve sampling and use of the same set of seed numbers across all simulations, the selection order of items (s_i) for a particular sample number h was the same across all modified sieve simulations. The result is that for any two modified sieve samples A and B with the same sample number h and $n_A > n_B$, sample A included all n_B sample items plus an additional $(n_A - n_B)$ sample items. Hence, lower stratum multi-stage inclusion probabilities for modified sieve selection can be judged by comparing chi-square goodness-of-fit tests for progressively larger single-stage sample sizes (presented in Tables 2). However, because the mechanics of systematic sampling are different, testing inclusion properties for various sized single-stage samples does not provide evidence about multi-stage inclusion properties.

3.7 Method of Analysis

To determine if single-stage selection counts for lower stratum line items exhibited PPS properties a chi-square goodness-of-fit test was performed for each simulation. Each chi-square test included a class category for each line item appearing in the lower stratum. For simulations without an upper stratum the numbers of classes were 8,282 and 4,033 for populations 1M and 4, respectively.

In cases where an upper stratum was present the numbers of classes were slightly less. Results of these tests were considered statistically significant for p-values of .01 or less. Given the significance criterion used, the number of classes used, and the number of samples selected in each simulation run, the power of the resulting chi-square tests to detect a small effect size $w = .10$ (see Cohen 1988, p. 216) were estimated to be .99 or greater (see Faul et al. 2007). Additionally, for each simulation the effect size index w was computed to provide an estimate of the potential magnitude of the deviation from PPS selection rates, if any.

To determine the observed beta risk (non-coverage) for a simulation the bound for each sample in that simulation was compared to the tolerable error amount. If the bound for a sample was less than, or equal to, tolerable error the audit decision was to accept the population. The observed beta risk (non-coverage) was computed as the proportion of samples resulting in an accept decision. Observed beta risk for a reliable procedure should be very near, but always below, the planned beta risk.

A procedure may be reliable, but perform poorly relative to other methods due to a lack of: (1) *tightness* and/or (2) *relative efficiency* in its bound (see Horgan 1997, p. 47-48; Plante et al. 1985, p. 45). In this context *tightness* refers to how close the bound is, on average, to the true overstatement error. *Tightness* is a relative measure, and is computed as the mean of the bound sampling distribution divided by the true overstatement error in the population. Reliable procedures with better tightness are less susceptible to incorrect rejection (alpha error) and thus avoid the cost of extending audit work to verify the presence of a material

error. To judge the relative performance of modified sieve and randomized systematic selection the tightness of bounds for these two methods were compared.

A procedure may be reliable and yield a *tight* bound, but exhibit suboptimal performance because the bound exhibits a higher degree of variability than bounds produced by another method. In this study the *relative efficiency* of modified sieve and randomized systematic sampling was computed as the standard deviation of the bound distribution for modified sieve divided by the standard deviation of the bound distribution for randomized systematic sampling. A measure less than one indicates greater relative efficiency for modified sieve sampling while a ratio greater than one indicates less relative efficiency.

4. RESULTS

4.1 Tests for PPS Selection

Table 2 reports observed line item sample size information (means and standard deviations) and p-values with related effect size estimates for chi-square goodness-of-fit tests that lower stratum selection rates are PPS. Data presented in this table are for both modified sieve (MSS) and randomized systematic (SYS) samples. Because of the similarity of results over all 512 simulation runs (256 each for modified sieve and randomized systematic selection) data presented in Table 2 are limited to results for the 32 simulation runs using population 1M and population 4 with a tolerable error rate of 5% and a planned error count of $k=0$. It should be noted that these results are the same regardless of the error bunching

pattern since only book values affect selection rates. Results for all simulation runs are available from the lead author on request.

