The 2016 Zambia Enterprise Skills Survey (ZESS)

I. Introduction

1. This document provides additional information on the data collected for the 2016 Zambia Firm-Level Skills Survey between 9th May 2016 and 19th September 2016, conducted by the Enterprise Analysis (DECEA) and the Education Global Practice (GEDDR) of the World Bank Group.

The objective of the survey is to collect firm-level data for a diagnostics of the composition and demand for skills and the relationship between skills (and/or skills constraints) and firm performance of selected economic sectors in Zambia. A detailed skills module was developed as part of a larger firm-level survey collecting information, among others, on the characteristics of firms and their owners, innovation and export activities, and firm performance.

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

II. Sampling Structure

2. The sample for the survey was selected using stratified random sampling, following a broadly similar methodology used in the World Bank's Enterprise Surveys (ES) – stratified random sampling[1]. However, it is important to note that the universe of inference for the Zambia Firm-Level Skills survey is not strictly comparable to that of the ES. For the ES, the universe of inference is private non-agricultural sectors in the country, excluding the following sectors: financial intermediation (group J[2]), real estate and renting activities (group K, except sub-sector 72, IT) and all public or utilities-sectors. For Zambia Firm-Level Skills Survey, however, the universe is firms in eight selected economic activities, viz., food processing (ISIC15), textile and garments (ISIC 17 & 18), fabricated metal products (ISIC 28), furniture (ISIC 36), construction (ISIC 45), hotel and restaurant (ISIC 55), transport (ISIC 60 - 64) and Information technology (ISIC 72)).

[1] 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). The complete text of the World Bank Enterprise Survey's sampling methodology can be found at http://goo.gl/EgMYXX.

[2] ISIC refers to the ISIC code revision 3.1.

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

4. The universe was stratified into ***eight economic activities*** (as noted above); ***three size*** stratification - small (5 to 19 employees), medium (20 to 99 employees), and large (more than 99 employees); and ***four regions*** (city and the surrounding business area): Kitwe, Livingstone, Lusaka, and Ndola.

III. Sampling Implementation

5. 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. However, the quality of the sample frames was not optimal and, therefore, some adjustments were needed to correct for the presence of ineligible units. These adjustments are reflected in the weights computation (*see below*).

6. Lusaka Probe Market Research was hired to implement the fieldwork.

7. Sample frame used for the survey is based on the 2010 Zambia Establishment census, collected and maintained by Zambia Statistical Office. This is the same sampling frame used for the 2013 Zambia Enterprise Survey conducted by the World Bank.

The sampling frame database contained the following information:

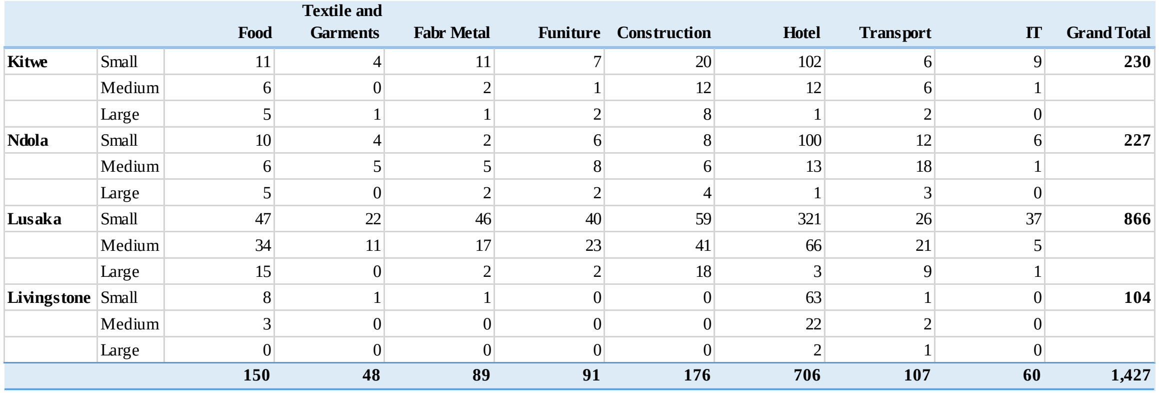
a) Detailed stratification variables;
b) Location identifiers- address, phone number, email; and
c) Contact name(s).

