The Guatemala 2017 Enterprise Surveys Data Set

I. Introduction

This document provides additional information on the data collected in Guatemala between October 2017 and May 2018. The objective of the Enterprise Survey is to gain an understanding of what firms experience in the private sector.

As part of its strategic goal of building a climate for investment, job creation, and sustainable growth, the World Bank has promoted improving the business environment as a key strategy for development, which has led to a systematic effort in collecting enterprise data across countries. The Enterprise Surveys (ES) are an ongoing World Bank project in collecting both objective data based on firms' experiences and enterprises' perception of the environment in which they operate.

The ES currently cover over 160,000 firms in 148 countries, of which 139 have been surveyed following the standard methodology. This allows for better comparisons across countries and across time. Data are used to create statistically significant business environment indicators that are comparable across countries. The ES are also used to build a panel of enterprise data that will make it possible to track changes in the business environment over time and allow, for example, impact assessments of reforms.

This report outlines and describes the sampling design of the data, the data set structure as well as additional information that may be useful when using the data, such as information on non-response cases and the appropriate use of the weights.

II. Sampling Structure

The sample for 2017 Guatemala ES was selected using stratified random sampling, following the methodology explained in the *Sampling Note*[1]. Stratified random sampling[2] was preferred over simple random sampling for several reasons[3]:

a. To obtain unbiased estimates for different subdivisions of the population with some known level of precision.

b. To obtain unbiased estimates for the whole population. The whole population, or universe of the study, is the non-agricultural economy. It comprises: all manufacturing sectors according to the group classification of ISIC Revision 3.1: (group D), construction sector (group F), services sector (groups G and H), and transport, storage, and communications sector (group I). Note that this definition excludes the following sectors: financial intermediation (group J), real estate and renting activities (group K, except sub-sector 72, IT, which was added to the population under study), and all public or utilities-sectors.

---

[1] The complete text can be found at http://www.enterprisesurveys.org/~/media/GIAWB/EnterpriseSurveys/Documents/Methodology/Sampling_Note.pdf

[2] A stratified random sample is one obtained by separating the population elements into non-overlapping groups, called strata, and then selecting a simple random sample from each stratum. (Richard L. Scheaffer; Mendenhall, W.; Lyman, R., "Elementary Survey Sampling", Fifth Edition).

[3] Cochran, W., 1977, pp. 89; Lohr, Sharon, 1999, pp. 95

c. To make sure that the final total sample includes establishments from all different sectors and that it is not concentrated in one or two of industries/sizes/regions.

d. To exploit the benefits of stratified sampling where population estimates, in most cases, will be more precise than using a simple random sampling method (i.e., lower standard errors, other things being equal.)

e. Stratification may produce a smaller bound on the error of estimation than would be produced by a simple random sample of the same size. This result is particularly true if measurements within strata are homogeneous.

f. The cost per observation in the survey may be reduced by stratification of the population elements into convenient groupings.

Three levels of stratification were used in this country: industry, establishment size, and region. The original sample design with specific information of the industries and regions chosen is described in Appendix C.

Industry stratification was designed in the way that follows: the universe was stratified as into manufacturing, retail and other services industries- Manufacturing (ISIC Rev. 3.1 codes 15 - 37), Retail (ISIC code 52) and Other Services (ISIC codes 45, 50, 51, 55, 60-64, and 72).

For the Guatemala ES, size stratification was defined as follows: small (5 to 19 employees), medium (20 to 99 employees), and large (100 or more employees).

Regional stratification for the Guatemala ES was done across two regions: Greater Guatemala City, and Rest of the country.

III. Sampling implementation

Given the stratified design, sample frames containing a complete and updated list of establishments as well as information on all stratification variables (number of employees, industry, and region) are required to draw the sample. Great efforts were made to obtain the best source for these listings.

Asies was the main contractor that implemented the Guatemala 2017 ES.

The sample frame consisted of listings of firms from two sources. For panel firms the list of 565 firms from the Guatemala 2010 ES was used and for fresh firms (i.e., firms not covered in 2010) firm data from Directorio Nacional de Empresas y sus Locales (DINEL), Banco de Guatemala (Banguat) was used.

