

The Peru 2017 Enterprise Surveys Data Set

I. Introduction

This document provides additional information on the data collected in Peru from March 2017 to March 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 Peru ES was selected using stratified random sampling, following the methodology explained in the *Sampling Note*¹. Stratified random sampling² was preferred over simple random sampling for several reasons³:

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 as follows: the universe was stratified into three manufacturing industries and two services industries- Food and Beverages (ISIC Rev. 3.1 code 15), Textiles and Garments (ISIC codes 17,18), Other Manufacturing (ISIC codes 16, 19-37), Retail (ISIC code 52) and Other Services (ISIC codes 45, 50, 51, 55, 60-64, and 72).

For the Peru 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 was done across five regions: Lima, Arequipa, Chiclayo, Trujillo and Piura.

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.

The Peru 2017 ES was implemented by Datum International, S.A.

The sample frame consisted of listings of firms from several sources. For panel firms the list of 1000 firms from the Peru 2010 ES was used, and for fresh firms (i.e., firms not covered in 2010) the lists obtained from Top 10mil 2011, Registro Mype Callao 2010, Registro Mype 2012 and SUNAT (Hacienda) 2011 were used.

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

		Food	Textiles and Garments	Other Manufacturing	Retail	Other Services	Grand Total
Lima	Small	470	926	2809	256	1797	10915
	Medium	274	325	903	193	1220	
	Large	192	236	492	123	699	
Arequipa	Small	62	112	209	312	1406	2517
	Medium	26	10	55	32	202	
	Large	11	12	20	6	42	
Chiclayo	Small	56	16	115	117	675	1136
	Medium	16	4	8	7	94	
	Large	5	0	1	6	16	
Trujillo	Small	57	49	181	222	1220	1988
	Medium	10	4	23	21	151	
	Large	10	1	5	2	32	
Piura	Small	39	3	36	4	34	164
	Medium	6	0	4	2	20	
	Large	3	1	1	1	10	
		1237	1699	4862	1304	7618	16720

Source: World Bank, Top 10mil 2011, Registro Mype Callao 2010, Registro Mype 2012 and SUNAT (Hacienda) 2011

Table 2: Peru Sample Frame (Panel)

		Food	Textiles and Garments	Other Manufacturing	Retail	Other Services	Grand Total
Lima	Small	22	26	92	13	9	707
	Medium	43	49	138	25	20	
	Large	50	61	119	17	23	
Arequipa	Small	7	9	20	7	11	127
	Medium	11	3	15	10	13	
	Large	3	5	6	3	4	
Chiclayo	Small	5	3	16	10	14	78
	Medium	5	2	3	2	13	
	Large	1	0	0	2	2	
Trujillo	Small	11	6	14	11	12	88
	Medium	3	0	9	4	11	
	Large	1	1	1	0	4	
		162	165	433	104	136	1000

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 33.9% (1855 out of 5477 establishments)⁴.

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

Table 3: Achieved Interviews (Fresh and Panel Combined)

		Food	Textiles and Garments	Other Manufacturing	Retail	Other Services	Grand Total
Lima	Small	39	45	55	31	19	535
	Medium	32	40	55	24	22	
	Large	47	38	45	25	18	
Arequipa	Small	14	17	9	23	34	168
	Medium	0	2	15	0	6	
	Large	0	5	6	0	17	
	Medium and Large	10	0	0	10	0	
Piura	All	7	0	8	2	23	40
Chiclayo and Trujillo	Small	15	8	28	54	88	260
	Medium	7	0	0	4	31	
	Large	3	0	0	2	8	
	Medium and Large	0	2	10	0	0	
		174	157	231	175	266	1003

⁴ Based on out of target and ineligible contacts

Table 4: Achieved Interviews (Panel)

		Food	Textiles and Garments	Other Manufacturing	Retail	Other Services	Grand Total
Lima	Small	10	14	34	7	3	251
	Medium	17	20	33	13	9	
	Large	24	23	28	6	10	
Arequipa	Small	3	2	6	2	5	57
	Medium	0	1	7	0	4	
	Large	0	4	4	0	4	
	Medium and Large	7	0	0	8	0	
Chiclayo and Trujillo	Small	5	3	11	11	11	66
	Medium	4	0	0	2	7	
	Large	1	0	0	1	2	
	Medium and Large	0	1	7	0	0	
		71	68	130	50	55	374

