DESIGN OF A NATIONAL SAMPLE FOR
AN IN-PERSON SURVEY OF SMALL BUSINESS

Prepared for:
Small Business Administration
Office of Advocacy

Contract No. SBA-4117-OA-89

Prepared by:
Harold Nisselson
Joseph Waksberg

May 1991

Westat, Inc.
1650 Research Boulevard
Rockville, MD 20850
June 3, 1991

David Hirschberg  
Office of Advocacy  
Small Business Administration  
409 Third Street, S.W.  
Washington Office Center - 5th Floor  
Mail Code 3110  
Washington, D. C. 20416

Dear Dave:

We are submitting herewith three copies of a report "Design of a National Sample for an In-Person Survey of Small Business." This report describes the research carried out under contract no. SBA-4117-OA-89.

If you have any comments or need any additional information, please let me know.

Sincerely,

Joe
Joseph Waksberg  
Chairman of the Board

JW:mlp  
Enclosures

cc:  D. Levine  
    H. Nisselson  
    J. TerMaat
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>INTRODUCTION .......................................................... Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INTRODUCTION ..................................................................... 1-1</td>
</tr>
<tr>
<td></td>
<td>1.1 Overview ....................................................................... 1-1</td>
</tr>
<tr>
<td></td>
<td>1.2 Scope of Work .................................................................. 1-1</td>
</tr>
<tr>
<td></td>
<td>1.3 Some Options for the Survey Design ................................ 1-2</td>
</tr>
<tr>
<td></td>
<td>1.4 Outline of Report ....................................................... 1-5</td>
</tr>
<tr>
<td>2</td>
<td>SURVEY ESTIMATES AND SAMPLING VARIANCE .................................. 2-1</td>
</tr>
<tr>
<td></td>
<td>2.1 Survey Model .................................................................... 2-1</td>
</tr>
<tr>
<td></td>
<td>2.2 Estimates from the Survey ................................................ 2-4</td>
</tr>
<tr>
<td></td>
<td>2.3 Sampling Variances .......................................................... 2-7</td>
</tr>
<tr>
<td>3</td>
<td>OPTIMIZATION OF THE DESIGNS ................................................. 3-1</td>
</tr>
<tr>
<td></td>
<td>3.1 Overview .......................................................................... 3-1</td>
</tr>
<tr>
<td></td>
<td>3.2 Independent Area Sample .................................................. 3-1</td>
</tr>
<tr>
<td></td>
<td>3.3 Dual Frame Sample ............................................................. 3-5</td>
</tr>
<tr>
<td></td>
<td>3.4 Relative Costs of the Two Alternative Designs ...................... 3-7</td>
</tr>
<tr>
<td>4</td>
<td>FINDINGS AND RECOMMENDATIONS ............................................. 4-1</td>
</tr>
<tr>
<td></td>
<td>4.1 Alternative Designs Considered ........................................... 4-1</td>
</tr>
<tr>
<td></td>
<td>4.2 Design Specifications ....................................................... 4-1</td>
</tr>
<tr>
<td></td>
<td>4.3 Future Research .............................................................. 4-2</td>
</tr>
</tbody>
</table>

## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Outline of stages in alternative sample designs .................... 2-2</td>
</tr>
<tr>
<td>3-1</td>
<td>Summary of rough costs estimates for a survey of small businesses: Single frame vs. dual frame approach .......... 3-8</td>
</tr>
</tbody>
</table>

## Appendices
1. INTRODUCTION

1.1 Overview

Recent legislation, including the Women’s Business Ownership Act of 1988, PL 100-533, call for current detailed business data, particularly data on women business owners. For a variety of reasons, the Small Business Data Base currently maintained by the Small Business Administration (SBA), and information available from Census Bureau surveys and studies of the various statistical agencies, do not meet all of SBA’s needs.1

Under contract to the SBA, Westat, Inc. carried out research to develop recommendations for the design of a national in-person survey of a probability sample of small businesses to provide better data on the demographic and financial characteristics of small businesses. It is contemplated that this type of survey will be repeated in the future at periodic intervals to be determined. The survey also would be intended to go beyond the limited legislative requirement in that it would provide an effective small business data collection program for all segments of small business.

1.2 Scope

The survey is to cover all small business firms with the exception of activities in SIC Division A and Division J. Excluded are:

- All agencies of federal, state, and local governments, and businesses operated by government bodies
- Businesses conducted out of private residences, unless a specific portion is clearly designated for business use with a separate entrance (e.g., doctor’s office in a residence with a separate entrance)2

---

1 See: David Hirschberg, “An Evaluation of the Small Business Data Base” (Processed), Small Business Administration, 1989

2 Coverage of small businesses operated by the self-employed but not in a visible business site may be achievable through an SBA-sponsored supplement to the Census Bureau’s Current Population Survey. See: David Hirschberg, “Special Census/CPS Data on Self Employed”
For purposes of this study a small business firm is a firm having fewer than 500 full-time employees. In the case of firms owned or controlled by other companies, the relevant criterion is total employment of all branches, subsidiaries, and establishments of the ultimate parent company.

1.3 Some Options for the Survey Design

An efficient sample design for an in-person survey is a multi-stage probability sample design suitable for effective deployment of field staff. The basic sampling unit is an individual nonresidential building with screening of the occupants for eligibility. Exceptions would be made for very large or complex structures such as office buildings and shopping malls which can be subsampled.

The first stage in a general multi-stage plan would be the selection of a stratified sample of PSUs (collections of counties or metropolitan areas referred to as primary sampling units.) The largest metropolitan areas would be included in the survey sample and samples of smaller PSUs. It appears that most of the available national samples of PSUs used by Federal statistical agencies or private statistical research organizations would be suitable. PSUs may represent very large geographic areas. To provide efficient clustering of sample sites for personal interviewing, PSUs would be subdivided into second stage units (SSUs) for subsampling. The second stage would be the selection within the sample PSUs of stratified samples of SSUs (collections of ZIP code areas or census tracts referred to as secondary sample units). Zip code areas appear to be preferable. Reasonable estimates of numbers of establishments of potential interest to SBA by ZIP code area, or other measures of relevant economic activity, can be constructed fairly easily.

There are four options or major paths that could be followed to obtain or develop lists of establishments for further subsampling for the survey. The survey design could be based on:

1. Sampling materials from and existing survey of buildings or establishments

   This approach would have four components:
   a. Re-canvass of buildings from the existing survey to develop an updated list of establishments
b. Listing and screening of establishments that were never actually used in buildings that were listed in sample segments from the survey

c. Listing (excess listings) in segments created for within-SSU sampling in the survey but not previously listed, and screening of establishments in them

d. Listing and screening of buildings that were not in existence as nonresidential buildings at the time of the survey (new or converted buildings)

It is highly attractive from a cost standpoint, and several existing potential list sources were explored in the study, but with negative results. The Energy Information Administration ruled out use of sampling materials from its NBECS/CBECS surveys. In discussion with the Bureau of Labor Statistics (BLS), the potential use of the Business Establishment List ("BEL") it is constructing was ruled out; and similarly for use of Bureau of the Census frames. The "Nonhousehold Establishment Frame" compiled by the University of Michigan Survey Research Center (SRC) for the U.S. Post Office in 1978 is considered by SRC to be too out of date to be useful. It is not used by SRC and was not recommended by SRC for current use in the study. Special purpose lists such as compiled for NCHS or NIH surveys are too limited in their coverage to be useful. Therefore, it was concluded that this is not a feasible option.

2. A commercial list or lists of businesses from which a sampling frame of businesses could be constructed.

This approach is relatively attractive from a cost standpoint. It would provide flexibility for sampling purposes since businesses in the frame could be stratified by characteristics shown (e.g., SIC, size) and sampled at varying rates appropriate for the objectives of a particular SBA survey.

However, this approach — relying solely on a commercial list — has serious drawbacks for SBA, the foremost being the substantial bias due to the fact that these lists tend to exclude many of the new businesses and the smallest small businesses.