8. The enumerated establishments with 5 employees or more were then used as the sample frame for the 2016 Zambia Firm-Level Skills Survey with the aim of obtaining interviews of 390 establishments.

9. The quality of the frame was assessed at the onset of the project through visits to a random subset of firms and local contractor knowledge. The sample frame was not immune from the typical problems found in establishment surveys: positive rates of non-eligibility, repetition, non-existent units, etc.

10. 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.

Counts from sample frames are shown below:



		Food	Textile and Garments	Fabr Metal	Funiture	Construction	Hotel	Transport	IT	Grand Total
Kitwe	Small	11	4	11	7	20	102	6	9	**230**
	Medium	6	0	2	1	12	12	6	1	
	Large	5	1	1	2	8	1	2	0	
Ndola	Small	10	4	2	6	8	100	12	6	**227**
	Medium	6	5	5	8	6	13	18	1	
	Large	5	0	2	2	4	1	3	0	
Lusaka	Small	47	22	46	40	59	321	26	37	**866**
	Medium	34	11	17	23	41	66	21	5	
	Large	15	0	2	2	18	3	9	1	
Livingstone	Small	8	1	1	0	0	63	1	0	**104**
	Medium	3	0	0	0	0	22	2	0	
	Large	0	0	0	0	0	2	1	0	
		150	**48**	**89**	**91**	**176**	**706**	**107**	**60**	**1,427**

Source: 2010 Zambia Census of Business Establishment, Zambia Statistical Office

IV. Data Base Structure:

11. Data is collected using single and standardized questionnaire administered to all firms. The questionnaire has eight sections; six main sections and two sections on control information.

12. 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). 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.

13. There is a unique establishment identifiers, variable name *id*. 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.

14. 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 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 using ISIC codes for the chosen industries for stratification. These codes include food processing (ISIC[3] 15), textile and garments (ISIC 17 & 18), fabricated metal products (ISIC 28), furniture (ISIC 36), construction (ISIC 45),

[3] ISIC refers to the ISIC code revision 3.1.

hotel and restaurant (ISIC 55), transport (ISIC 60 - 64) and Information technology (ISIC 72)).

15. The surveys were implemented following a two-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. In some cases, when the phone numbers were unavailable in the sample frame, the enumerators applied the screeners in person. The variables *a4b* and *a6b* contain the industry and size of the establishment from the screener questionnaire.

16. Note that the fiscal years vary by firm as there is no standard for all firms in Zambia. The start and end dates for the fiscal year for each firm can be found in the *fymonb, fyyearb, fymone* and *fyyeare* variables in the dataset. This information is of particular help in determining the appropriate exchange rate period to use to convert all monetary values from the local currency to, for instance, the US$.

V. Universe Estimates

17. Universe estimates for the number of establishments in each cell (i.e., region-industry-size) were produced for the strict, weak and median eligibility definitions. The estimates were the multiple of the relative eligible proportions. Appendix A, B and B1 provides the estimates the universe based on the strict, median and weak eligibility assumptions respectively (*see below for definition of the three eligibility assumptions*). Appendix E provides definition of eligibility codes.

18. 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.

19. Three sets of assumptions on establishment eligibility are used to construct sample adjustments using the status code information. Appendix E provides the definition of eligibility and firm counts.

