

The Malaysia 2015 Enterprise Surveys Data Set

I. Introduction

This document provides additional information on the data collected in Malaysia between March 2015 and May 2016. 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 130,000 firms in 135 countries, of which 121 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 2015 Malaysia 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 in the way that follows: the universe was stratified into five manufacturing industries and two services industries- Food and Beverages (ISIC Rev. 3.1 code 15), Garments (ISIC code 18), Chemicals (ISIC code 24), Electronic Products (ISIC codes 31 and 32), Other Manufacturing (ISIC codes 16,17,19-23,25-29,30,33-37), Retail (ISIC code 52) and Other Services (ISIC codes 45, 50, 51, 55, 60-64, and 72).

For the Malaysia 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 Malaysia ES was done across five regions: Central, South, North, East Coast and East Malaysia.

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.

Mekong Economics was the main contractor, Kadence International was the subcontractor and Conversation Zone was the sub-subcontractor that implemented the Malaysia 2015 ES.

The sample frame consisted of listings of firms from Malaysia Department of Statistics (DOS); no panel firms were included in the sample frame for the Malaysia 2015 ES.

Table 1: Malaysia ES Sample Frame (Fresh)

		Food Products	Chemical Products	Wearing Apparel	Electronic Products	Other Manufacturing	Retail Trade	Other Services	Grand Total
Central	Small	33	36	42	69	87	69	69	985
	Medium	33	36	36	36	54	36	54	
	Large	33	41	17	36	48	45	75	
South	Small	33	33	51	33	24	27	33	634
	Medium	33	39	44	33	21	15	15	
	Large	42	34	28	39	21	21	15	
North	Small	45	33	54	42	45	24	21	659
	Medium	39	33	36	39	27	18	18	
	Large	30	33	17	45	18	21	21	
East Coast	Small	33	16	36	36	15	15	15	399
	Medium	40	17	13	17	15	15	15	
	Large	30	15	3	5	15	18	15	
East Malaysia	Small	42	42	66	53	25	30	30	645
	Medium	63	24	6	28	25	31	25	
	Large	57	5	1	15	18	31	28	
		586	437	450	526	458	416	449	3,322

Source: Malaysia Department of Statistics.

The quality of the frame was enhanced by the verification process conducted by Mekong Economics. However, the sample frame was not immune from 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 3.1% (82 out of 2672 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 2: Achieved Interviews (Fresh)

		Food Products	Chemical Products	Wearing Apparel	Electronic Products	Other Manufacturing	Retail Trade	Other Services	Grand Total
Central	Small	11	13	13	14	20	20	22	292
	Medium	11	12	12	12	13	12	17	
	Large	11	12	12	12	10	15	18	
South	Small	11	12	12	15	8	6	7	200
	Medium	11	12	13	13	7	5	5	
	Large	8	12	11	14	6	7	5	
North	Small	14	12	13	13	10	7	10	212
	Medium	15	12	12	13	9	5	5	
	Large	13	12	6	12	6	8	5	
East Coast	Small	13	5	12	9	5	6	6	154
	Medium	9	9	10	11	6	6	5	
	Large	9	8	3	3	6	6	7	
East Malaysia	Small	11	7	12	14	5	6	9	142
	Medium	9	6	1	5	3	10	6	
	Large	11	0	1	2	8	10	6	
		167	144	143	162	122	129	133	1,000

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 “EA” or “MYA” indicate questions specific to Malaysia and other countries in East Asia and Pacific 2015 and, 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. Variables *a8* to *a11* contain additional information and were also collected in the screening phase.

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 firms interviewed had several fiscal years. Most firms had January 2014 to December 2014 as their last complete fiscal year. Variables *eea3a3w* (starting month of last complete fiscal year) and *eea3a3y* (last complete fiscal year) can be used to obtain the last complete fiscal year for each firm.

For questions pertaining to monetary amounts, the unit is the Malaysian Ringgit.

V. Universe Estimates

Universe estimates for the number of establishments in each cell in Malaysia 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.

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*.