[Insert Table 2 here]

As disclosed in Table 2, observed mean line item sample sizes for modified sieve and randomized systematic selection were identical when rounded to the nearest integer. Although details are not provided in Table 2, observed mean line item sample sizes for both selection methods equaled planned sample sizes except in cases where the presence of upper stratum elements led to a reduction in observed sample size. Theoretically this occurs when the sample size is larger than $\Sigma(Y)/\max(Y_i)$. This tended to occur at the smallest tolerable error rate (1.25%) and low levels of planned beta risk. Modified sieve selection exhibited no variation in observed sample size over the 100,000 samples in each simulation, hence the zero standard deviation. For randomized systematic selection, in most cases observed sample sizes fluctuated between the planned sample size and the planned sample size plus one additional item, hence the nonzero standard deviation. These results indicate that modified sieve selection yields about the same average line item sample size as randomized systematic selection but is more stable.

Data presented in Table 2 also disclose that none of the chi-square goodness-of-fit tests, for modified sieve or randomized systematic selection, were statistically significant (p-values ranged from .093 to .976). Hence, for each simulation a null hypothesis that lower stratum selection rates are PPS cannot be rejected. For modified sieve and randomized systematic selection the estimated

effect sizes were similar and ranged from .02 to .08. These values are quite small and below Cohen's suggested threshold of $w = .10$ for a small effect size (Cohen 1988, p. 224). It should be noted that a lack of statistical significance implies the true effect size is zero (see Cohen 1988, 10). Given the method of computing w (see Cohen 1988, 216) and the similarity of values for modified sieve and systematic selection (which is known to yield PPS inclusion rates), the effect size estimates for modified sieve sampling reported in Table 2 are likely the result of simulation noise rather than deviation from PPS properties.

These results were expected for randomized systematic sampling since this method is known to produce PPS selections. For modified sieve sampling, given the estimated power of the chi-square tests (in excess of .99) to detect small deviations from PPS and the very small estimated effect sizes, these data provide strong evidence that lower stratum selections are materially the same as PPS. While modified sieve sampling is known to slightly favor larger line items, these data indicate that any selection bias is generally too small to be detected in realistic sampling applications. Because of the list-sequential nature of modified sieve sampling these results suggest that both single- and multi-stage samples exhibit inclusion probabilities which are indistinguishable from PPS selection.

4.1 Beta Risk, Tightness, and Relative Efficiency

Table 3 reports observed beta risk for modified sieve and randomized systematic selection. Also reported are measures of tightness and relative efficiency. Because of the similarity of results over all simulations data presented in Table 3 are limited to results for the 32 simulations using population 1M and

population 4 (with low error bunching in audit values) with a tolerable error rate of 5% and a planned error count of $k=0$. Table 4 reports the results for the same simulations using populations with uniform error bunching. Results for all simulation are available from the lead author on request.

[insert Tables 3 and 4 here]

As disclosed in Table 3 modified sieve sampling performed well, yielding observed beta risk values below planned levels and similar to those for randomized systematic sampling. For both modified sieve and randomized systematic sampling it should be recognized that these values overstate (slightly) observed beta risk since the test populations were seeded with error amounts just below the materiality threshold. Given the known tendency of modified sieve to favor larger line items, the low error bunching and 100 percent taintings used in Table 3 simulations represent unrealistic worst-case tests. Even under these severe conditions modified sieve sampling yielded reliable results.

Table 4 beta risk values, based on a rigorous but less extreme testing scenario of uniform error bunching with 100% taints, yield similar conclusions. Here, observed beta risk values for modified sieve and randomized systematic sampling are similar and below the planned level of beta risk. However, with uniform error bunching beta risk values for both sampling methods exhibit a slightly higher degree of conservatism (deviation below the level of planned beta risk).

Tables 3 and 4 also present comparative tightness measures for modified sieve and randomized systematic sampling. As one would expect, the tightness

measures decreased in magnitude as planned beta risk increased. And, at each level of planned beta risk the tightness measures for modified sieve and randomized systematic sampling were very similar indicating that modified sieve and randomized systematic bounds were comparable.

To compare the variability of bounds resulting from modified sieve and randomized systematic sampling the ratio of their standard deviations (relative efficiency ratio) was computed for each level of planned beta risk. As disclosed in Tables 3 and 4 all of these relative efficiency measures were near one indicating that modified sieve and randomized systematic bounds exhibited comparable variability.