It would be necessary to trace sample businesses that could not be located at the address given in the frame, in order to determine whether or not they were currently out-of-business. If still in business but at a different location, follow-up would be needed to obtain interviews from them. Although the costs of these activities can be controlled by techniques such as subsampling, they still may be relatively large. It is anticipated that a substantial proportion of the sample would be found to be out-of-business or not locatable, also leading to biases in the coverage of small businesses.

With these two possibilities ruled out, attention focused on two other options.
3. **An independent area sample approach**

This approach would develop a new sample of buildings and the small businesses in them independent of existing surveys (although there might be some chance overlap). It would be based on an independent field canvass of a sample of area segments, defined within the SSUs (ZIP code areas), for which lists of businesses can be cost-effectively compiled in the field for selecting the final sample.

The segments would be defined on the basis of Census geography, i.e., blocks or block groups. Two alternatives to Census geography for further subsampling were studied. One promising alternative would be to use 9-digit Zip codes ("ZIP + 4") to define subsampling units. The other alternative considered was areas served by 3-digit telephone exchanges. Both alternatives were discarded after fairly extensive exploration. It was found that a substantial proportion of ZIP + 4's represented post office boxes and the Post Office would not release the identity of the box holder. Moreover, in many instances they did not represent compact areas. Westat's experience also indicated that addresses served by more than one 3-digit exchange were intermingled within a given geographic area, and the addresses served by a given 3-digit exchange might be scattered over an area that could be considered relatively large for field data collection operations. Therefore these alternatives were not pursued.

Field listing has been used to build the most complete frame of commercial buildings. Listers construct lists by driving or walking through the sampled area recording addresses as they go. The problems are more complex when this technique is applied at the establishment level as would be necessary for a survey of small businesses. In larger buildings or shopping malls, the lister's job may be quite extensive and may require consulting managers or rental agents to obtain complete listings including unoccupied units.

4. **A dual frame approach**

This approach involves developing a sample of small businesses by creating a sample of buildings or locations (sites) the small business occupants of which will constitute a sample of small businesses. A dual frame sample will consist of two components:

a. A sample of businesses (listings) from a commercial business list or lists. The physical locations of these businesses would be treated as a sample of buildings or locations within a building as appropriate. The file sampled will be referred to as the List Frame, and the sample will be referred to as the List Sample.

This represents the first-known attempt to develop a nationwide sample of businesses by developing a nationwide sample of business locations. It avoids many of the problems created by treating the businesses selected from a commercial list as a sample of businesses.
b. A sample of area segments (small geographic areas such as blocks) to be canvassed for small businesses not in the list frame or list sample cases that could not be identified in the field. The universe of all areas segments will be referred to as the Area Frame, and the sample will be referred to as the area sample.

Commercial list such as Dun’s National Business List or lists from other suppliers such as American Business Lists or John Klein & Associates could reasonably be considered for a dual frame approach. All such sources appear to include a 5-digit Zip code as part of the establishment address. A suitable frame of businesses for sampling could, therefore, be created by purchasing the listing for a sample of Zip code areas established by SBA. A survey design for SBA, then, would have the following components:

(1) Selection of sample of ZIP code areas within a national PSU sample. Measures of size for the stratification of ZIP code areas for sampling would be based on Bureau of the Census County Business Patterns data by ZIP code, factors derived from the 1987 Economic Censuses for nonemployers, and data from the list source.

(2) A sample of nonresidential buildings or locations that would consist of the buildings or locations within buildings occupied by a sample of businesses selected from the commercial list for the sample ZIP code areas. If firms are identified in the lists, consideration could be given to sampling firms rather than individual establishments.

(3) A subsample of area segments from the sample ZIP code areas in which all nonresidential buildings would be field listed and their occupants identified. The following modification might be considered:

a) Within sample buildings that match buildings in the list frame, list occupants and check whether they are in frame. Subsample those not in the frame to provide the supplement to the list frame.

b) In area segments, only list occupants of buildings not in frame (rather than occupants of all buildings)

1.4 Outline of Report

The following chapters summarize the survey design research that were carried out for the two alternatives. Section 2 presents the statistical estimators for each of the two designs studied (the independent area sample and the dual frame approach) and their sampling variances. Section 3 describes the optimization of each of the two designs and an evaluation of the alternatives. Section 4 summarizes the findings and recommendations.
2. SURVEY ESTIMATES AND SAMPLING VARIANCES

2.1 Survey Model

Figure 2-1 provides an outline of the sampling stages in each of the two alternative sample design approaches. The sampling operations shown create the equivalent of an unbiased frame of business addresses/sites large enough to provide samples for successive surveys of small businesses over time. This is a cost-effective approach since it minimizes the field listing and sampling operations over successive surveys. There is an issue as to whether the current occupants of the business locations should be administered a set of screening questions as part of the listing operations to identify those which are small businesses and characteristics such as their SICs, or whether such screening should be part of the survey interview. Since significant turnover of the occupants is to be expected over time, because of the dynamic nature of the small business population, it is recommended that screening be carried out as part of the survey interview. The survey sample itself can be made large enough to compensate for the anticipated losses due to out-of-business cases and businesses found not to be eligible at the time of the survey.

As indicated in Figure 2-1, the first three stages of the survey sample are the same under both the independent area sample design and the dual frame design, although the allocation of the sample at each of the three stages may differ between the two designs. We assume a survey design in which, at the first stage, certain PSUs would be selected with certainty. For example, these might be large or parts of very large metropolitan areas such as large Metropolitan Statistical Areas (MSAs) or Primary Metropolitan Statistical Areas (PMSAs) within Consolidated Metropolitan Statistical Areas (CMSAs) designated by the Bureau of the Census. These are referred to as "self representing" PSUs. All other PSUs would be stratified for sampling, and we assume for this research a design with one PSU per stratum selected with probability proportionate to a measure of size. These first stage units are referred to as "nonself-representing" PSUs.

In both of the alternative designs, the second stage sampling units (SSUs) would be five-digit zip code areas or groups of five-digit zip code areas constructed to be of approximately equal size in terms of number of businesses of potential interest. We assume these to be selected by simple random sampling. However, consideration may be given to sampling SSUs with stratification within PSUs and possibly different sampling rates in different strata.
FIRST-STAGE SAMPLE
A stratified sample of PSUs (primary sampling units) appropriate for a national survey. The PSUs would be large metropolitan areas, and counties or groups of counties outside those areas. The first-stage sample would be the same under either design.

SECOND-STAGE SAMPLE
A (stratified) sample of SSUs (secondary sampling units) within the PSUs selected in the first-stage sample. The SSUs would be five-digit ZIP code areas. A single sample of ZIP codes would be used to develop further components of each of the alternative designs.

THIRD-STAGE SAMPLE
Within the selected ZIP codes recode the addresses of businesses from the business list sources to 1990 Census geography: Blocks/Block Groups.

Within the selected ZIP codes select a sample of businesses from the list source whose addresses are to be used to establish (by field check) a current list of business addresses.

Within a subsample of the selected ZIP codes define "area segments" on the basis of Census geography to be field listed to supplement the sample of addresses from the list source.

Within the selected ZIP codes define "area segments" on the basis of Census geography to be field listed.

Samples of business addresses for surveys of small businesses.

Samples of business addresses for surveys of small businesses.

Figure 2-1. Outline of stages in alternative sample designs
2.1.1 Independent Area Sample

In this approach, the third stage would be the subdivision of the selected SSUs into area segments for which lists of business sites can be cost effectively compiled in the field for selecting, at the fourth stage, the final sample of businesses to be surveyed.

The area segments would be based on Census geographic areas (e.g., blocks or block groups) associated with the SSU. For this purpose Census geographic areas would be associated with the zip code of the post office that serves the area. Area segments would be constructed to be of approximately equal size in terms of number of small businesses to the extent feasible. To assist this process, the addresses of businesses shown in the list source would be computer coded to 1990 census geography, and segments would be made of approximately equal size in terms of list source units. These would be subject to field review. Area segments would be sampled by random sampling, possibly stratified random sampling as would be efficient. All businesses and their physical locations within the selected area segments would be listed by field canvass. The fourth stage units would be individual businesses.