20. 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) / Total

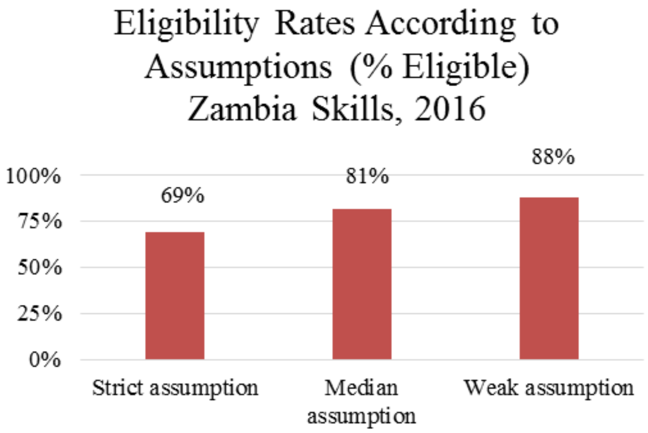
21. Median assumption: eligible establishments are those for which it was possible to directly determine eligibility and those that rejected the screener questionnaire or an answering machine or fax was the only response. The resulting weights are included in the variable *wmedian*.

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

22. 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 1,2,3,4,91,92,93,10,11,12,&13) / Total

23. The following graph shows the different eligibility rates calculated for firms in the sample frame under each set of eligibility assumptions:



24. 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

25. 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.)[4]

[4] This is equivalent to the weighted average of the estimates for each stratum, with weights equal to the population shares of each stratum.

26. Three versions of sampling weights are provided based on the three eligibility assumptions noted above, i.e., strict, median and weak weights. 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, establishments with less than 5 employees, 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[5], 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

28. 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. For estimations with weighting, we recommend the use of the median weights.

29. 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 a 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 has the advantage of providing an estimate that is independent of the sample design.[6]

VIII. Non-response

30. 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. The 2016 Zambia Firm-Level Skills Survey suffers from both problems and different strategies were used to address these issues.

31. Item non-response was addressed by re-contacting firms. That is, establishments with incomplete information were re-contacted in order to complete this information, whenever necessary. However, there were clear cases of low response. The response rates

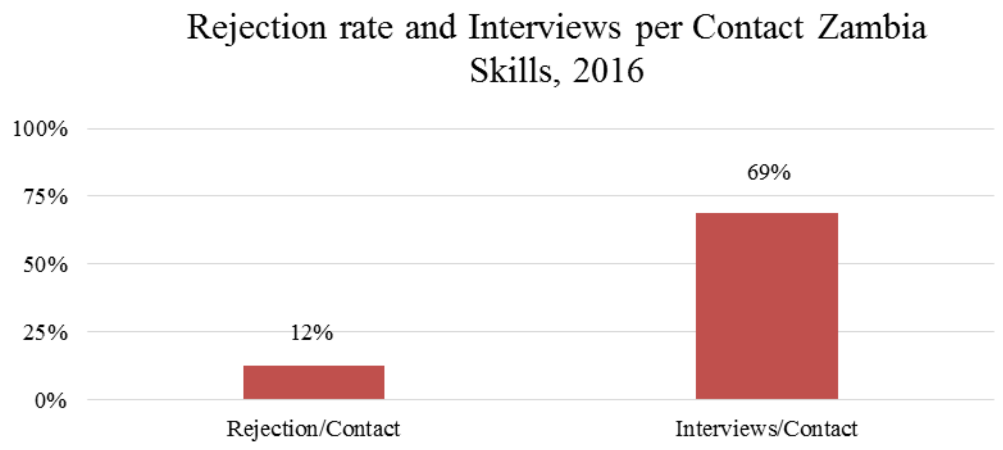
[5] For the surveys that implemented a screener over the phone.

[6] 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.

are particularly low for questions about the names and locations of the main universities and schools attended by the establishment's recent hires. Nevertheless, utmost effort was made to recover responses through repeated callbacks.

32. 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.

33. As the following graphs show, the number of realized interviews per contact contacted establishments was 0.69.[7] 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 number of rejections per contact was 0.12.