5. DISCUSSION

PPS selection methods commonly used in audit practice share two important limitations: (1) initial line item samples cannot generally be augmented while maintaining PPS properties and (2) line-item sample size exhibits random variation. In addition, the most widely used form of PPS selection (systematic) can yield unreliable control of beta risk when the population exhibits unusual error clustering. Modified sieve selection overcomes these limitations and provides reliable control of beta risk.

To test the performance of modified sieve sampling simulations consisting of 100,000 replications each were conducted. These simulations used: (1) realistic accounting populations, (2) common sampling plans, and (3) varied combinations of tolerable error and planned beta risk. Tests are considered worst-case because: (1) populations were seeded with error amounts slightly below the materiality

threshold, (2) all errors in the low bunching condition were seeded in the smallest line items, and (3) all taintings were 100 percent. Analyses of these simulations disclosed no statistically significant evidence of deviation from PPS selection despite the use of tests with high power. Similar results were observed when errors were seeded uniformly in the population. Thus, while modified sieve is known to favor larger population elements, in realistic and extreme tests the magnitude of the bias was too small to be detected. Given the lack of detectable selection bias modified sieve sampling represents an effective means of obtaining single- and multi-stage PPS samples.

Theoretically, modified sieve sampling yields selections that are not exactly PPS. The method favors line items with larger expected PPS selection probabilities, while under selecting line items with smaller expected PPS selection probabilities. This selection bias decreases as the variance in line item expected selection probabilities decreases. In practice, two factors operate to yield a small variance in expected selection rates and thus insignificant selection bias. First, auditors normally stratify line items by recorded value (size) and conduct a census in the upper stratum containing individually significant items. This stratification process reduces the variance of recorded values in the lower stratum and reduces the variance in lower stratum expected selection probabilities. Second, accounting populations subject to monetary unit sampling tend to be large, normally consisting of more than 1,000 line items (populations consisting of more than 10,000 line items are common). The presence of such a large number of line items

in the lower stratum tends to reduce the variance in recorded values and thereby reduces the variance in lower stratum expected selection probabilities.

Simulation results also disclosed that modified sieve sampling yields stable line item sample sizes and single-stage observed beta risk values at or below planned levels (as did randomized systematic sampling). Comparisons with randomized systematic sampling based on bound tightness and relative efficiency suggest similar performance. Testing did not address multi-stage beta risk and no inferences about multi-stage risk should be drawn from this study.

Modified sieve sampling should be attractive to auditors for a number of reasons. In contrast to methods currently used in practice modified sieve sampling permits sample augmentation while effectively maintaining PPS selection rates, and yields a more stable line item sample. Additionally, because a random determination is made for each line item, sampling risk properties of the method are more reliable than those of the widely-used systematic sampling.

REFERENCES

- ACL Services Ltd (2006). *ACL Getting Started*, Vancouver: ACL Services Ltd.
- AICPA (1999). *Audit Practice Release: Audit Sampling*, New York: American Institute of Certified Public Accountants.
- Anderson, R. and Teitlebaum, A.D. (1973). *Dollar-Unit Sampling: A Solution to the Audit Dilemma*, Canadian Chartered Accountant, Vol. 102, No. 4, pp. 30-39.
- Arens, A., and Loebbecke, J. (1981). *Applications of Statistical Sampling to Auditing*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.
- Arkin, H. (1982). *Sampling Methods for the Auditor: An Advanced Treatment*. McGraw-Hill: New York.
- CaseWare IDEA Inc (2004). *IDEA 2004 User Guide*, Toronto: CaseWare IDEA Inc.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition, Hillsdale: Lawrence Erlbaum.
- Faul, F., Erdfelder, E., Lang, A.-G., and A. Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39: 175-191.
- Gentle, J. (1998). *Random Number Generation and Monte Carlo Methods*. Springer: New York.
- Hall, T. W., Hunton, J. E., and Pierce, B. J. (2002). Sampling Practices of Auditors in Public Accounting, Industry, and Government, *Accounting Horizons*, Vol. 16, No. 2, pp. 125-136.
- Horgan, J.M. (1997). Stabilizing the Sieve Sample Using PPS, *Auditing: A Journal of Practice & Theory*, Vol. 16, No. 2, pp. 40-51.
- Horgan, J.M. (1998). Stabilized Sieve Sampling: A Point-Estimator Analysis, *Journal of Business & Economic Statistics*, Vol. 16, No. 1, pp. 42-51.
- IFAC (2006). *IFAC Handbook: Technical Pronouncements*. New York: International Federation of Accountants.
- Jones, P. (1999). *Statistical Sampling and Risk Analysis in Auditing*. Aldershot, UK: Gower.