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

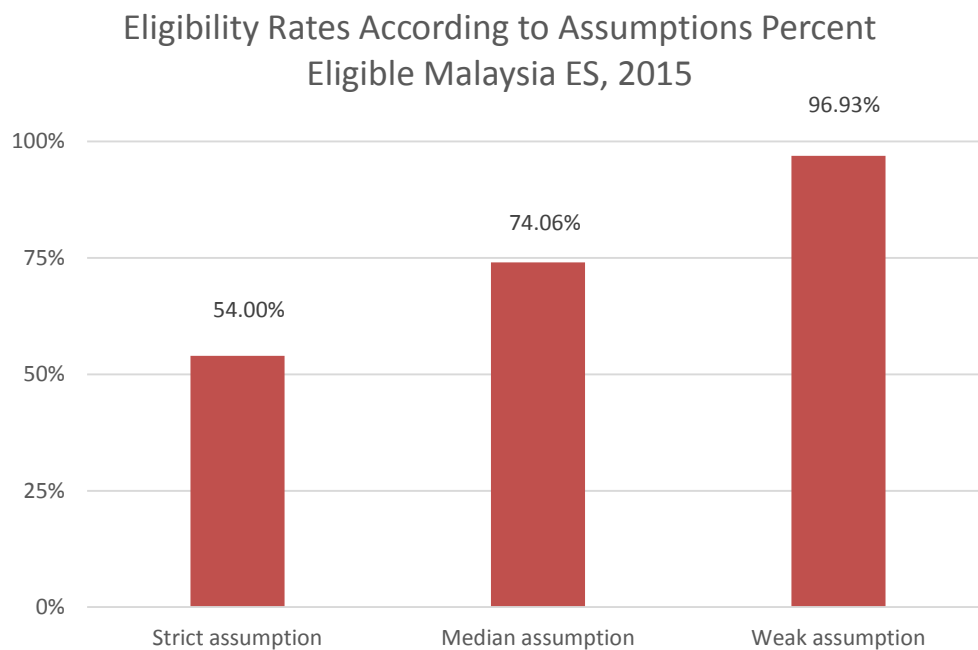
Median assumption: eligible establishments are those for which it was possible to directly determine eligibility and those that rejected the screener questionnaire or an answering machine or fax was the only response. The resulting weights are included in the variable *wmedian*.

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

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 *wwweak*.

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

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 Malaysia 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 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 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 use 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

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.

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.

as structural relationships or behavioral models that may vary for different parts of the population, then, there is no reason to use weights.

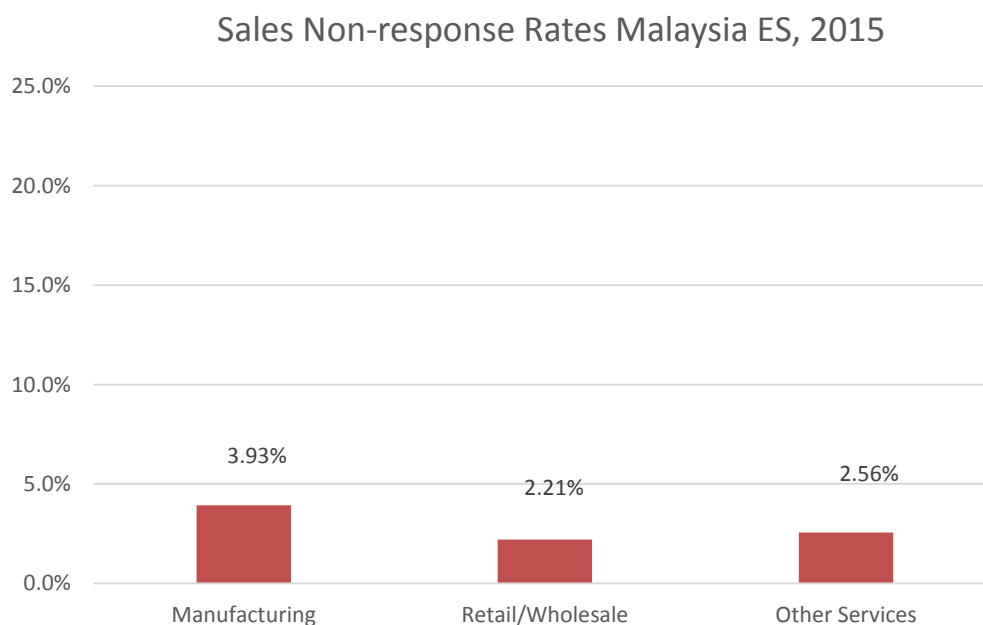
VIII. Non-response

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:

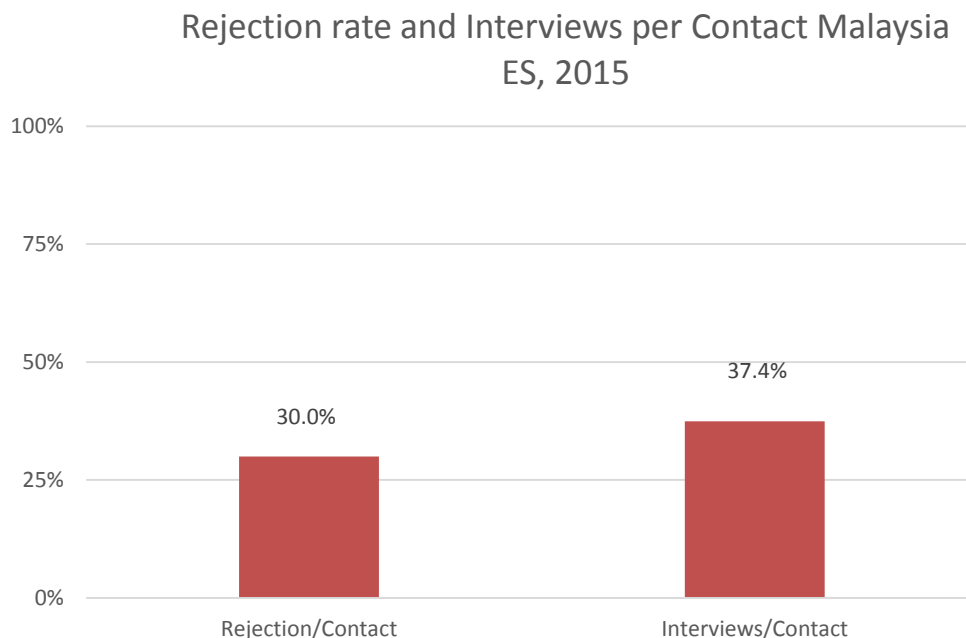
a- For sensitive questions that may generate negative reactions from the respondent, such as corruption or tax evasion, enumerators were instructed to collect the refusal to respond (-8) as a different option from don't know (-9).

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.



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 0.37.⁹ 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 0.3.



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 Malaysia. 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.

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

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
1443	Eligible	1. Eligible establishment (Correct name and address)	1440
		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)	0
		4. Eligible establishment (Moved and traced)	3
		16. Eligible establishment (Panel Firm - now less than five employees; this code applies only to panel firms.)	0
405	Screener refusal	13. Refuses to answer the screener	405
80	Ineligible	5. The establishment has less than 5 permanent full time employees	4
		616. The firm discontinued businesses - (Establishment went bankrupt)	6
		617.	0
		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)	0
		620. The firm discontinued businesses - (It was impossible to determine for what reason)	6
		621. The firm discontinued businesses - (Other)	0
		7. Not a business: Private household	2
		8. Ineligible activity: Education, Agriculture, Finances, Government, etc.	57
2	Out of target	151. Out of target - outside the covered regions	1
		152. Out of target - moved abroad	1

		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	0
		155. Out of target - establishment was not in operation for the entirety of last fiscal year	0
		156. Duplicated firm within the sample	0
742	Unobtainable	91. No reply after having called in different days of the week and in different business hours	493
		92. Line out of order	47
		93. No tone	8
		94. Phone number does not exist	14
		10. Answering machine	119
		11. Fax line- data line	12
		12. Wrong address/ moved away and could not get the new references	49

2672	Total contacted
-------------	------------------------

Response Outcomes : Malaysia ES 2015:

Target and totals	Sample target	1000
	Sample target completion rate	100.0%
	Total contacts available in frame	3322
	Total contacts issued	2752
	Total contacts contacted	2672

Screening phase	Screening in process	0
	Eligibles	1443
	Screeners refusal	405
	Ineligible + out of target	82
	Unobtainable	742
Interview phase (only if eligible)	Complete interviews without extra module	1000
	Complete interviews with extra module	0
	Eligible in process + incomplete interviews	0
	Interview refusal	396

Percent breakdown (relative to total contacted)	Screening in process rate	0.0%
	Screeners refusal rate	15.2%
	Ineligible + out of target rate	3.1%
	Unobtainable rate	27.8%
	Interview conversion rate	37.4%
	Eligible in process + incomplete interviews rate	0.0%
	Interview refusal rate	14.8%

Appendix B: Universe Estimate Based on Sampling Weights

Strict Universe Estimates – Fresh:

		Food Products	Chemical Products	Wearing Apparel	Electronic Products	Other Manufacturing	Retail Trade	Other Services	Grand Total
Central	Small	92	159	221	218	1,756	11,352	18,835	40,553
	Medium	61	131	46	144	999	884	3,686	
	Large	50	61	10	107	514	154	1,073	
South	Small	53	43	74	76	710	3,109	5,930	12,624
	Medium	46	45	34	55	500	186	1,144	
	Large	31	21	17	68	276	29	176	
North	Small	110	55	123	142	1,104	4,860	8,743	18,464
	Medium	95	50	29	65	605	241	1,483	
	Large	35	27	13	96	292	40	257	
East Coast	Small	84	5	55	23	456	2,088	5,218	9,255
	Medium	39	13	9	11	175	147	709	
	Large	19	12	2	4	63	16	108	
East Malaysia	Small	33	14	48	32	296	2,537	5,019	9,811
	Medium	42	9	3	11	170	239	969	
	Large	40	0	0	3	75	35	238	
		829	646	684	1,055	7,991	25,915	53,587	90,706

Median Universe Estimates – Fresh:

		Food Products	Chemical Products	Wearing Apparel	Electronic Products	Other Manufacturing	Retail Trade	Other Services	Grand Total
Central	Small	126	222	292	298	2,418	14,537	26,161	54,716
	Medium	83	183	60	195	1,369	1,127	5,094	
	Large	64	81	13	138	666	185	1,404	
South	Small	71	59	96	101	953	3,879	8,025	16,640
	Medium	62	60	44	72	667	231	1,540	
	Large	39	27	21	85	349	34	224	
North	Small	163	84	175	210	1,647	6,740	13,152	27,039
	Medium	140	76	41	96	898	332	2,219	
	Large	49	39	18	133	410	52	363	
East Coast	Small	109	7	68	30	593	2,522	6,836	11,855
	Medium	50	17	11	14	226	177	924	
	Large	23	14	3	5	77	19	133	
East Malaysia	Small	54	23	75	52	486	3,877	8,318	15,832
	Medium	69	15	4	17	278	363	1,597	
	Large	61	0	1	5	117	51	371	
		1,162	907	922	1,452	11,153	34,125	76,362	126,082

Weak Universe Estimates – Fresh:

		Food Products	Chemical Products	Wearing Apparel	Electronic Products	Other Manufacturing	Retail Trade	Other Services	Grand Total
Central	Small	171	287	392	434	3,415	19,718	34,945	73,689
	Medium	113	235	81	283	1,928	1,524	6,788	
	Large	85	101	17	195	913	244	1,820	
South	Small	99	79	133	153	1,393	5,447	11,098	23,243
	Medium	87	80	61	109	973	324	2,125	
	Large	53	35	28	125	495	46	301	
North	Small	184	90	195	254	1,929	7,581	14,567	30,256
	Medium	157	81	46	115	1,048	373	2,452	
	Large	53	40	19	156	467	57	391	
East Coast	Small	125	8	78	37	708	2,894	7,726	13,499
	Medium	57	18	12	18	269	202	1,041	
	Large	26	15	3	6	89	21	146	
East Malaysia	Small	87	34	119	89	808	6,186	13,073	25,043
	Medium	110	23	7	29	461	577	2,504	
	Large	95	0	1	8	188	78	566	
		1,502	1,127	1,192	2,011	15,083	45,272	99,543	165,730

Appendix C: Original Sample Design

Original Sample Design (Fresh)

		Food Products	Chemical Products	Wearing Apparel	Electronic Products	Other Manufacturing	Retail Trade	Other Services	Grand Total
Central	Small	11	12	14	13	20	23	23	295
	Medium	11	12	12	12	13	12	18	
	Large	11	12	12	12	10	15	17	
South	Small	11	11	13	11	8	5	5	189
	Medium	11	11	12	11	7	5	5	
	Large	10	12	12	11	6	7	5	
North	Small	11	11	13	12	11	6	6	195
	Medium	11	11	12	11	8	5	5	
	Large	10	11	12	11	6	7	5	
East Coast	Small	11	8	12	11	5	5	5	161
	Medium	10	11	12	10	5	5	5	
	Large	10	11	3	6	5	6	5	
East Malaysia	Small	10	11	13	11	5	5	5	160
	Medium	11	11	7	10	5	7	5	
	Large	11	5	1	8	6	7	6	
		160	160	160	160	120	120	120	1,000