TOOL 5 VALUE CHAIN SAMPLING GUIDELINES

JOBS IN VALUE CHAINS SURVEYS

LET'S WORK
A global partnership to create more & better private sector jobs

WORLD BANK GROUP
Jobs

SAMPLING NOTE

This tool provides an overview of two important techniques that will be critical for implementing the sampling for the 'Jobs in Value Chains' analysis. First, it will discuss Block Enumeration techniques; second it will provide a basic overview of approaches and options for standard sampling, once the population is identified.

BLOCK ENUMERATION: STANDARD METHODOLOGY[1]

Obtain aerial map of city, divide into blocks, and classify blocks using local knowledge

For each city, using an aerial map such as Google Earth Pro, divide the city into "blocks" and then using local knowledge, each block should be classified into strata defined by the predominant spatial use. Blocks should be roughly the same size and ideally should be defined by practical boundaries such as roads and natural geographic phenomena (e.g. a river, etc.).

Sometimes lists may be available for a city and these can be incorporated into the block enumeration process if they will save time for the block enumerator. Lists can include a list of blocks themselves or perhaps lists of establishments (telephone directory, sample frame, lists from Chambers of Commerce, etc.). For example, business "yellow pages" may provide useful contact information such as address and telephone numbers although that particular information may not be as up-to-date as the contact information obtained directly by the enumerator.

Block strata can include:

a. Industrial (manufacturing)

b. Commercial (retail/services)

c. Offices

d. Hotels/Restaurants

e. Mixed without Residential (Industrial/Commercial/Office/Hotels)

f. Mixed with Residential (Industrial/Commercial/Office/Hotels/Residential)

g. Residential

h. Other (Government/Education/Nature/Park)

For Block strata e. and f. which are both Mixed, these categories should be used if it is not easy to determine the predominant spatial use between different non-Residential types (e.) and Residential and non-Residential types (f.). Each city will have a list of blocks, each block is classified as Block strata a, b, c,...or g.

[1] Prepared by the World Bank Enterprise Analysis Unit

The map’s information should be recorded in an Excel spreadsheet, following the example in the chart below:

			Block number	
Block strata	**Neighborhood**	**City**	**From**	**To**
B	AAA	aaa	1	40
A	BBB	ccc	301	475
B	AAA	eee	1001	2005
D	CCC	eee	41	300
D	BBB	eee	476	1000
B	BBB	aaa	178	329

Depending on the size of the city and the distribution of business activities in a city, it may be useful to first categorize the city into broad clusters where each cluster contains many blocks (say each cluster contains 40-100 blocks) and using local knowledge, each cluster is identified as one of the eight types of Block strata. Then by definition, since the cluster is a homogenous composition of blocks pertaining to a specific type of activity, ALL blocks within that cluster are labelled as the Block strata identified with that cluster.

Randomly select 10 pilot blocks from the list among the blocks that are Block strata a-f and undertake full enumeration

Undertaking full enumeration of 10 pilot blocks in each of the cities covered in each region serves three purposes: 1) starts the process of block enumeration, 2) allows one to double-check the classification of blocks in Step 1 for those 20 blocks and may cause reclassifications of other blocks and 3) allows one to have a sense of on average, how many eligible sector-specific establishments exist in each block type. Please refer to the ISIC rev. 3.1 codes (http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=17).

The population of industries to be included in the Enterprise Surveys and Micro Survey, the Universe of the study, includes the following list (according to ISIC, revision 3.1): all manufacturing sectors (group D), construction (group F), services (groups G and H), transport, storage, and communications (group I), and subsector 72 (from Group K).

The pilot will be carried out by a combination of visual surveying on the part of the enumerator along with face-to-face interviews. The questionnaire for the pilot and for the block enumeration carried out in Step 4 of this document is designed to seek the following set of information:

BLOCK INFORMATION:

A. Block number

B. Block type

ADDRESS/CONTACT INFORMATION:

C. Address:

1. Company name on door phone / door / entrance
2. No. apartment/office
3. Neighborhood

P. Telephone

Q. Email

R. Official Company name

BUSINESS CHARACTERISTICS:

D. Type of premises

E. Activity description

F. The response regarding the activity description was: (asked to an employee/asked to another person/enumerator own judgment)

G. Is the main business activity undertaken in these premises?

H. Is the business activity currently ongoing?

I. How many other business structures belong to the company in this neighborhood?

L. Type of firms

M. Ownership

N. Number of workers

O. The response regarding the number of workers was: asked to an employee/asked to another person/enumerator own judgment)

The activity description should be written down exactly as provided by the respondent and/or visual surveying. After enumerating a block, activity descriptions (point E) need to be recoded into two columns- 1) the first column is sector (Manufacturing, Retail, etc.) and 2) the second column is enumerator determination of Formal vs. Informal activity.