- Leslie, D.A., Teitlebaum, A.D. and Anderson, R.J. (1979). *Dollar-Unit Sampling: a Practical Guide for Auditors*. Toronto: Copp, Clark, Pitman.
- Neter, J. and Loebbecke, J.K. (1975). *Behavior of Major Statistical Estimators in Sampling Accounting Populations – An Empirical Study*. New York: American Institute of Certified Public Accountants.
- Plante, R., Neter, J., and Leitch, R. (1985). Comparative Performance of Multinomial, Cell, and Stringer Bounds, *Auditing: A Journal of Practice & Theory*, Vol. 5, No. 1, pp. 40-56.
- Rietveld, C. (1978). De Zeefmethode Als Selectiemethode Voor Statistische Steekproeven in de Controlepraktijk (I), *Compact: Computer en Accountant*, Vol. 15, pp. 2-11 (Amsterdam: Klynveld Kraayenhof & Co.).
- Rietveld, C. (1979a). De Zeefmethode Als Selectiemethode Voor Statistische Steekproeven in de Controlepraktijk (II en III), *Compact: Computer en Accountant*, Vol. 16, pp. 2-13 (Amsterdam: Klynveld Kraayenhof & Co.).
- Rietveld, C. (1979b). De Zeefmethode Als Selectiemethode Voor Statistische Steekproeven in de Controlepraktijk (IV), *Compact: Computer en Accountant*, Vol. 17, pp. 9-18 (Amsterdam: Klynveld Kraayenhof & Co.).
- Roberts, D. (1978). *Statistical Auditing*. New York: American Institute of Certified Public Accountants.
- Visual Numerics (2003). *IMSL Fortran Library User's Guide: STAT/LIBRARY*, Volume 2, San Ramon, CA: Visual Numerics, Inc.
- Wright, D.W. (1991). Augmenting a Sample Selected with Probabilities Proportional to Size, *Auditing: A Journal of Practice & Theory*, Vol. 10, No. 1, pp. 145-158.
- Wurst, J., Neter, J., and Godfrey, J. (1989a). Efficiency of Sieve Sampling in Auditing. *Journal of Business & Economic Statistics*, Vol. 7, No. 2, pp. 199-205.
- Wurst, J., Neter, J., and Godfrey, J. (1989b). Comparison of Sieve Sampling with Random and Cell Sampling of Monetary Units. *The Statistician*, Vol. 38, No. 4, pp. 235-249.
- Wurst, J. Neter, J., and Godfrey, J. (1991). Effectiveness of Rectification in Audit Sampling. *The Accounting Review*, Vol. 66, No. 2, pp. 333-346.

Table 1. Characteristics of Book Values
for Accounting Populations

Characteristic	Population 1M	Population 4
Total book value (\$)	334,207.70	7,502,957.20
Mean book value (\$)	40.35	1,860.39
Standard deviation (\$)	72.32	3,865.13
Skewness	6.24	3.08
Kurtosis	48.29	11.40
Maximum (\$)	948.28	24,928.60
Minimum (\$)	0.50	0.10
Total number of line items	8,282	4,033

Table 2. Observed Sample Sizes and Results of Tests for PPS Selection
(TDR=5%, Planned Error=0)