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 “ASC” indicate questions specific to Peru and other countries in Latin America 2017, therefore, they may not be found in the implementation of the rollout in other countries. All other variables are global and are present in all country surveys over the world. All variables are numeric except for 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 end date of the last complete fiscal year is identified by variables *a20y*, *a20m*, and *a20d*, collecting information on respectively, year, month, and day. For questions pertaining to monetary amounts, the unit is the Peruvian Sol, PEN.

V. Universe Estimates

Universe estimates for the number of establishments in each cell in Peru were produced without adjusting 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.

Three sets of assumptions on establishment eligibility are usually 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*.

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

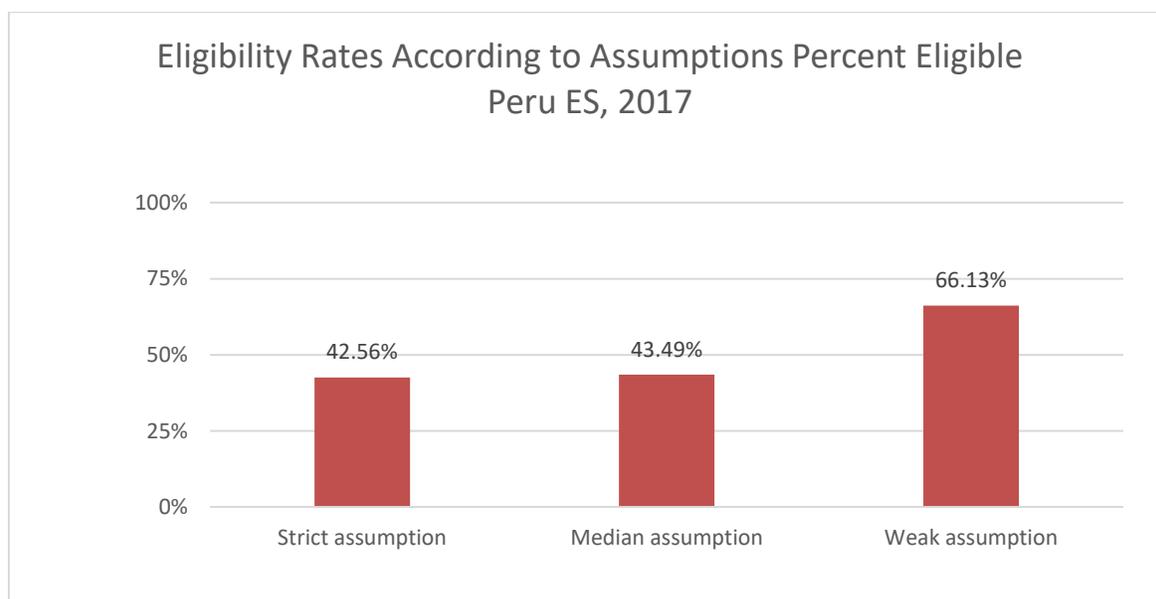
Median assumption: eligible establishments are those for which it was possible to directly determine eligibility and those that rejected the screener questionnaire or an answering machine or fax was the only response. The resulting weights are included in the variable *wmedian*.

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

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*.

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

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 Peru were produced without adjusting 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 line⁶, 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 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 consider 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.

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.