A pilot report will be prepared once the pilot is completed. The number, Block strata and sector of eligible establishments should be summarized as follows. In the example table below, the counts in black color are counts at the establishment level, the counts in red are at the block level:

Establishments enumerated and eligible for the survey in 10 blocks

Formal Sectors	Industrial	Commercial	Offices	Hotels/ Restaurants	Mixed without Residential	Mixed with Residential	Total
Manufacturing	4	2	5	6	9	4	30
Retail	16	15	17	12	17	11	88
Other services	10	11	14	17	14	10	76
Total	30	28	36	35	40	25	194
# of blocks enumerated	4	2	4	2	6	2	20

a. Determine how many total blocks should be fully enumerated

For a specific sector, say Retail, let's say the target number of interviews is 120. To allow a response rate of 25% (conservative estimate) enough blocks need to be enumerated such that a list of 480 Retail establishments is available. Based on the average number of Retail establishments in Block strata types b., e. and f. (called *y*), the number of blocks of type b., e. or f. that need to be enumerated is 480 divided by *y*. Enumerating this number of blocks would yield approximately 480 Retail establishments. Similar calculations should be done for other sectors and overlapping sectors within a block should be considered.

Hence for each of the Block strata, enough blocks should be randomly sample to produce at least 4 times the target number of interviews for each survey strata.

b. Randomly select blocks from the list for full enumeration

Once the total number of blocks to be enumerated (distributed among the different Block strata) is determined, full enumeration of these blocks will be undertaken. Similar to the pilot where a few blocks considered as "g. Residential" were enumerated, a small percentage of blocks considered as Residential should be enumerated in order to eliminate any potential bias created by completely ignoring Residential blocks (oftentimes in urban areas Residential blocks can contain Retail establishments).

If the city is large enough so that simple random selection of blocks will yield a subset of blocks that are prohibitively far from each other so as to be extraordinarily expensive for the enumerator, a first stage of random selection of clusters and then a second stage of random selection of blocks with the selected clusters can be performed. This would yield a random selection of blocks within a constrained geographic area (the cluster) so as to help reduce costs and travel time.

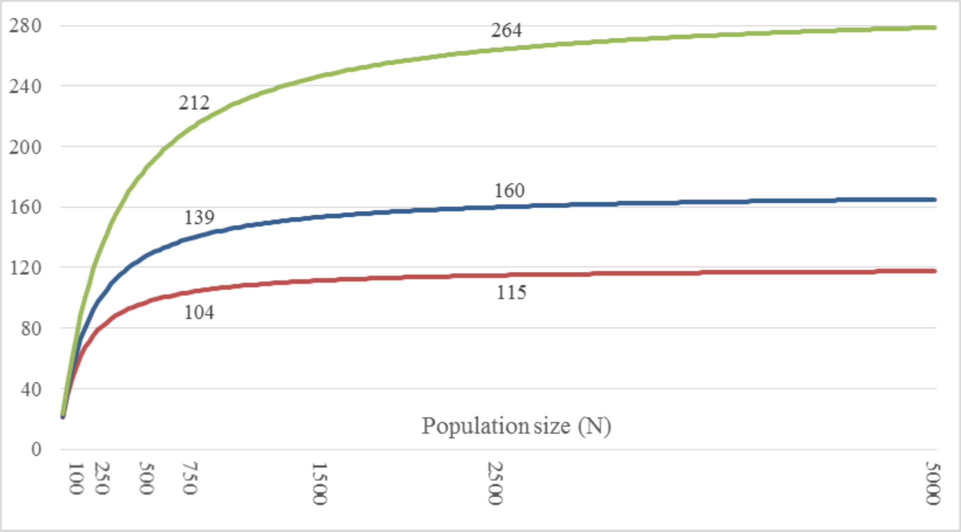
STANDARD RANDOM SAMPLING

Determining sample size

Once we identify the population, we are in a position to determine the required sample size, which depends on the size of the population, how uniform the underlying population is known (or assumed) to be[2], and of course the project's requirements for precision. To give these concepts more concrete examples, consider a stratified sample design, where the underlying population sizes are known. This will frequently be the case for larger and downstream firms in value chains. Assume that the initial budget is not explicitly limited, but rather that the project team is setting out to determine the sampling requirements for reliable and precise estimates.