Table 1: Guatemala ES Sample Frame (Fresh and Panel Combined)

	size	Manufacturing	Retail	Other Services	Grand Total
Greater Guatemala City	Small	716	3,753	6,407	**16,754**
	Medium	301	1,580	2,698	
	Large	86	448	765	
Rest of the country	Small	439	2,295	3,919	**10,249**
	Medium	185	967	1,650	
	Large	52	274	468	
		1,779	**9,317**	**15,907**	**27,003**

Source: World Bank and Directorio Nacional de Empresas y sus Locales (DINEL), Banco de Guatemala (Banguat)

Table 2: Guatemala Sample Frame (Panel)

		Manufacturing	Retail	Other Services	Grand Total
Greater Guatemala City	Small	90	27	16	**444**
	Medium	97	32	26	
	Large	99	21	36	
Rest of the country	Small	38	21	13	**121**
	Medium	13	6	7	
	Large	12	3	8	
		349	**110**	**106**	**565**

Necessary measures were taken to ensure the quality of the frame; however, the sample frame was not immune to the typical problems found in establishment surveys: positive rates of non-eligibility, repetition, non-existent units, etc.

Given the impact that non-eligible units included in the sample universe may have on the results, adjustments may be needed when computing the appropriate weights for individual observations. The percentage of confirmed non-eligible units as a proportion of the total number of sampled establishments contacted for the survey was 13.1% (192 out of 1468 establishments)[4].

Breaking down by industry and size, the following sample targets were achieved (based on the sampling information):

[4] Based on out of target and ineligible contacts

Table 3: Achieved Interviews (Fresh and Panel Combined)

		Manufacturing	Retail	Other Services	Grand Total
Greater Guatemala City	Small	40	29	16	**265**
	Medium	39	29	22	
	Large	52	14	24	
Rest of the country	Small	8	10	11	**80**
	Medium	9	14	21	
	Large	2	2	3	
		150	**98**	**97**	**345**

Table 4: Achieved Interviews (Panel)

		Manufacturing	Retail	Other Services	Grand Total
Greater Guatemala City	Small	26	8	5	**133**
	Medium	27	10	7	
	Large	30	9	11	
Rest of the country	Small	5	7	2	**24**
	Medium	1	1	3	
	Large	2	2	1	
		91	**37**	**29**	**157**

IV. Data Base Structure:

The structure of the data base reflects the fact that 2 different versions of the survey instrument were used for all registered establishments. Questionnaires have common questions (*core* module) and respectfully additional manufacturing- and services-specific questions. The eligible manufacturing industries have been surveyed using the ***Manufacturing*** questionnaire (includes the *core* module, plus manufacturing specific questions). Retail firms have been interviewed using the ***Services*** questionnaire (includes the *core* module plus retail specific questions) and the residual eligible services have been covered using the ***Services*** questionnaire (includes the *core* module). Each variation of the questionnaire is identified by the index variable, *a0*.

All variables are named using, first, the letter of each section and, second, the number of the variable within the section, i.e. *a1* denotes section *A*, question *1* (some exceptions apply due to comparability reasons). Variable names preceded by the prefix "LAC" indicate questions specific to Guatemala and other countries in Latin America 2016, therefore, they may not be found in the implementation of the rollout in other countries. All other suffixed variables are global and are present in all country surveys over the world. All variables are numeric with the exception of those variables with an "x" at the end of their names. The suffix "x" denotes that the variable is alpha-numeric.

There are 2 establishment identifiers, *idstd* and *id*. The first is a global unique identifier. The second is a country unique identifier. The variables *a2* (sampling region), *a6a* (sampling establishment's size), and *a4a* (sampling sector) contain the establishment's classification into the strata chosen for each country using information from the sample frame. The strata were defined according to the guidelines described above.

There are three levels of stratification: industry, size and region. Different combinations of these variables generate the strata cells for each industry/region/size combination. A distinction should be made between the variable a4a and d1a2 (industry expressed as ISIC rev. 3.1 code). The former gives the establishment's classification into one of the chosen industry-strata based on the sample frame, whereas the latter gives the establishment's actual industry classification (four digit code) based on the main activity at the time of the survey.