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

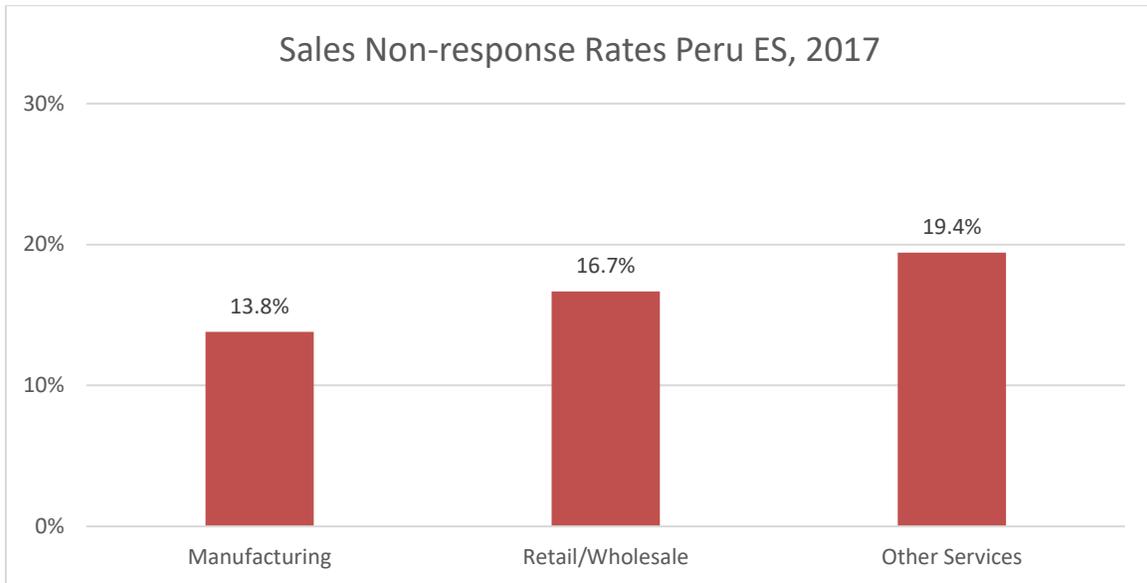
Survey non-response must be differentiated from item non-response. The former refers to refusals to participate in the survey altogether whereas the latter refers to the refusals to answer some specific questions. Enterprise Surveys suffer from both problems and different strategies were used to address 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 share of interviews per contacted establishments was 0.18⁹ 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 0.24.

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 Peru. 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. *Sampling: 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
2331	Eligible	1. Eligible establishment (Correct name and address) 2251 2. Eligible establishment (Different name but same address - the new firm/establishment bought the original firm/establishment) 0 3. Eligible establishment (Different name but same address - the firm/establishment changed its name) 48 4. Eligible establishment (Moved and traced) 32 16. Eligible establishment (Panel Firm - now less than five employees; this code applies only to panel firms.) 0	
0	Screener refusal	13. Refuses to answer the screener	0
1403	Ineligible	5. The establishment has less than 5 permanent full time employees 46 616. The firm discontinued businesses - (Establishment went bankrupt) 2 618. The firm discontinued businesses - (Original establishment disappeared and is now a different firm) 5 619. The firm discontinued businesses - (Establishment was bought out by another firm) 9 620. The firm discontinued businesses - (It was impossible to determine for what reason) 1309 621. The firm discontinued businesses - (Other) 3 71. Ineligible legal status: not a business, but private household 1 72. Ineligible legal status: cooperatives, non-profit organizations, etc. 4 8. Ineligible activity: Education, Agriculture, Finances, Government, etc. 24	
452	Out of Target	151. Out of target - outside the covered regions 14 152. Out of target - moved abroad 1 153. Out of target - Not registered with Statistical Authority 135 154. Out of target - establishment is HQ without production or sales of goods or services 0 155. Out of target - establishment was not in operation for the entirety of last fiscal year 9 156. Duplicated firm within the sample 293	
1291	Unobtainable	91. No reply after having called in different days of the week and in different business hours 778 92. Line out of order 19	

		93. No tone	20
		94. Phone number does not exist	418
		10. Answering machine	50
		11. Fax line- data line	1
		12. Wrong address/ moved away and could not get the new references	5
5477	Total contacted		

Response Outcomes : Peru ES 2017

Target and totals	Sample target	1000
	Sample target completion rate	100.3%
	Total contacts available in frame	16720
	Total contacts issued	5950
	Total contacts contacted	5477

Screening phase	Screening in process	0
	Eligibles	2331
	Screener refusal	0
	Ineligible + out of target	1855
	Unobtainable	1291
Interview phase (only if eligible)	Complete interviews without extra module	1003
	Complete interviews with extra module	0
	Eligible in process + incomplete interviews	0
	Interview refusal	1319