First, consider a desired level of precision, say 7.5%, and confidence interval of 90%. This means that the project is seeking to obtain figures and indicators which can be said to be within 7.5% of the true – but unknown – population value, with 90% confidence. Secondly, some assumptions about the uniformity (or variance) of the underlying population must be made. For simplicity, consider the case of one dimension (and therefore an indicator based on one question) that is expressed a proportion. Take the question "At the present time does this establishment have its own website?" as a YES/NO question.[3] Here (as is the case with all proportion questions) then the variance is bounded, and in fact, reaches its maximum when exactly half of the population answers "YES" and the other half answers "NO".[4]

Required Sample (7.5% precision) at 90%, 95%, and 99% Confidence



The orange line above shows the required sample sizes for a 7.5% precision level with 90% confidence under these conditions. The x-axis represents the overall population size (N). Conveniently as this population size increases, the required sample converges to 120 and this is underlying principle for

[2] Formally the required sample size (n) for discrete variables is given by: $= [\frac{1}{N} + \frac{N-1}{N} * \frac{1}{PQ} (\frac{k}{z_{.5(1-\alpha)}})^2]^{-1}$. Where N is the size of the relevant population; P is the proportion of the population taking on a certain value; Q = 1-P; k is desired precision level; and z is the z-score for the confidence level $1 - \alpha$. This formula assumes a normal distribution and is expressed as a function of a finite population, N.

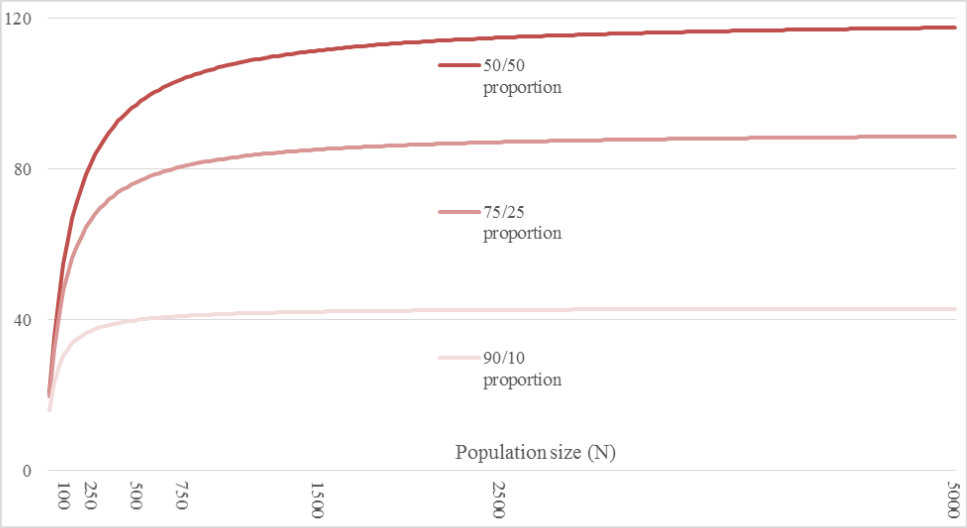
[3] Taken from the World Bank Group's Enterprise Surveys: www.enterprisesurveys.org

[4] In this case P=0.5 and Q=1-P=0.5.

sampling for several WBG projects such as the Enterprise Surveys.[5] What is more, the blue (95%) and grey lines (99%) show that at higher levels of confidence this sampling requirement increases quickly. For example, for a population of 750 a 90% confidence level requires a sample of 104, while a 95% one requires 139, and a sample of 212 is required for 99% confidence. For a population of 2,500, these samples are 115, 160, and 264, respectively. These requirements moreover are multiplicative: that is they must be set out for each analytical cut of interest. For an analysis interested at developing indicators for three levels of the value chain, the varying level of confidence from 90% to 99% more than doubles the sampling requirement, from a maximum of 360 to 885.