2.1.2 Dual Frame Sample

The dual frame sample would be selected in three stages for the list frame and four stages for the area sample frame.

List Frame

For the list sample frame, the third stage sampling units would be businesses shown in the list source or sources with addresses located in the zip code areas of the selected SSUs. The addresses of these businesses would be treated as a sample of sites (buildings or locations within buildings). The third stage units may be selected as a simple random sample, or as a stratified sample with possibly different sampling rates in different strata. The survey would conduct interviews for the current business occupants of the selected sites.
Area Sample Frame

For the area sample frame, the third stage sampling units would be area segments within the selected SSU's. As in the independent area sample approach, the area segments would be based on Census geographic areas (blocks or block groups) associated with the SSUs. As indicated in Figure 2-1, an area sample to supplement the list sample in the dual frame approach need not necessarily be selected in every sample zip code, but only in a subsample of them designed to provide an unbiased national estimate for the area frame.

All potential building sites in the selected area segments would be identified by a field listing operation. The listers would attempt to match the sites listed with the businesses and their addresses shown in the list source(s) as located in the segment. Sites listed that do not appear in the list source would constitute the area sample supplement. These listings or subsamples of them would be the fourth stage units. For the moment, it is assumed that businesses at sites that are represented in the list frame would not be considered for interview in future surveys. However, this rule need not necessarily be observed for future surveys.

2.2 Estimates from the Survey

The survey will be used to estimate statistics such as the following examples:

1. Aggregates such as the number of small businesses or employment or sales; and
2. Ratios such as the proportion of small businesses that have particular characteristics (based on either all or only selected types of small businesses), an average per firm, or the ratio of two variables of interest such as sales per employee or the ratio of sales to debt.

Let $X$ denote the total value of some characteristic of interest for small businesses, and $x'$ the simple inflation estimate of $X$ from a survey. Let $R$ denote the ratio of two variables, say $X$ and $Y$ for small businesses. Then, we can express $R$ as the ratio

$$ R = \frac{X}{Y} $$

and its estimator from a survey as

$$ r = \frac{x'}{y}. $$
We assume a survey design with \( G \) strata of which \( G_S \) are self representing, and \( G_N \) are nonself representing with one sample PSU per stratum selected with probability proportional to size. Let \( P_{hi} \) denote the probability of selection of the \( i \)th PSU, say PSU\( (hi) \), in stratum \( h \). Within PSU\( (hi) \) assume that there are \( M_{hi} \) SSUs of which \( m_{hi} \) are selected by simple random sampling. Then, for either of the alternative designs, an unbiased estimator of \( X \) may be expressed as

\[
x' = \sum_{h=1}^{G} \sum_{i=1}^{1} \frac{1}{P_{hi}} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} x_{hij} \tag{1}
\]

where \( x_{hij} \) is an unbiased estimator of the total of \( X \), say \( X_{hij} \) for the \( j \)-th SSU in PSU \( (hi) \).

2.2.1 Independent Area Sample

Within SSU \( (hij) \) assume that there are \( T_{hij} \) segments of which \( t_{hij} \) are selected by simple random sampling. Further, within the \( k \)-th segment, say segment \( (hijk) \), assume that there are \( N_{hijk} \) sites listed of which \( n_{hijk} \) are selected by simple random sampling.

Then the estimator \( x_{hij} \) from the independent area sample may be expressed as

\[
x_{Ahij} = \frac{T_{hij}}{t_{hij}} \sum_{k=1}^{N_{hijk}} \frac{n_{hijk}}{n_{hijk}} \sum_{\ell=1}^{n_{hijk}} x_{hijk\ell} \tag{2}
\]

Note that the quantity

\[
\frac{N_{hijk}}{n_{hijk}} \sum_{\ell=1}^{n_{hijk}} x_{hijk\ell} = N_{hijk} \bar{x}_{hijk} \tag{3}
\]

where \( \bar{x}_{hijk} \) denotes the mean value of \( x \) for the sample of businesses in segment \( (hijk) \) is an unbiased estimator of the total of \( X \), say \( X_{hijk} \), for the \( k \)-th segment in SSU \( (hij) \). Similarly, the quantity
\[
\frac{T_{hij}}{t_{hij}} \sum_{t_{hij}} = N_{hijk} \bar{x}_{hijk}
\]

(4)

is an unbiased estimator of the total of \(X\), say \(X_{hij}\), for the \(j\)-th SSU in PSU (hi).

### 2.2.2 Dual Frame Sample

The total of \(X\), say \(X_{hij}\), for the \(j\)-th SSU in PSU (hi) can be expressed as the sum

\[
X_{hij} = X_{Lhij} + X_{Ahij}
\]

(5)

where \(X_{Lhij}\) and \(X_{Ahij}\) are the total values for small businesses represented in the list frame and the supplemental area frame, respectively, in SSU (hij). Then we can, correspondingly, express \(\tilde{x}_{hij}\) as the sum

\[
\tilde{x}_{hij} = \tilde{x}_{Lhij} + \tilde{x}_{Ahij}
\]

(6)

where \(\tilde{x}_{Lhij}\) and \(\tilde{x}_{Ahij}\) are estimators of \(X_{Lhij}\) and \(X_{Ahij}\), respectively.

Within SSU (hij) assume that there are \(N_{Lhij}\) businesses shown in the list source(s) of which \(n_{Lhij}\) are selected by simple random sampling. Then, the estimator \(\tilde{x}_{Lhij}\) may be expressed as

\[
\tilde{x}_{Lhij} = \frac{N_{Lhij}}{n_{Lhij}} \sum_{k=1}^{n_{Lhij}} X_{Lhijk}
\]

(7)

The estimator \(x\) may be expressed, similarly to equation (2) as

\[
x_{Ahij} = \frac{T_{Ahij}}{t_{Ahij}} \sum_{k=1}^{n_{Ahijk}} \frac{N_{Ahijk}}{n_{Ahijk}} \sum_{\ell=1}^{n_{Ahijk}} x_{Ahijk}\ell
\]

(8)
where the quantities in this expression are defined as in (2) but refer only to the supplemental area frame.

2.3 Sampling Variances

As can be seen from equation (1) the sampling variances of survey estimates under either design alternative will have three components corresponding to:

1. The sampling of nonself representing PSUs within strata
2. The sampling of SSUs within PSUs
3. The sampling within SSUs

The structure of the first two components will be the same under either design, since the two alternatives differ only in the design of the sampling within SSUs.

Let \( \text{Var}(x') \) denote the sampling variance of \( x' \), with a subscript 1 to denote the independent area sample approach and 2 to denote the dual frame approach. Then we can write

\[
\text{Var}(x_1') = \text{Var}_1 + \text{Var}_2 \pm \text{Var}_{13} \quad (9)
\]

and

\[
\text{Var}(x_2') = \text{Var}_1 + \text{Var}_2 \pm \text{Var}_{23} \quad (10)
\]

where \( \text{Var}_1 \) denotes the contribution of sampling at the first stage, \( \text{Var}_2 \) the contribution of sampling at the second stage, and \( \text{Var}_{13} \) the contribution at the third stage.

In the following text, detailed equations are presented for each of these components so that as survey data become available they may be used to estimate these components to help guide improvements in the survey design.

\( \text{Var}_1 \) arises only from the nonself representing strata of PSUs. It is
where \( S_h^2 \) is the variance of \( x \) between PSUs in stratum \( h \), and is

\[
S_h^2 = \frac{D_h}{\sum_{i=1}^{D_h} \frac{P_{hi}}{f_{hi}}} \left( \frac{X_{hi}}{P_{hi}} - X_h \right)^2
\]  

(12)

where \( D_h \) denotes the number of PSUs in stratum \( h \).