All of the following variables contain information from the sampling frame. They may not coincide with the reality of individual establishments as sample frames may contain inaccurate or outdated information. The variables containing the sample frame information are included in the data set for researchers who may want to further investigate statistical features of the survey and the effect of the survey design on their results.

-*a2* is the variable describing sampling regions

-*a6a*: coded using the same standard for small, medium, and large establishments as defined above.

-*a4a*: coded following the stratification by sector as defined above.

The surveys were implemented following a 2 stage procedure. Typically first a screener questionnaire is applied over the phone to determine eligibility and to make appointments. Then a face-to-face interview takes place with the Manager/Owner/Director of each establishment. However, sometimes the phone numbers were unavailable in the sample frame, and thus the enumerators applied the screeners in person. The variables *a4b* and *a6b* contain the industry and size of the establishment from the screener questionnaire.

Note that there are variables for size (*l1*, *l6* and *l8*) that reflect more accurately the reality of each establishment. Advanced users are advised to use these variables for analytical purposes. Variables *l1* (number of permanent full-time workers at the end of the last complete fiscal year), *l6* (number of full-time seasonal workers employed during last complete fiscal year) and *l8* (average length of employment of full-time temporary employees during last complete fiscal year) were designed to obtain a more accurate measure of employment accounting for permanent and temporary employment. Special efforts were made to make sure that this information was not missing for most establishments.

The last complete fiscal year is January to December 2016. For questions pertaining to monetary amounts, the unit is the Guatemalan Quetzal.

V. Universe Estimates

Universe estimates for the number of establishments in each cell in Guatemala were produced for the strict, weak and median eligibility definitions described below. The estimates were the multiple of the relative eligible proportions.

For some establishments where contact was not successfully completed during the screening process (because the firm has moved and it is not possible to locate the new location, for example), it is not possible to directly determine eligibility. Thus, different assumptions about the eligibility of establishments result in different adjustments to the universe cells and thus different sampling weights.

In order to account for extreme-value weights of certain panel firms, the same average weight was applied for panel contacts in the following cell: Small retail firms in rest of the country.

Three sets of assumptions on establishment eligibility are used to construct sample adjustments using the status code information.

Strict assumption: eligible establishments are only those for which it was possible to directly determine eligibility. The resulting weights are included in the variable *wstrict*.

Strict eligibility = (Sum of the firms with codes 1,2,3,4,&16) / Total

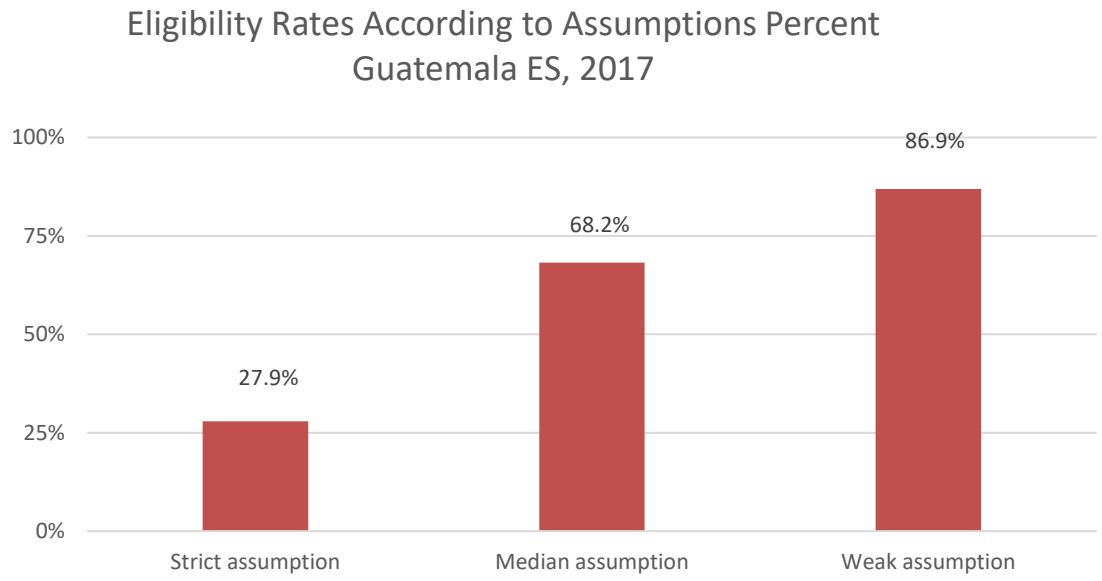
Median assumption: eligible establishments are those for which it was possible to directly determine eligibility and those that rejected the screener questionnaire or an answering machine or fax was the only response. The resulting weights are included in the variable *wmedian*.