This case assumes the maximum variance, however, and it may be reasonable in certain cases to assume more uniformity in the underlying population. The figure above shows the calculated sample size required (for 7.5% and 90% confidence) assuming different proportions of the population. Using the question above for comparison, the dark orange line assumes that the true value in the population at-large of establishments with a website is 50% (and thus 50% answer "NO"); the middle line assumes this proportion is 75% showing that the required sample drops substantially (at its limit to 90), and this continues if 90% is the true proportion (with a required sample of 43).

Required Sample, Population Proportion at 50%, 75%, and 90%



Some additional caveats are necessary. First, the above examples rely on the case of one indicator (based on one question); however, any extensive survey will feature several questions and variables and thus the choice of assumptions for sample size should rely on one or a handful of central questions. It is good practice to utilize a conservative approach. Researchers can rely on previous surveys or administrative sources to provide a priori estimates of the underlying proportions, or when this is not possible, rely on best-available data. Secondly, while the variance on proportion (i.e., categorical)

[5] For details on ES sampling methodology, see: http://www.enterprisesurveys.org/~/media/GIAWB/EnterpriseSurveys/Documents/Methodology/Sampling_Note.pdf

questions is by its structure bounded, the variance on continuous variables is un-bounded.[6] High-variance variables thus, will demand further sample expansion.

For the purposes of this 'jobs in value chains' exercise, we will adopt a strategy of 50/50 proportion with a 7.5% precision and 90% confidence interval.

Finally, note that the sample calculations above present the number of required non-missing responses. That is, a sample requirement of n=120 demands at least that number of observations; however, survey implementation faces challenges of ineligible respondents, survey non-response (i.e., refusal to the survey), and item non-response (i.e., refusal to the question). In such cases, several sampling strategies (or their combination) can be adapted:

- **Realized sample**: utilizes replicates (that is, additional contacts) that are issued and/or made available to implementers until the required sample targets are achieved. A successful strategy for ineligibility and survey non-response, not often practical for item non-response. Outcomes of survey attempts (e.g., successful, ineligible, refusal) should be recorded in detail;
- **Inflation rates:** issues an initial sample, assuming relevant sample loss rates (such as survey non-response, item non-response, ineligibility) that will yield the desired sample targets. That, is relevant inflation rates are built-in to the sample. For example, assume a sample target of 120, an eligibility rate of 85%, and a survey response rate of 50%, and an item response rate of 90%. This gives[7]:

SAMPLE LOSS AND INFLATION

Target interviews	**120**
x [1/ eligibility rate]	
x 1/.85	142
x [1/ response rate]	
x 1/.50	284
x [1/ item response rate]	
x 1/.90	316

[6] For continuous variables, this is $n = [\frac{1}{N} + (\frac{k}{CV * z_{.5(1-\alpha)}})^2]^{-1}$, where CV indicates the coefficient of variation of the variable in question, given by $CV = {}^{\sigma}/_{\mu}$, with δ indicating the standard deviation and μ, the population mean.

[7] Adapted from Valliant, R., J.A. Dever, and F. Kreuter. Practical Tools for Designing and Weighting Survey Samples (New York: Springer, 2013). See pp. 174-177 for further discussion.

This can be a successful strategy for ineligibility, item, and survey non-response, however it may result in a loss of control over final sample distribution, often requiring additional replicates be issued. This can be time-intensive process as well, particularly if item response requirements extend across several variables (in which case it is the joint non-response rate that is relevant);

- **Mixed sampling methodologies***:* the above examples assume a simple one-stage design where a best estimate of the population is known. However, such strategies may not be feasible in cases where the population size is unknown and/or a two- or three-stage sampling strategy is more appropriate. This is likely to be the case for upstream portions of the value chain (e.g. cassava farmers) that can be effectively targeted through primary sampling units (PSU) including villages, markets, geographic clustering, etc. In these cases, researchers should allow for sufficient PSUs in order to obtain a sufficient number of observations (e.g. secondary sampling units [SSUs]). Full consideration should also be made in these cases for appropriate calculations of variance and weighting (see below).