\( \text{Var}_2 \) arises from the sampling within all PSUs, both self representing and nonself representing. It is

\[
\text{Var}_2 = \frac{G}{\sum_{h=1}^{G} \frac{D_h}{\sum_{i=1}^{D_h} \frac{P_{hi}}{M_{hi}}} \left( 1 - f_{hi} \right) S_{hi}^2}
\]

(13)

where \( f_{hi} \) is the sampling rate of SSUs within PSU \((hi)\)

\[
f_{hi} = \frac{m_{hi}}{M_{hi}}
\]

(14)

and \( S_{hi}^2 \) is the variance of \( x \) between SSUs in PSU\((hi)\) with

\[
S_{hi}^2 = \frac{1}{M_{hi} - 1} \sum_{j=1}^{M_{hi}} (X_{hij} - \bar{X}_{hi})^2
\]

(15)

and

\[
\bar{X}_{hi} = X_{hi}/M_{hi}
\]

(16)

= average value of \( x \) per ZIP area in PSU\((hi)\)

Note that for the self representing PSUs \( D_h = 1 \) and \( P_{hi} = 1 \).
2.3.1 Independent Area Sample

In the independent area sample approach \( \text{Var}_{13} \) is the sum of two components, say \( \text{Var}_{131} \) and \( \text{Var}_{132} \). \( \text{Var}_{131} \) arises from the sampling of segments within SSUs and is

\[
\text{Var}_{131} = \sum_{h=1}^{G} \sum_{i=1}^{D_h} \frac{1}{P_{hi}} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{M_{hi}} \frac{T_{hij}^2}{t_{hij}} (1 - f_{hij}) S_{1hij}^2
\]

where \( f_{hij} \) is the sampling rate of segments within SSU(hij)

\[
f_{hij} = \frac{t_{hij}}{T_{hij}}
\]

and \( S_{1hij}^2 \) is the variance of \( X \) between area segments in SSU(hij) with

\[
S_{1hij}^2 = \frac{1}{T_{hij} - 1} \sum_{k=1}^{T_{hij}} (X_{hijk} - \bar{X}_{1hij})^2
\]

and

\[
\bar{X}_{1hij} = \frac{X_{1hij}}{T_{hij}}
\]

= average value of \( X \) per area segment in SSU(hij)

\( \text{Var}_{132} \) arises from the sampling of small businesses within segments and is

\[
\text{Var}_{132} = \sum_{h=1}^{G} \sum_{i=1}^{D_h} \frac{1}{P_{hi}} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{M_{hi}} \frac{T_{hij}}{t_{hij}} \sum_{h=1}^{N_{hijk}} \frac{N_{hijk}^2}{n_{hijk}} (1 - f_{hijk}) S_{2hijk}^2
\]

where \( f_{hijk} \) is the sampling rate of small businesses within segment (hijk)

\[
f_{hijk} = \frac{n_{hijk}}{N_{hijk}}
\]

and \( S_{2hijk}^2 \) is the variance of \( X \) between small businesses within segment (hijk) with
\( S_{1hijk}^2 = \frac{1}{N_{hijk} - 1} \sum_{l=1}^{N_{hijk}} (X_{hijkl} - \bar{X}_{1hijk})^2 \)  

and

\( \bar{X}_{hijk} = \frac{X_{hijk}}{N_{hijk}} \)  

= average value of \( X \) per small business in segment \((hijk)\)

2.3.2 Dual Frame Sample

In the dual frame approach \( V_{23} \) is the sum of three components: one from the sampling of businesses from the list frame, say \( V_{231} \); and one from the sampling of the supplemental area sample, say \( V_{232} \) that itself has two components, say \( V_{2321} \) and \( V_{2322} \), which parallel the components of the independent area sampling within SSUs.

\( V_{231} \), the contribution of the third stage sampling of businesses from the list frame, is

\[
V_{231} = \sum_{h=1}^{G} \sum_{i=1}^{D_h} \frac{1}{P_{hi}} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{M_{hi}} \frac{N_{hij}^2}{n_{hij}} (1 - f_{2hij}) S_{2hij}^2
\]

where \( f_{2hij} \) is the sampling rate of businesses from the list frame in SSU \((hij)\)

\( f_{2hij} = \frac{n_{hij}}{N_{hij}} \)  

and \( S_{2hij}^2 \) is the variance of \( X \) between list frame businesses within SSU \((hij)\) with

\[
S_{2hij}^2 = \frac{1}{N_{hij} - 1} \sum_{j=1}^{N_{hij}} (X_{hij} - \bar{X}_{2hij})^2
\]

and

\( \bar{X}_{2hij} = \frac{X_{2hij}}{N_{hij}} \)  

= average value of \( X \) per list frame business in SSU \((hij)\)
The components $\text{Var}_{2321}$ and $\text{Var}_{2322}$ have formulas that parallel those for $\text{Var}_{131}$ and $\text{Var}_{132}$, respectively, but referring to segments in the area sample supplement.
3. OPTIMIZATION OF THE DESIGNS

3.1 Overview

The optimization of a sample design has the goal of achieving a desired precision at least cost or, alternatively, greatest precision at a fixed cost. Since there is little evidence currently available as to unit costs of the sampling and survey operations or the variance factors involved, the analysis here will serve primarily to provide guidelines to reasonable national survey designs.

In a multistage sample design, optimization is achieved by optimizing the sampling at each of the successive stages, beginning with the last stage. If this is carried through to the PSU level the number of PSUs that can be included in the sample is then determined by the cost per PSU and the specification of the total cost target. Therefore, in the independent area sample approach we will start with the sum of the within-PSU components in equation (9) of Section 2.

\[ \text{Var}_2 + \text{Var}_{13} \]

and for the dual frame sample the sum of the components in equation (10)

\[ \text{Var}_2 + \text{Var}_{23} \]

We assume that the primary sampling units are fixed in advance by the national sample selected and not subject to optimum determination.

3.2 Independent Area Sample

In this approach the within-PSU sampling represents a three-stage design in which the first stage is the sampling (within sample PSUs) of ZIP code areas. The second stage is the sampling of area segments within selected ZIP code areas and the third stage is the sampling of small businesses within selected segments. For purposes of the optimization, it is convenient to work with the relvariance of an estimated PSU total from this design, \( X_{hi}^2 \), rather than the variance itself. The relvariance is the ratio

\[ \text{Var}(X_{hi}^2)/X_{hi}^2 \]

(1)
where \( \hat{X}_{hi} \) is estimator of \( X_{hi} \). The within PSU component of the sampling relvariance of the estimated total for a PSU, on an average per PSU basis, can be approximated by the following equation:

\[
\frac{\sigma^2}{m t \bar{n}} \left[ \delta_2 \bar{t} \bar{n} + 1 + \delta_3 (\bar{n} - 1) + a \right]
\]

where

- \( \sigma^2 \) = relvariance of small businesses within PSU
- \( m \) = average number of sample ZIP codes per PSU
- \( \bar{t} \) = average number of segments in the sample per ZIP code
- \( \bar{n} \) = average number of small businesses in the sample per segment
- \( \delta_2 \) = intraclass correlation of small businesses within ZIP codes
- \( \delta_3 \) = intraclass correlation of small businesses within segments
- \( a \) = a term that reflects the variability in size and subsampling rates of segments

The factor in square brackets in (2) represents the design effect due to clustering of the sample within PSU.

The following simple cost function reflects the variable cost of a three-stage design such as that within PSU

\[
c = c_1m + c_2m \bar{t} + c_3m \bar{t} \bar{n}
\]

where

- \( c_1 \) = variable cost per ZIP code in sample
- \( c_2 \) = variable cost per segment selected in a ZIP code
- \( c_3 \) = variable cost per interview for a small business selected in a segment

\[\text{1 See for example, "Sample Survey Methods and Theory" by Hansen, Hurwitz and Madow, Vol. I, Ch. 6}\]
\( c_1 \) includes the cost of sampling and of obtaining maps of the boundaries of the selected ZIP code areas. \( c_2 \) includes the cost of coding the commercial list source addresses in the selected ZIP codes to Census geography, establishing the segment boundaries, selecting the sample segments and listing businesses in the sample segments for the sampling of businesses for interview. \( c_3 \) includes the cost of interviewing to achieve a completed interview, as well as the necessary callbacks to achieve a satisfactory response rate.