Pop.	Planned Beta	Observed Line Item Sample Size				Chi-Square Test that Lower Stratum Selection is PPS			
		Mean		Std. Dev.		p-value		Effect Size Estimate w	
		MSS	SYS	MSS	SYS	MSS	SYS	MSS	SYS
1M	.01	90	90	0.0	.01	.893	.525	.03	.03
	.04	63	63	0.0	.01	.305	.758	.04	.04
	.05	59	59	0.0	.01	.348	.728	.04	.04
	.07	52	52	0.0	<.01	.530	.976	.04	.04
	.15	37	37	0.0	0.0	.782	.753	.05	.05
	.29	25	25	0.0	.01	.847	.412	.06	.06
	.37	20	20	0.0	<.01	.620	.900	.06	.06
	.47	15	15	0.0	<.01	.889	.093	.07	.08
4	.01	90	90	0.0	<.01	.532	.856	.02	.02
	.04	63	63	0.0	<.01	.811	.244	.03	.03
	.05	59	59	0.0	<.01	.439	.820	.03	.03
	.07	52	52	0.0	<.01	.292	.841	.03	.03
	.15	37	37	0.0	0.0	.940	.264	.04	.03
	.29	25	25	0.0	0.0	.902	.875	.04	.04
	.37	20	20	0.0	0.0	.966	.679	.04	.04
	.47	15	15	0.0	0.0	.195	.979	.05	.05

Key: MSS=Modified Sieve Sampling; SYS=Systematic Sampling

Table 3. Observed Beta Risk, Tightness, and Relative Efficiency
 Modified Sieve versus Systematic Selection
 (TDR=5%, Planned Error=0, Error Bunching=Low)

Pop.	Planned Beta	Observed Beta		Tightness Measure		Relative Efficiency Measure
		MSS	SYS	MSS	SYS	
1M	.01	.0097	.0093	2.603	2.604	.999
	.04	.0388	.0391	2.500	2.498	.998
	.05	.0477	.0481	2.471	2.466	.998
	.07	.0698	.0687	2.437	2.438	1.001
	.15	.1496	.1478	2.324	2.328	.999
	.29	.2785	.2768	2.136	2.136	1.006
	.37	.3608	.3574	2.057	2.064	.998
	.47	.4648	.4636	1.977	1.985	.994
4	.01	.0092	.0089	2.601	2.601	1.003
	.04	.0392	.0379	2.490	2.500	.997
	.05	.0468	.0473	2.469	2.469	.999
	.07	.0694	.0687	2.429	2.433	.999
	.15	.1484	.1457	2.322	2.334	.992
	.29	.2759	.2765	2.139	2.133	1.001
	.37	.3601	.3594	2.052	2.058	.996
	.47	.4655	.4627	1.977	1.980	1.002

Key: MSS=Modified Sieve Sampling; SYS=Systematic Sampling

Table 4. Observed Beta Risk, Tightness, and Relative Efficiency
 Modified Sieve versus Systematic Selection
 (TDR=5%, Planned Error=0, Error Bunching=Uniform)

Pop.	Planned Beta	Observed Beta		Tightness Measures		Relative Efficiency Measure
		MSS	SYS	MSS	SYS	
1M	.01	.0083	.0088	2.603	2.600	.997
	.04	.0361	.0354	2.501	2.501	1.002
	.05	.0448	.0446	2.473	2.470	1.003
	.07	.0649	.0655	2.437	2.438	.994
	.15	.1438	.1455	2.327	2.325	.997
	.29	.2721	.2724	2.200	2.139	.997
	.37	.3572	.3579	2.059	2.052	1.008
	.47	.4620	.4592	1.975	1.983	.996
4	.01	.0053	.0052	2.596	2.594	1.002
	.04	.0279	.0278	2.491	2.489	1.005
	.05	.0336	.0345	2.459	2.463	.993
	.07	.0526	.0518	2.433	2.432	1.000
	.15	.1251	.1277	2.320	2.314	.999
	.29	.2495	.2510	2.137	3.133	.997
	.37	.3317	.3316	2.056	2.056	1.005
	.47	.4365	.4359	1.983	1.984	1.004

Key: MSS=Modified Sieve Sampling; SYS=Systematic Sampling