Stratification

In preparing the sample frame, identifying the stratification that will needed to meet the objectives is critical to determine the nature of the sample frame and the size of the overall sample. In the case of value chains, we may want to stratify for a number of things, including: i) the stage of the value chain; ii) firm size category); iii) position inside and outside the established (lead firm) value / supply chain; and, in some cases iv) region within the country (particularly in large countries and where there are distinct value chains operating in different regions).

Note that given the relatively small sample sizes that will exist in many value chains, the sample size will end up being a relatively large proportion of the population.

Weighting

Finally, if a standard approach to sampling is likely to result in a sample that does not reflect accurately the population characteristics (i.e. the sample frame is unable to capture what is known to be true about the general population), then it may be necessary to weight the sample to address this. This may involve, for example, ‘oversampling’ a particular region, or microenterprises, or women-owned enterprises. This can be done in the surveying stage or, if not identified or acted on at that stage, even post-survey by weighting the survey results higher for certain types of units. For example, if microenterprises are known to represent 80% of the population in a value chain, but the survey results only resulted in 50% of responses coming from microenterprises, the results from microenterprises could be adjusted by weighting each of them the equivalent of 1.6 times those of other firms (0.80/0.50=1.6).

BOX 1: DETAILED EXAMPLE OF SAMPLING STRATEGY CONSIDERATIONS FOR A BASIC VALUE CHAIN

To consider a practical example, consider a very streamlined and simple value chain with three key stages: say, fish processing. These include (1) a large relatively uniform population of primary input suppliers, that is fishermen; (2) a relatively variant and medium-sized population of fish processing plants; and (3) a large and highly variant population of final distributors and traders, namely retailers and wholesalers. Assume that for a key variable (discrete), for example if the business exports, the proportion for group (1) is YES=0.15 (NO=0.85); for group (2) it is YES=0.6 (NO=0.4); and for group (3) it is YES=0.5 (NO=0.5). Group 1 has a population of 25,000; group 600; and group 3 a population of 2,000. For the sake of this one relevant indicator, assume an item response rate of 90%. The required sample targets (with 7.5% precision at 90% confidence) are given by:

Group		Universe (N)	% YES (P)	% NO (Q)	Precision	z-score	n		Response rate		Inflated n
1	very large, low variance	25,000	0.15	0.85	7.50%	1.645	61		0.9		68
2	medium, medium variance	600	0.60	0.40	7.50%	1.645	97				108
3	large, high variance	2,000	0.50	0.50	7.50%	1.645	113				127
							272				303

Sample Design and Budget Constraints

While the required sample sizes can be calculated using the above steps, in reality budget constraints may limit the achievable sample. The above steps outline preliminary estimates of the required sample sizes to ensure a minimal level of precision within a given confidence interval; working within budget and time constraints, however, a full sample design will need to be developed. There are several ways to accommodate these constraints, the choice of which is subject to the specific constraints and priorities of the project. Three potential methods are presented below, for illustration, extend the example used above, given a budget for 200 total surveys.

- **Simple proportional allocation***:* obtained by simply allocating the total sample available based on their share of the population. This method closely approximates simple random sampling of course, and so sample can likely be without the consideration of sampling weights. It is also useful is standard deviations (if known) are roughly equal across groups but risks insufficient sample (see groups 2 and 3 in the table below).

Group	Universe (N)	Proportion of N	Sample
1	25,000	91%	181
2	600	2%	4
3	2,000	7%	15
	27,600		200

- **Proportional allocation by targets***:* a second, more robust method is to proportionately allocate sample based on the distribution of sample targets. Notably, this method allocates more sample to groups 2 and 3 as their variance increases sample target requirements. This substantially changes the probability of selection for each establishment (element) in each group, and thus for accurate estimates, survey weights should be used. In their simplest form, these weights are expressed as the inverse of the probability of selection, a so-called base weight given by N/n (shown below).[8] This method is commonly employed and in more complex designs (with say two or three-dimensional stratification) can be developed using optimization programs such as Solver in Excel.[9]

Group	Universe (N)	Sample Target	Proportion of target	Sample (n)	Base Weight (N/n)
1	25,000	61	23%	45	554.9
2	600	97	36%	71	8.4
3	2,000	113	42%	84	23.9
		272		200	