Following standard theory, from (2) and (3), the optimum value of the design parameter \( \pi \) is given by the equation

\[
\frac{\pi}{\pi} = \sqrt{\frac{c_2}{c_3}} \frac{1 - \delta_3 + a}{\delta_3}
\]

and the number of segments per ZIP code by the equation

\[
\frac{\pi}{\pi} = \sqrt{\frac{c_1}{c_2}} \frac{\delta_3}{\delta_2}
\]

We assume segments from which the \( \pi \) will be sampled will average 60 listings. To be conservative, we take \( \delta_3 = 0.10 \) and \( \delta_2 = 0.0125 \), and based on experience in other surveys we take \( a = 0.5 \).

For cost factors we assume 3 minutes per listing for the sample segments, consistent with the experience in the pilot study, and a 45 minute interview. With the other cost factors involved we take

\[
c_2/c_3 = 210/45 = 4.7
\]

and

\[
c_1/c_2 = 120/210 = 0.57
\]

Then

\[
\frac{\pi}{\pi} = \sqrt{(13.5)(4.7)} = \sqrt{63.45} = 8
\]

and
\[ \frac{^\wedge}{l} = \sqrt{(8)(.57)} = \sqrt{4.56} = 2 \]

Accordingly, the independent area sample design would sample 8 small businesses per segment and 2 segments per ZIP code area. Assuming a target of 2,000 completed interview this would imply

\[ \frac{2,000}{8} = 250 \text{ segments and 125 ZIP code areas.} \]

The target of 2,000 completed interviews was selected as a moderate size sample for a first survey, that still would provide national estimates of most characteristics of small business with acceptable reliability. (Note that to achieve an average of 8 completed interviews, a somewhat larger sample should be selected to provide for noninterviews. For example, an 80 percent completion rate would require an average of 10 businesses per segments designated for the sample.)

A sample of 2,000 small businesses would provide adequate reliability not only for analysis of all small businesses, but also for 5 or 6 subclasses of approximately equal size. The subclasses could be SIC groups, regional totals, etc. For example, a 5-way breakdown would result in approximately 400 businesses per subclass being analyzed. This would provide a standard error of between about 1.5 and 2 percent on an estimate of a characteristic that amounts to 10 percent of all businesses in that subclass. Similarly, a 50 percent estimate would be subject to a standard error of about 2.5 to 3 percent. We consider these levels of precision adequate, but the higher sampling errors arising from smaller sample sizes makes the value of analyses of smaller subclasses somewhat dubious.

It should be recognized that although the sample will provide reasonable estimates of the number of small businesses owned and operated by minorities, women etc., estimates of characteristics of these businesses may be subject to large sampling errors. This should not discourage analysts from carrying out such analyses since the data may provide useful insights on special problems these businesses face, in spite of large sampling errors. Analysts, however, should be aware of the possible limitations of the data.
3.3 Dual Frame Sample

The design of a dual frame sample requires a slightly different analysis. We start with the within-ZIP code component Var23. We can express Var23 in relvariance form as

\[
\frac{P_L^2 V_L^2}{n_L} + \frac{P_A^2 V_A^2}{n_A}
\]  

(6)

where

\[P_L = \text{proportion of small businesses represented in the list source}\]

\[P_A = \text{proportion of small businesses represented by the area sample supplement}\]

\[n_L, n_A = \text{the sample sizes from the two sample sources}\]

The sample within a ZIP code may be viewed as a stratified sample with two strata: (1) the list frame and (2) the area sample frame. The first term in (6) represents the relvariance component arising from the list stratum and the second term the component from the area sample stratum.

Let

\[C_L = \text{variable cost per interview for a list sample case}\]

\[C_A = \text{variable cost per interview for a sample case not represented in the list source}\]

\[C_L\] includes the cost of locating a list source business site and interviewing the current business occupant. \[C_A\] includes the cost of listing a segment, and matching the listings to the list source frame to identify any small businesses not in the list source frame.

In accordance with standard theory for stratified sampling, the optimum allocation of the dual frame sample of a ZIP code is to take \(n_L\) proportional to

\[
\frac{P_L V_L}{\sqrt{C_L}}
\]  

(7)

and \(n_A\) proportional to
\[ \frac{P_AV_A}{\sqrt{C_A}} \]  

so

\[ \frac{n_L}{n_A} = \frac{P_LV_L}{P_AV_A} \sqrt{\frac{C_A}{C_L}} \]  

Since this list sample is not clustered within ZIPS, while the area frame sample is a two-stage sample within ZIPS, we can write approximately

\[ V^2_L = V^2 \]  

\[ V^2_A = V^2 (DEF)_A \]

where \( V^2 \) is the relvariance between small businesses within a ZIP and \( (DEF)_A \) is the design effect for the area sample supplement (of equation (2) in this section). Then we can express the ratio (9) as

\[ \frac{n_L}{n_A} = \frac{P_L}{P_A} \frac{1}{\sqrt{DEF}_A} \sqrt{\frac{C_A}{C_2}} \]  

It is reasonable to estimate that the list frame might cover 75 percent of small businesses and the area frame 25 percent. Then

\[ PL/PA = 3 \]

If the same segment design is used as described for the independent area sample, segments of 8 businesses would average \( \frac{1}{4}(8) = 2 \) businesses not represented in the list frame. Also, it is reasonable to estimate that in a list frame sample perhaps 20 percent might be lost due to post office addresses, business sites no longer in existence, and other reasons. Therefore the list frame sample should designate 25 percent more business than the target to provide for losses.

Then with the cost factors used we take

\[ C_A/C_L = 1.6 \]
\[ \sqrt{\frac{C_A}{C_L}} = 1.3 \]

Also, with only an average of 2 businesses added per segment \((\text{DEF})_A\) would be 1.10 and \(\sqrt{\text{(DEF)}_A}\) would be 1.05. So

\[ \frac{\hat{n}_L}{\hat{n}_A} = (3)(\frac{1}{1.05})(1.3) = 3.7 \]

Then the area frame sample proportion of a total allocation would be

\[ \frac{\hat{n}_A}{\hat{n}_A + \hat{n}_L} = \frac{1}{1 + 3.7} = \frac{1}{4.7} = 0.21 \]

or 20 percent, say, with the list frame accounting for 80 percent.

With a total of 2,000 interviews, then, 1,600 would come from the list frame and about 400 from the area frame.

### 3.4 Relative Costs of the Two Alternative Designs

As a guide to the SBA, Figure 3.1 shows rough estimated costs for a 2,000 interview sample survey under each of the two designs.

These indicate a 13 percent savings in the cost for the same sample size with a dual frame design. If the two designs were matched as to the sampling errors of survey estimates, the smaller sample required with the dual frame design would create additional savings. Again, these figures should be used only as guides.

After the first survey, much more information will become available for estimating design parameters, and the sample design can be refined.
We also note that a dual frame sample provides the flexibility for oversampling kinds of business that are identified on the business lists. This obviously includes SIC, region of the country, and indicators of business size. This added flexibility is another advantage of a dual frame design.