Median eligibility = (Sum of the firms with codes 1,2,3,4,16,10,11, & 13) / Total

Weak assumption: in addition to the establishments included in points a and b, all establishments for which it was not possible to contact or that refused the screening questionnaire are assumed eligible. This definition includes as eligible establishments with dead or out of service phone lines, establishments that never answered the phone, and establishments with incorrect addresses for which it was impossible to find a new address. Under the weak assumption only observed non-eligible units are excluded from universe projections. The resulting weights are included in the variable *wweak*.

Weak eligibility= (Sum of the firms with codes, 1,2,3,4,16,10,11,13,91,92,93,94,12) / Total

The indicators computed for the ES website use the median weights. The following graph shows the different eligibility rates calculated for firms in the sample frame under each set of assumptions.



Universe estimates for the number of establishments in each industry-region-size cell in Guatemala were produced for the strict, weak and median eligibility definitions. Appendix B shows the universe estimates of the numbers of registered establishments that fit the criteria of the ES.

Once an accurate estimate of the universe cell projection was made, weights for the probability of selection were computed using the number of completed interviews for each cell.

VI. Weights

Since the sampling design was stratified and employed differential sampling, individual observations should be properly weighted when making inferences about the population. Under stratified random sampling, unweighted estimates are biased unless sample sizes are proportional to the size of each stratum. With stratification the probability of selection of each unit is, in general, not the same. Consequently, individual observations must be weighted by the inverse of their probability of selection (probability weights or *pw* in Stata.)[5]

Special care was given to the correct computation of the weights. It was imperative to accurately adjust the totals within each region/industry/size stratum to account for the presence of ineligible units (the firm discontinued businesses or was unattainable, education or government establishments, no reply after having called in different days of the week and in different business hours, no tone in the phone line, answering machine, fax line[6], wrong address or moved away and could not get the new references). The information required for the adjustment was collected in the first stage of the implementation: the screening process. Using this information, each stratum cell of the

[5] This is equivalent to the weighted average of the estimates for each stratum, with weights equal to the population shares of each stratum.
[6] For the surveys that implemented a screener over the phone.

universe was scaled down by the observed proportion of ineligible units within the cell. Once an accurate estimate of the universe cell (projections) was available, weights were computed using the number of completed interviews.

VII. Appropriate use of the weights

Under stratified random sampling, weights should be used when making inferences about the population. Any estimate or indicator that aims at describing some feature of the population should take into account that individual observations may not represent equal shares of the population.

However, there is some discussion as to the use of weights in regressions (see Deaton, 1997, pp.67; Lohr, 1999, chapter 11, Cochran, 1953, pp.150). There is not strong large-sample econometric argument in favor of using weighted estimation for a common population coefficient if the underlying model varies per stratum (stratum-specific coefficient): both simple OLS and weighted OLS are inconsistent under regular conditions. However, weighted OLS have the advantage of providing an estimate that is independent of the sample design. This latter point may be quite relevant for the ES as in most cases the objective is not only to obtain model-unbiased estimates but also design-unbiased estimates (see also Cochran, 1977, pp 200 who favors the used of weighted OLS for a common population coefficient.)[7]

From a more general approach, if the regressions are descriptive of the population then weights should be used. The estimated model can be thought of as the relationship that would be expected if the whole population were observed.[8] If the models are developed as structural relationships or behavioral models that may vary for different parts of the population, then, there is no reason to use weights.