Percent breakdown (relative to total contacted)	Screening in process rate	0.0%
	Screener refusal rate	0.0%
	Ineligible + out of target rate	33.9%
	Unobtainable rate	23.6%
	Interview conversion rate	18.3%
	Eligible in process + incomplete interviews rate	0.0%
	Interview refusal rate	24.1%

Appendix B: Universe Estimate Based on Sampling Weights

Strict Universe Estimates – Fresh:

		Food	Textiles and Garments	Other Manufacturing	Retail	Other Services	Grand Total
Lima	Small	177	298	1050	107	716	4529
	Medium	136	138	447	110	644	
	Large	78	81	195	56	296	
Arequipa	Small	23	36	78	130	559	1035
	Medium	0	4	27	0	106	
	Large	0	5	8	0	19	
	Medium and Large	19	0	0	21	0	
Piura	All	40	0	34	7	57	139
Chiclayo and Trujillo	Small	32	16	85	107	574	976
	Medium	10	0	0	12	97	
	Large	4	0	0	3	15	
	Medium and Large	0	3	17	0	0	
		520	581	1942	552	3085	6679

Strict Universe Estimates – Median:

		Food	Textiles and Garments	Other Manufacturing	Retail	Other Services	Grand Total
Lima	Small	183	303	1111	112	737	4799
	Medium	147	148	498	122	698	
	Large	82	82	208	59	308	
Arequipa	Small	23	35	79	130	553	1036
	Medium	0	4	29	0	111	
	Large	0	5	8	0	19	
	Medium and Large	18	0	0	21	0	
Piura	All	39	0	34	6	56	136
Chiclayo and Trujillo	Small	31	16	87	106	566	972
	Medium	10	0	0	13	101	
	Large	4	0	0	3	15	
	Medium and Large	0	3	17	0	0	
		538	597	2072	571	3164	6942

Strict Universe Estimates – Weak:

		Food	Textiles and Garments	Other Manufacturing	Retail	Other Services	Grand Total
Lima	Small	294	507	1824	170	1244	7257
	Medium	190	197	649	148	934	
	Large	119	125	309	80	468	
Arequipa	Small	36	57	125	191	897	1576
	Medium	0	6	37	0	143	
	Large	0	6	12	0	27	
	Medium and Large	20	0	0	21	0	
Piura	All	42	0	37	6	62	148
Chiclayo and Trujillo	Small	62	32	172	200	1167	1902
	Medium	16	0	0	19	166	
	Large	8	0	0	5	28	
	Medium and Large	0	4	23	0	0	
		786	933	3187	840	5137	10883

Appendix C: Original Sample Design

Original Sample Design (Fresh)

		Food	Textiles and Garments	Other Manufacturing	Retail	Other Services	Grand Total
Lima	Small	2	18	20	2	8	195
	Medium	13	19	15	7	9	
	Large	20	20	16	19	7	
Arequipa	Small	6	17	2	13	20	100
	Medium	5	2	9	6	2	
	Large	2	2	4	1	9	
Chiclayo	Small	18	4	12	30	13	110
	Medium	3	1	2	1	18	
	Large	1	0	1	1	5	
Trujillo	Small	12	14	3	23	26	112
	Medium	2	2	5	5	6	
	Large	3	0	1	1	9	
Piura	Small	16	2	15	2	7	65
	Medium	3	0	2	1	8	
	Large	2	1	1	1	4	
		108	102	108	113	151	582

Original Sample Design (Panel)

		Food	Textiles and Garments	Other Manufacturing	Retail	Other Services	Grand Total
Lima	Small	3	18	20	3	8	200
	Medium	14	20	16	7	10	
	Large	20	20	17	17	7	
Arequipa	Small	7	9	3	7	11	80
	Medium	6	2	9	7	3	
	Large	3	3	4	2	4	
Chiclayo	Small	5	3	12	10	13	70
	Medium	4	1	2	2	13	
	Large	1	0	0	2	2	
Trujillo	Small	11	6	3	11	12	68
	Medium	2	0	5	4	7	
	Large	1	1	1	0	4	
Piura	Small	0	0	0	0	0	0
	Medium	0	0	0	0	0	
	Large	0	0	0	0	0	
		77	83	92	72	94	418