<table>
<thead>
<tr>
<th>Type</th>
<th>Single frame*</th>
<th>Dual frame**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listing/sample design: fixed</td>
<td>$239,800</td>
<td>$156,600</td>
</tr>
<tr>
<td>Listing/sample design: non-fixed</td>
<td>171,700</td>
<td>32,800</td>
</tr>
<tr>
<td>Listing/sample design: Total</td>
<td>411,500</td>
<td>189,300</td>
</tr>
<tr>
<td>Interviewing/processing: fixed</td>
<td>336,100</td>
<td>362,300</td>
</tr>
<tr>
<td>Interviewing/processing: non-fixed</td>
<td>363,000</td>
<td>413,700</td>
</tr>
<tr>
<td>Interviewing/processing: total</td>
<td>699,100</td>
<td>776,000</td>
</tr>
<tr>
<td>Total fixed:</td>
<td>575,900</td>
<td>518,800</td>
</tr>
<tr>
<td>Total non-fixed</td>
<td>534,700</td>
<td>446,500</td>
</tr>
<tr>
<td>Total</td>
<td>1,110,600</td>
<td>965,300</td>
</tr>
</tbody>
</table>

*Assumes 2,000 interviews and 15,000 listed lines

**Assumes 1,600 “list” interviews and 400 “area” interviews

Figure 3-1. Summary of rough costs estimates for a survey of small businesses: Single frame vs. dual frame approach
4. FINDINGS AND RECOMMENDATIONS

4.1 Alternative Designs Considered

- Simple and inexpensive sampling frames that provide satisfactory coverage of almost all small business are not available. For a variety of reasons described in Section 1.3 of this report, U.S. Government agencies samples of businesses or of buildings containing businesses cannot be provided to SBA, nor can frames compiled for various administrative purposes. Commercial lists of businesses, although useful for many purposes, are not sufficiently complete in their coverage of small businesses. Their coverage is particularly low for new small businesses. It is clear that other methods of creating sampling frames are necessary to fulfill SBA's needs.

- Area samples to identify businesses have been used successfully and therefore constitute a plausible sample design for the SBA surveys. However, they are quite expensive and lack the ability to oversample selected segments of businesses.

- A dual frame survey consisting of an integrated commercial list and area sample has been tried in a field test, as part of this research. Westat's report on the field test is attached as an appendix to this report. The methodology tested included having the list represent a set of business addresses, rather than the businesses themselves. This eliminates the need to find the new location of businesses that may have moved from their original sites, as well as giving the businesses that replaced the movers a chance of selection from the list sample. It represents the first known attempt to develop a nationwide sample of businesses through a sample of business locations. It avoids many of the problems created by treating the businesses selected from a commercial list as a sample of businesses.

- The test indicated that such a dual frame sample design was practical, and it could provide the framework for a national probability business sample.

4.2 Design Specifications

- The research developed parameters for the two alternative sample designs considered - an independent area sample and a dual frame design consisting of an integrated list and area sample. Specifications for the sample, and sampling and subsampling units and sizes that would provide efficient sample designs for the two alternatives considered were developed. Operating characteristics and costs of the two samples were compared.

- On the basis of their comparison we recommend a dual frame sample design for SBA's planned surveys. Considering SBA's current resources and that this would be the first survey conducted in this manner, we recommend use of a moderate sample size. This would ensure that the first survey would not exceed the available budget. It would also provide better information on
costs than is currently available, permitting better planning of future surveys.

- We suggest a sample of 2,000 small businesses for the first year. This sample size would provide adequate precision for estimates of nationwide statistics, and for a limited number of subclasses based on industry or selected other characteristics.

- We recommend the specifications for the sample shown below. They should be considered approximations rather than rigid constraints. If minor modifications would simplify the operations, this would be acceptable.

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of PSU's</td>
<td>60 to 70</td>
</tr>
<tr>
<td>Number of Segments</td>
<td>250</td>
</tr>
<tr>
<td>Number of Zip Codes</td>
<td>125</td>
</tr>
<tr>
<td>Number of Completed</td>
<td></td>
</tr>
<tr>
<td>Interviews</td>
<td>2,000</td>
</tr>
<tr>
<td>From List Sample</td>
<td>400</td>
</tr>
<tr>
<td>From Area Sample</td>
<td>1,600</td>
</tr>
</tbody>
</table>

- Considerable efforts should be made to achieve as high response rates as possible. This should include a reasonable number of call backs for owners and managers who are frequently not available for interviews, letters from high level SBA officials indicating the importance of the survey, etc.

4.3 Future Research

The design specifications described above are based on fairly limited information on costs and variances, and on a small field test to study the feasibility of a dual frame design. After the first survey has been completed, much better information will be available which can be used to refine the design specifications. This involves some advance planning, as well as putting aside a small amount of the budget for these activities, mostly to be used at the completion of the first year's survey. The most important subjects to be covered in the research are:

- Estimates of components of variances; this can be prepared after a clean data tape becomes available. Some minor planning to insure that the data tape contains the necessary geographic codes is all that has to be done earlier.

- Estimates of unit costs; this requires keeping record of costs in fair detail. Past experience indicates that it is probably not possible to keep exact records of unit costs for all elements of the survey operations. As much detail as feasible should be kept.

- Possible problems in matching the list sample with what is found in listing an area segment; this is not an issue relating to sample design specifications, but the result could influence the survey procedures and the interviewers' training. Interviewers should be required to document any matching problems encountered, describing the issues in reasonable detailed form. It
would probably be useful to ask interviewers for comments on whether there were any difficulties in matching at the completion of field work.

- Businesses not readily visible in listing area segments; interviewers should be required to document whether any businesses were difficult to recognize and describe each such case encountered. This should be done separately for each segment listed. We note that businesses operated out of individual homes are not in scope for this survey, but other hard-to-recognize business locations may exist. Interviewers' experiences could be helpful in alerting the staff on how to improve the implementation of future surveys.
APPENDIX

Illustrative Intraclass Correlations for ZIP Codes Within PSU for the Statistic: Number of Small Businesses by SIC Division
Notes

The intraclass correlations are denoted as "DEFF STATISTICS"

The PSUs are illustrative cases that might be found in a 60 PSU national design as follows:

- 20 PSUs are examples of self representing large metropolitan areas (1-20)
- 26 PSU are examples of nonself representing smaller metropolitan areas (21-46)
- 14 PSUs are examples of nonself representing nonmetropolitan areas