VIII. Non-response

Survey non-response must be differentiated from item non-response. The former refers to refusals to participate in the survey altogether whereas the latter refers to the refusals to answer some specific questions. Enterprise Surveys suffer from both problems and different strategies were used to address these issues.

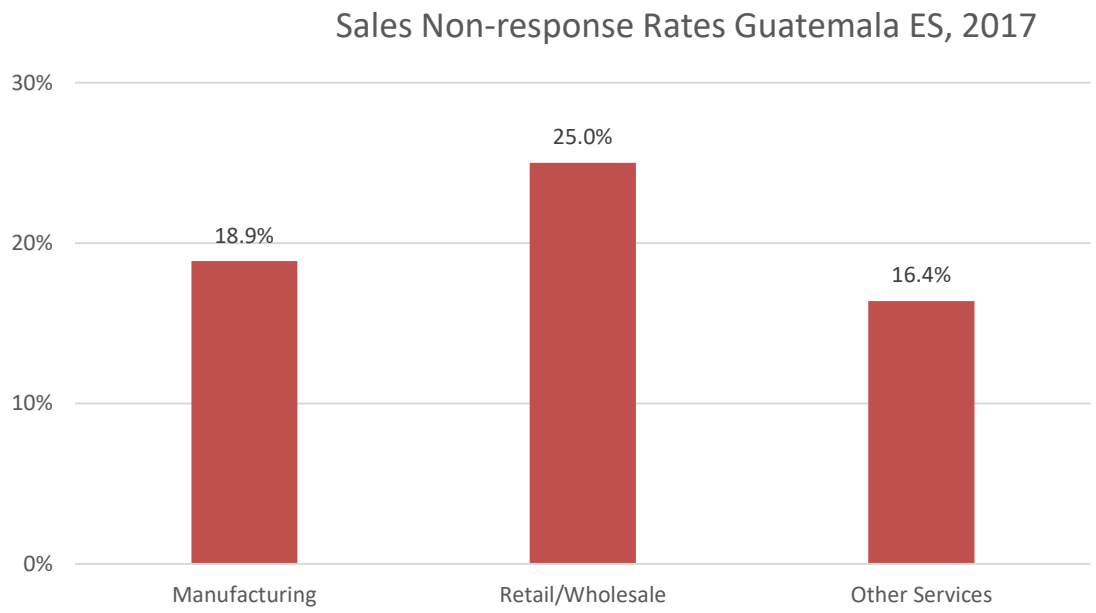
Item non-response was addressed by two strategies:

a- For sensitive questions that may generate negative reactions from the respondent, such as corruption or tax evasion, enumerators were instructed to collect the refusal to respond (-8) as a different option from don't know (-9).

b- Establishments with incomplete information were re-contacted in order to complete this information, whenever necessary. However, there were clear cases of low response. The following graph shows non-response rates for the sales variable, *d2*, by sector. Please, note that for this specific question, refusals were not separately identified from "Don't know" responses.

[7] Note that weighted OLS in Stata using the command regress with the option of weights will estimate wrong standard errors. Using the Stata survey specific commands *svy* will provide appropriate standard errors.