- **Cost-Constrained allocation***:* Both the proposed methods above assume that the cost of collecting data is the same across three groups; of course, this may not be the case. A final method (based on what is called the Neyman allocation) determines sample design based on this varying cost, taking into account the underlying variance (which is rarely known beforehand).[10] Consider that surveys for group 1 may be substantially cheaper ($40 unit cost), followed by group 3 ($75) and the most expensive being group 2 ($90). Using proportions based on target samples, the table below gives a slight re-allocation, shifting surveys from (more expensive) group 2 (fish processing) to the cheaper survey (group 1, fishermen):

Group	Universe (N)	Proportion	Unit cost (USD)	Sample (n)	Base Weight (N/n)
1	25,000	23%	40	54	463.0
2	600	36%	90	63	9.5
3	2,000	42%	75	83	24.1
				200	

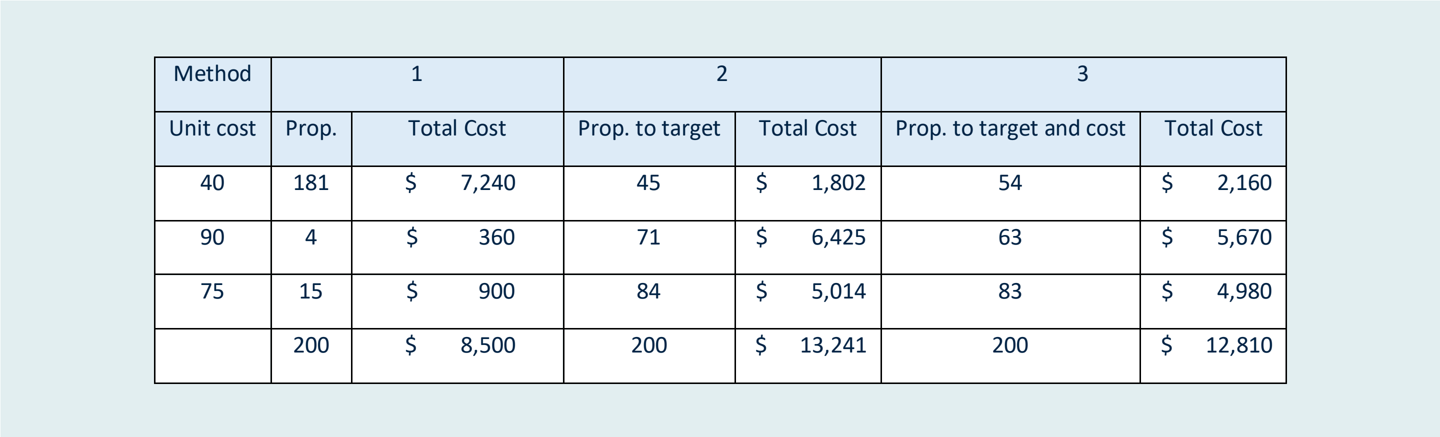
It is worth considering the overall cost implications of these designs. The cheapest option is given by option 1. But it is inefficient, only providing 4 surveys in group 2 and 15 in group 3. To increase observations in group 2 and 3 to a minimal 45 would cost an additional $5,500, which is more expensive than other options. There is a slight cost savings by using method 3, however in this example this is minimal. Project teams should consider the considerable additional information required to implement this method and if cost savings are sufficient, with possibilities of using those savings to allocate more sample.

[8] Note that this measure treats N as the most reliable measure of the population. In reality, however, through implementation, ex post information on the accuracy of these figures may be available. In those cases, and eligibility adjustment, e, that is the proportion of eligible establishments can be applied. This is expressed as

$w_h = \frac{N_h}{n_h} * e_h$ where, h, indicates that these calculations are done over each stratum or group.

[9] For instance, this is the method used by the Enterprise Surveys.

[10] Given by: $\frac{n_h}{n} = \frac{W_h S_h * c_h^{-.5}}{\sum_h^H W_h S_h * c_h^{-.5}}$ where c_h is the cost by group and S_h is the standard deviation and W_h is the share accounted for that stratum.



Method	1		2		3	
Unit cost	Prop.	Total Cost	Prop. to target	Total Cost	Prop. to target and cost	Total Cost
40	181	$ 7,240	45	$ 1,802	54	$ 2,160
90	4	$ 360	71	$ 6,425	63	$ 5,670
75	15	$ 900	84	$ 5,014	83	$ 4,980
	200	$ 8,500	200	$ 13,241	200	$ 12,810