The counts of numbers of businesses were developed from a commercial list source.
<table>
<thead>
<tr>
<th>PSU</th>
<th>DEFF_B</th>
<th>DEFF_C</th>
<th>DEFF_D</th>
<th>DEFF_E</th>
<th>DEFF_F</th>
<th>DEFF_G</th>
<th>DEFF_H</th>
<th>DEFF_I</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.000965</td>
<td>-0.000643</td>
<td>0.007341</td>
<td>-0.002949</td>
<td>0.011142</td>
<td>0.00327</td>
<td>0.002987</td>
<td>0.01527</td>
</tr>
<tr>
<td>2</td>
<td>-0.001674</td>
<td>-0.001018</td>
<td>0.00536</td>
<td>-0.00722</td>
<td>-0.000444</td>
<td>0.00122</td>
<td>-0.001139</td>
<td>0.00046</td>
</tr>
<tr>
<td>3</td>
<td>-0.002439</td>
<td>-0.001991</td>
<td>-0.001742</td>
<td>-0.002305</td>
<td>-0.001795</td>
<td>-0.00113</td>
<td>-0.001196</td>
<td>-0.00035</td>
</tr>
<tr>
<td>4</td>
<td>-0.002984</td>
<td>-0.002528</td>
<td>-0.002576</td>
<td>-0.002843</td>
<td>-0.002517</td>
<td>-0.00167</td>
<td>-0.002205</td>
<td>-0.00055</td>
</tr>
<tr>
<td>5</td>
<td>-0.003247</td>
<td>-0.001871</td>
<td>-0.001910</td>
<td>-0.002211</td>
<td>-0.001749</td>
<td>-0.00113</td>
<td>-0.001417</td>
<td>0.00060</td>
</tr>
<tr>
<td>6</td>
<td>-0.004766</td>
<td>-0.003683</td>
<td>-0.004233</td>
<td>-0.004492</td>
<td>-0.003733</td>
<td>-0.00179</td>
<td>-0.003206</td>
<td>0.00117</td>
</tr>
<tr>
<td>7</td>
<td>-0.002272</td>
<td>-0.001389</td>
<td>-0.000929</td>
<td>-0.001558</td>
<td>-0.000876</td>
<td>-0.00001</td>
<td>-0.001232</td>
<td>0.00209</td>
</tr>
<tr>
<td>8</td>
<td>-0.001409</td>
<td>-0.001118</td>
<td>-0.000848</td>
<td>-0.001207</td>
<td>-0.000780</td>
<td>-0.00051</td>
<td>-0.000724</td>
<td>0.00125</td>
</tr>
<tr>
<td>9</td>
<td>-0.001965</td>
<td>-0.001370</td>
<td>-0.001061</td>
<td>-0.001759</td>
<td>-0.001168</td>
<td>-0.00064</td>
<td>-0.000990</td>
<td>0.00091</td>
</tr>
<tr>
<td>10</td>
<td>-0.003279</td>
<td>-0.002306</td>
<td>-0.002485</td>
<td>-0.002894</td>
<td>-0.001948</td>
<td>-0.00132</td>
<td>-0.001539</td>
<td>0.00096</td>
</tr>
<tr>
<td>11</td>
<td>-0.001666</td>
<td>-0.000561</td>
<td>-0.00127</td>
<td>-0.001315</td>
<td>-0.000344</td>
<td>0.00077</td>
<td>0.000473</td>
<td>0.00345</td>
</tr>
<tr>
<td>12</td>
<td>-0.004569</td>
<td>-0.001546</td>
<td>-0.001904</td>
<td>-0.002645</td>
<td>-0.001466</td>
<td>-0.00094</td>
<td>-0.001177</td>
<td>0.00179</td>
</tr>
<tr>
<td>13</td>
<td>-0.002064</td>
<td>-0.0002127</td>
<td>-0.002672</td>
<td>-0.002819</td>
<td>-0.002560</td>
<td>-0.00148</td>
<td>-0.001864</td>
<td>0.00436</td>
</tr>
<tr>
<td>14</td>
<td>-0.001422</td>
<td>-0.0001382</td>
<td>-0.001350</td>
<td>-0.001709</td>
<td>-0.000696</td>
<td>0.00004</td>
<td>0.000734</td>
<td>0.00064</td>
</tr>
<tr>
<td>15</td>
<td>-0.002421</td>
<td>-0.0001342</td>
<td>-0.000648</td>
<td>-0.002067</td>
<td>-0.000597</td>
<td>-0.00066</td>
<td>-0.000813</td>
<td>0.00211</td>
</tr>
<tr>
<td>16</td>
<td>-0.001951</td>
<td>-0.0001642</td>
<td>-0.001711</td>
<td>-0.001472</td>
<td>-0.000227</td>
<td>-0.00066</td>
<td>-0.001126</td>
<td>0.00053</td>
</tr>
<tr>
<td>17</td>
<td>-0.002772</td>
<td>-0.0001230</td>
<td>-0.000256</td>
<td>-0.002241</td>
<td>-0.001686</td>
<td>0.000124</td>
<td>0.000130</td>
<td>0.00045</td>
</tr>
<tr>
<td>18</td>
<td>-0.001325</td>
<td>-0.0001763</td>
<td>-0.000773</td>
<td>-0.001263</td>
<td>-0.000788</td>
<td>-0.00043</td>
<td>-0.001054</td>
<td>-0.00008</td>
</tr>
<tr>
<td>19</td>
<td>-0.001403</td>
<td>-0.0001254</td>
<td>-0.000773</td>
<td>-0.001263</td>
<td>-0.000788</td>
<td>-0.00043</td>
<td>-0.001054</td>
<td>-0.00008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PSU</th>
<th>DEFF_B</th>
<th>DEFF_C</th>
<th>DEFF_D</th>
<th>DEFF_E</th>
<th>DEFF_F</th>
<th>DEFF_G</th>
<th>DEFF_H</th>
<th>DEFF_I</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>-0.0096113</td>
<td>0.019118</td>
<td>0.008829</td>
<td>-0.001544</td>
<td>0.016143</td>
<td>0.13242</td>
<td>0.019851</td>
<td>0.16220</td>
</tr>
<tr>
<td></td>
<td>-0.0035595</td>
<td>0.014254</td>
<td>0.004422</td>
<td>0.004084</td>
<td>0.012034</td>
<td>0.09261</td>
<td>0.028928</td>
<td>0.13471</td>
</tr>
<tr>
<td></td>
<td>-0.0077864</td>
<td>0.009309</td>
<td>-0.000909</td>
<td>-0.003521</td>
<td>-0.002105</td>
<td>0.04534</td>
<td>0.004006</td>
<td>0.08390</td>
</tr>
<tr>
<td>60</td>
<td>-0.0070229</td>
<td>0.047160</td>
<td>0.043364</td>
<td>0.013285</td>
<td>0.017833</td>
<td>0.13073</td>
<td>0.032826</td>
<td>0.20001</td>
</tr>
</tbody>
</table>
Plot of DEFF_B*R_B. Symbol used is '*'.

NOTE: 2 obs hidden.
Plot of DEFF_C*R_C. Symbol used is '*'.

NOTE: 1 obs hidden.
Plot of DEFF_D*R_D. Symbol used is '*'.

LARGEST METROPOLITAN AREAS
DIVISION D

Plot of DEFF_E*R_E. Symbol used is '✳'.

LARGEST METROPOLITAN AREAS
DIVISION E

Plot of DEFF_F*R_F. Symbol used is '*'.

NOTE: 1 obs hidden.
Plot of $\text{DEFF}_G \times R_G$. Symbol used is 'e'.

0.004

0.003

0.002

0.001

0.000

-0.001

-0.002

0.18 0.20 0.22 0.24 0.26 0.28 0.30 0.32 0.34

$R_G$
Plot of DEFF_H*R_H. Symbol used is ‘*’.
LARGEST METROPOLITAN AREAS
DIVISION I

Plot of DEFF_1^R_1. Symbol used is '•'.

R_1
SMALLER METROPOLITAN AREAS
DIVISION B

Plot of DEFF_B*R_B. Symbol used is '*'.

NOTE: 1 obs had missing values. 1 obs hidden.
Plot of DEFF_C*R_C. Symbol used is '*'.

NOTE: 1 obs hidden.
Plot of DEFF_D*R_D. Symbol used is '★'.

-0.005 -0.003 0.000 0.003 0.005 0.008 0.010 0.013 0.015 0.018

0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11

R_D
Plot of DEFF_E*R_E. Symbol used is '*'.

DEFF_E
0.000
-0.001
-0.003
-0.004
-0.005

R_E
0.020 0.025 0.030 0.035 0.040 0.045 0.050 0.055 0.060 0.065
Plot of DEFF_F^R_F. Symbol used is '*'.
Plot of DEFF_G*R_G. Symbol used is '*'.
AREA SAMPLE COMPARISON FORM  
SBA FIELD TEST  

CIRCLE  
ZIP: 20006  
20007  
20705  
20910  
20912  
22024  

SAMPLE #: ___ OF ___ LISTER: _________  
LINE #: ___ DATE: ___________  
NAME: _________________________  
ADDRESS: _______________________

RESULTS OF COMPARISON OF SAMPLED LISTING LINES WITH DMI AND TRW  

PLACE A CHECKMARK BY THE RESULT BELOW WHICH BEST DESCRIBES WHAT HAPPENED WHEN YOU ATTEMPTED TO RECONCILE THE SAMPLED BL YOU LISTED WITH THE DMI AND TRW LISTS.  

<table>
<thead>
<tr>
<th>DMI</th>
<th>TRW</th>
</tr>
</thead>
</table>
|     |     | 1. MATCH: Name and specific address match.  
|     |     | 2. MATCH: Specific address match, but name is different (verified with new occupant).  
|     |     | 3. MATCH: Specific address match but name is different (not verified with new occupant).  
|     |     | 4. MATCH: Address not a specific match but able to match based on name.  
|     |     | 5. NO MATCH: Could not match name or address.  
|     |     | 6. NO MATCH: Address not a specific match and cannot match name.  
|     |     | 7. NO MATCH: BL was on DMI/TRW and would be match but was listed incorrectly.  
|     |     | 8. NO MATCH: Could not match for other reason (SPECIFY BELOW)  

FOR ANY RESULT OTHER THAN "1", DESCRIBE THE SITUATION BELOW. IF YOU CONTACTED ANYONE FOR INFORMATION, BE SURE TO REPORT THE CONTACT BELOW.  