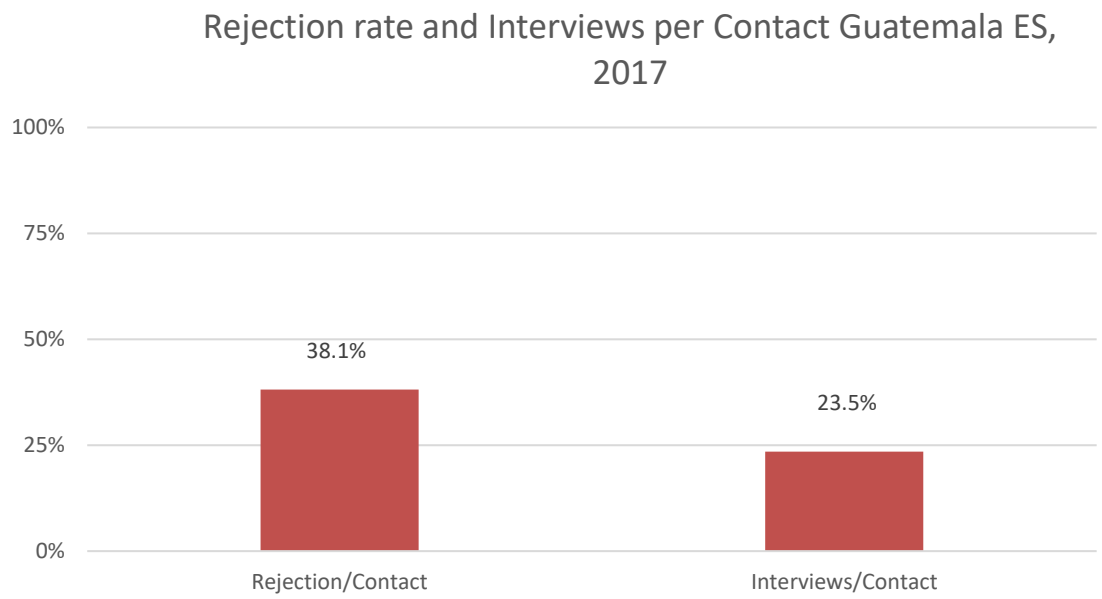
[8] The use weights in most model-assisted estimations using survey data is strongly recommended by the statisticians specialized on survey methodology of the JPSM of the University of Michigan and the University of Maryland.



Survey non-response was addressed by maximizing efforts to contact establishments that were initially selected for interview. Attempts were made to contact the establishment for interview at different times/days of the week before a replacement establishment (with similar strata characteristics) was suggested for interview. Survey non-response did occur but substitutions were made in order to potentially achieve strata-specific goals; whenever this was done, strict rules were followed to ensure replacements were randomly selected within the same stratum. Further research is needed on survey non-response in the Enterprise Surveys regarding potential introduction of bias.

As the following graph shows, the number of interviews per contacted establishments was 0.24.[9] This number is the result of two factors: explicit refusals to participate in the survey, as reflected by the rate of rejection (which includes rejections of the screener and the main survey) and the quality of the sample frame, as represented by the presence of ineligible units. The share of rejections per contact was 0.38.

[9] The estimate is based on the total no. of firms contacted including ineligible establishments.



Details on the rejection rate, eligibility rate, and item non-response are available at the level strata. This report summarizes these numbers to alert researchers of these issues when using the data and when making inferences. Item non-response, selection bias, and faulty sampling frames are not unique to Guatemala. All enterprise surveys suffer from these shortcomings, but in very few cases they have been made explicit.

References:

Cochran, William G., Sampling Techniques, New York, New York: John Wiley & Sons, 1977.

Deaton, Angus, The Analysis of Household Surveys, Baltimore, Maryland: Johns Hopkins University Press, 1998.

Levy, Paul S. and Stanley Lemeshow, Sampling of Populations: Methods and Applications, New York, New York: John Wiley & Sons, 1999.

Lohr, Sharon L. Samping: Design and Techniques, Boston, Massachusetts: Brookes/Cole, 1999.

Scheaffer, Richard L.; Mendenhall, W.; Lyman, R., Elementary Survey Sampling, Fifth Edition, 1996.

Appendix A

Status Codes Enterprise Survey (ES):

0	**Screening in process**	14. In process (the establishment is being called/ is being contacted - previous to ask the screener)	0
409	**Eligible**	1. Eligible establishment (Correct name and address)	373
		2. Eligible establishment (Different name but same address - the new firm/establishment bought the original firm/establishment)	6
		3. Eligible establishment (Different name but same address - the firm/establishment changed its name)	9
		4. Eligible establishment (Moved and traced)	21
		16. Eligible establishment (Panel Firm - now less than five employees; this code applies only to panel firms.)	0
556	**Screener refusal**	13. Refuses to answer the screener	556
183	**Ineligible**	5. The establishment has less than 5 permanent full time employees	8
		616. The firm discontinued businesses - (Establishment went bankrupt)	40
		.	0
		618. The firm discontinued businesses - (Original establishment disappeared and is now a different firm)	6
		619. The firm discontinued businesses - (Establishment was bought out by another firm)	1
		620. The firm discontinued businesses - (It was impossible to determine for what reason)	66
		621. The firm discontinued businesses - (Other)	3
		71. Ineligible legal status: not a business, but private household	33
		72. Ineligible legal status: cooperatives, non-profit organizations, etc.	3
		8. Ineligible activity: Education, Agriculture, Finances, Government, etc.	23
9	**Out of target**	151. Out of target - outside the covered regions	6
		152. Out of target - moved abroad	0
		153. Out of target - Not registered with Statistical Authority	0
		154. Out of target - establishment is HQ without production or sales of goods or services	1

		155. Out of target - establishment was not in operation for the entirety of last fiscal year	0
		156. Duplicated firm within the sample	2
311	**Unobtainable**	91. No reply after having called in different days of the week and in different business hours	167
		92. Line out of order	39
		93. No tone	4
		94. Phone number does not exist	0
		10. Answering machine	32
		11. Fax line- data line	4
		12. Wrong address/ moved away and could not get the new references	65

1468	**Total contacted**

Response Outcomes : Guatemala ES 2017:

Target and totals	Sample target	360
	Sample target completion rate	95.8%
	Total contacts available in frame	1927
	Total contacts issued	1927
	Total contacts contacted	1468

Screening phase	Screening in process	0
	Eligibles	409
	Screener refusal	556
	Ineligible + out of target	192
	Unobtainable	311
Interview phase (only if eligible)	Complete interviews without extra module	345
	Complete interviews with extra module	0
	Eligible in process + incomplete interviews	0
	Interview refusal	4

Percent breakdown (relative to total contacted)	Screening in process rate	0.0%
	Screener refusal rate	37.9%
	Ineligible + out of target rate	13.1%
	Unobtainable rate	21.2%
	Interview conversion rate	23.5%
	Eligible in process + incomplete interviews rate	0.0%
	Interview refusal rate	0.3%

Appendix B: Universe Estimate Based on Sampling Weights

Strict Universe Estimates – Fresh:

		Manufacturing	Retail	Other Services	Grand Total
Greater Guatemala City	Small	853	3,412	1,297	**9,194**
	Medium	412	1,646	626	
	Large	236	516	196	
Rest of the country	Small	312	2,601	917	**5,114**
	Medium	98	814	287	
	Large	20	6	58	
		1,931	**8,995**	**3,381**	**14,308**

Median Universe Estimates – Fresh:

		Manufacturing	Retail	Other Services	Grand Total
Greater Guatemala City	Small	853	3,412	1,297	**9,196**
	Medium	412	1,646	627	
	Large	236	516	197	
Rest of the country	Small	312	2,601	917	**5,129**
	Medium	98	814	287	
	Large	32	10	58	
		1,943	**8,999**	**3,383**	**14,325**

Weak Universe Estimates – Fresh:

		Manufacturing	Retail	Other Services	Grand Total
Greater Guatemala City	Small	853	3,412	1,297	**9,197**
	Medium	412	1,646	628	
	Large	236	516	197	
Rest of the country	Small	312	2,601	917	**5,133**
	Medium	98	814	287	
	Large	35	11	58	
		1,946	**9,000**	**3,384**	**14,330**

Appendix C: Original Sample Design

Original Sample Design (Fresh)

		Manufacturing	Retail	Other Services	Grand Total
Guatemala City	Small	7	25	11	**114**
	Medium	5	14	12	
	Large	17	5	18	
Rest of the Country	Small	5	20	8	**83**
	Medium	13	7	14	
	Large	4	6	6	
		51	**77**	**69**	**197**

Original Sample Design (Panel)

		Manufacturing	Retail	Other Services	Grand Total
Guatemala City	Small	10	14	8	**126**
	Medium	21	8	13	
	Large	23	11	18	
Rest of the Country	Small	2	6	4	**37**
	Medium	7	2	4	
	Large	6	2	4	
		69	**43**	**51**	**163**