__________________________________________________________________________  
__________________________________________________________________________  
__________________________________________________________________________  
__________________________________________________________________________  
__________________________________________________________________________  
__________________________________________________________________________  
__________________________________________________________________________  
__________________________________________________________________________
LIST SAMPLE LOCATION FORM
SBA FIELD TEST

ZIP: 20006
20007
20705
20910
20912
22024

SAMPLE #: ___ LISTER: _________
LINE #: ___ DATE : __________

NAME: ______________________
ADDRESS: ___________________

LIST: DMI
TRW

RESULTS OF ATTEMPT TO LOCATE BL FROM LIST

PLACE A CHECKMARK BY THE RESULT BELOW WHICH BEST DESCRIBES WHAT HAPPENED WHEN YOU ATTEMPTED TO LOCATE THE LISTING GIVEN ABOVE.

1. LOCATED BL: Name and specific address match.

2. LOCATED BL: Specific address match, but name is different (verified with new occupant).

3. LOCATED BL: Specific address match, but name is different (not verified with new occupant).

4. LOCATED BL: Address not specific but able to identify BL based on name.

5. COULD NOT LOCATE BL: Address on list does not match specific BL and cannot match name.

6. COULD NOT LOCATE BL: Could not find at all/ Address does not exist/ PO Box / Address otherwise Insufficient

7. COULD NOT LOCATE BL: Other reason (SPECIFY BELOW).

FOR ANY RESULT OTHER THAN "1", DESCRIBE THE SITUATION BELOW. IF YOU CONTACTED ANYONE FOR INFORMATION, BE SURE TO REPORT THE CONTACT BELOW.

__________________________________________

__________________________________________

__________________________________________
The following four forms were designed for use in the area sample phase of a national survey. They were not used in the pilot study since the segmenting was done by central office staff, and field staff used the forms shown as Attachments 6 and 7 earlier in this appendix.
<table>
<thead>
<tr>
<th>STREET</th>
<th>SIDE OF STR.</th>
<th>BETWEEN INTERSECTING STREETS</th>
<th>READILY COUNTABLE SITES</th>
<th>SITES THAT ARE DIFFICULT TO COUNT (1 PER LINE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>INDICATE NUMBER AND TYPE</td>
<td>BUILDING</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>INDUSTRIAL</td>
<td>RETAIL</td>
</tr>
</tbody>
</table>
# Survey of Small Businesses

## Block Summary Form

**Indicate Block Outcome:**

<table>
<thead>
<tr>
<th>OFFIC</th>
<th>ENTER BLOCK ONLY #</th>
<th>NO SITES PRESENT</th>
<th>SITES PRESENT:</th>
<th>ONE OR MORE ADDED</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**PSU:***

**SEG#:***

**CITY:***

**STATE:***

**ZIP:***

**Interviewer:**

**Region #:***
<table>
<thead>
<tr>
<th>ONLY BLOCK #</th>
<th>D &amp; B LISTING</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hank's Sporting Goods</td>
</tr>
<tr>
<td></td>
<td>1107 Main St.</td>
</tr>
<tr>
<td>2</td>
<td>Shoes Unlimited</td>
</tr>
<tr>
<td></td>
<td>Midtown Center</td>
</tr>
<tr>
<td></td>
<td>1112 Main St.</td>
</tr>
<tr>
<td>3</td>
<td>Al's Used Books</td>
</tr>
<tr>
<td></td>
<td>Midtown Center</td>
</tr>
<tr>
<td></td>
<td>1112 Main St.</td>
</tr>
<tr>
<td>4</td>
<td>Fosdick's Department Store</td>
</tr>
<tr>
<td></td>
<td>Midtown Center</td>
</tr>
<tr>
<td></td>
<td>1112 Main St.</td>
</tr>
<tr>
<td>5</td>
<td>All-Hours Convenience</td>
</tr>
<tr>
<td></td>
<td>1116 Main St.</td>
</tr>
<tr>
<td>6</td>
<td>Bill's Tap</td>
</tr>
<tr>
<td></td>
<td>505 Miller Rd.</td>
</tr>
<tr>
<td>7</td>
<td>Union Drywalling</td>
</tr>
<tr>
<td></td>
<td>Miller Rd.</td>
</tr>
<tr>
<td>8</td>
<td>John Smith, D.D.S</td>
</tr>
<tr>
<td></td>
<td>104 Nathan St.</td>
</tr>
<tr>
<td></td>
<td>Suite #1</td>
</tr>
<tr>
<td>9</td>
<td>Baylor Accounting Systems</td>
</tr>
<tr>
<td></td>
<td>104 Nathan St.</td>
</tr>
<tr>
<td></td>
<td>Suite #2</td>
</tr>
<tr>
<td>10</td>
<td>Tom Smith, Attorney</td>
</tr>
<tr>
<td></td>
<td>104 Nathan St.</td>
</tr>
<tr>
<td></td>
<td>Suite #3</td>
</tr>
<tr>
<td>11</td>
<td>Weller Manufacturing</td>
</tr>
<tr>
<td></td>
<td>Norton Rd.</td>
</tr>
<tr>
<td>12</td>
<td>Acme Furniture</td>
</tr>
<tr>
<td></td>
<td>204 Pollard St.</td>
</tr>
<tr>
<td>13</td>
<td>Mattress Land</td>
</tr>
<tr>
<td></td>
<td>204 Pollard St.</td>
</tr>
<tr>
<td>14</td>
<td>206 1/2 Pollard St.</td>
</tr>
</tbody>
</table>

**Comments**
<table>
<thead>
<tr>
<th>OFFICE SITE USE BLOCK SEQ.</th>
<th>NAME</th>
<th>STREET ADDRESS</th>
<th>OTHER ADDRESS</th>
<th>DESIGNATION</th>
<th>CODE:</th>
<th>REASON</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONLY # # #</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TYPE:**
1. INDUSTRIAL
2. RETAIL
3. COMMERCIAL
4. OFFICE
5. OTHER

**ADD REASON:**
1. MISSED
2. NEW SITE
3. DON'T KNOW
Plot of $\text{DEFF}_H \times R_H$. Symbol used is '*'.

The x-axis represents $R_H$ ranging from 0.060 to 0.120, and the y-axis represents $H$ ranging from -0.005 to 0.025.
Plot of DEFF_1*R_1. Symbol used is '*'.

NOTE: 1 obs hidden.
Plot of $\text{DEFF}_B \times R_B$. Symbol used is '•'.

NOTE: 1 obs had missing values.
Plot of DEFF_C*R_C. Symbol used is ‘*’. 
Plot of DEFF*D*R.D. Symbol used is 'x'.

NON-METROPOLITAN AREAS
DIVISION D

Plot of DEFF*D*R.D. Symbol used is 'x'.

NON-METROPOLITAN AREAS
DIVISION D
Plot of DEFF * R_E. Symbol used is '•'.

DEFF_E

R_E

0.015
0.010
0.005
0.000
-0.005
-0.010
-0.015
-0.020
0.025 0.030 0.035 0.040 0.045 0.050 0.055 0.060 0.065 0.070

...
Plot of DEFF_F*R_F. Symbol used is '*'.

NON-METROPOLITAN AREAS
DIVISION F
Plot of DEFF_G*R_G. Symbol used is '*'.

NON-METROPOLITAN AREAS
DIVISION G
Plot of DEFF_H*R_H. Symbol used is '*'.

-0.02 -0.01 0.00 0.01 0.02 0.03 0.04
-0.05 0.00 0.05 0.10 0.15 0.20 0.25 0.30

0.040 0.045 0.050 0.055 0.060 0.065 0.070 0.075 0.080 0.085 0.090 0.095

R_H
Plot of \( \text{DEFF}_I \times \text{R}_I \). Symbol used is \('\ast\').