34. Details on the rejection rate, eligibility rate, and item non-response are available at the level of 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 the 2016 Zambia Firm-Level Skills Survey. All firm-level surveys suffer from these shortcomings, but in very few cases they have been made explicit. Appendix F provides further detail.

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

References:

Cochran, William G., Sampling Techniques, 1977.

Deaton, Angus, The Analysis of Household Surveys, 1998.

Levy, Paul S. and Stanley Lemeshow, Sampling of Populations: Methods and Applications, 1999.

Lohr, Sharon L. Samping: Design and Techniques, 1999.

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

Appendix A

Universe Estimates based on the Strict Eligibility Assumption

		Food	Textile and Garments	Fabr Metal	Funiture	Construction	Hotel	Transport	IT	Grand Total
Kitwe	Small	8	3	8	6	15	85	4	6	**177**
	Medium	4	0	1	1	8	9	4	0	
	Large	4	1	0	2	6	1	2	0	
Ndola	Small	9	4	2	6	7	101	11	5	**213**
	Medium	5	4	4	8	5	12	15	1	
	Large	4	0	2	2	4	1	3	0	
Lusaka	Small	28	13	27	28	35	219	16	19	**543**
	Medium	18	6	9	15	23	41	12	2	
	Large	9	0	1	1	11	2	6	0	
Livingstone	Small	6	1	1	0	0	51	1	0	**80**
	Medium	2	0	0	0	0	16	1	0	
	Large	0	0	0	0	0	2	0	0	
		96	**31**	**55**	**70**	**114**	**541**	**74**	**33**	**1,013**

Appendix B

Universe Estimates based on the median Eligibility Assumption

		Food	Textile and Garments	Fabr Metal	Funiture	Construction	Hotel	Transport	IT	Grand Total
Kitwe	Small	9	3	10	7	17	98	5	8	**214**
	Medium	5	0	2	1	11	12	6	0	
	Large	5	1	0	2	8	1	2	0	
Ndola	Small	9	3	2	6	7	98	11	5	**220**
	Medium	5	4	5	9	6	13	17	1	
	Large	5	0	2	2	4	1	3	0	
Lusaka	Small	31	13	33	32	39	237	18	24	**624**
	Medium	23	7	13	19	28	51	15	3	
	Large	11	0	2	2	14	3	7	0	
Livingstone	Small	7	1	1	0	0	59	1	0	**97**
	Medium	3	0	0	0	0	22	2	0	
	Large	0	0	0	0	0	2	0	0	
		113	**32**	**69**	**80**	**133**	**598**	**88**	**42**	**1,155**

Appendix B1

Universe Estimates based on the median Eligibility Assumption

		Food	Textile and Garments	Fabr Metal	Funiture	Construction	Hotel	Transport	IT	Grand Total
Kitwe	Small	10	3	11	7	18	98	6	9	**219**
	Medium	6	0	2	1	11	12	6	0	
	Large	5	1	0	2	8	1	2	0	
Ndola	Small	9	3	2	6	7	95	12	6	**217**
	Medium	6	4	5	8	5	12	18	1	
	Large	5	0	2	2	4	1	3	0	
Lusaka	Small	38	15	39	36	46	266	22	31	**714**
	Medium	27	8	14	21	32	54	18	4	
	Large	13	0	2	2	15	3	8	0	
Livingstone	Small	7	1	1	0	0	58	1	0	**94**
	Medium	3	0	0	0	0	20	2	0	
	Large	0	0	0	0	0	2	0	0	
		127	**34**	**78**	**87**	**147**	**622**	**98**	**50**	**1,244**

Appendix C

Original Sample Design, Zambia Firm-Level Skills Survey

		Food	Textile and Garments	Fabr Metal	Funiture	Construction	Hotel	Transport	IT	Grand Total
Kitwe	Small	4	2	4	3	6	18	3	5	**71**
	Medium	3	0	1	1	4	2	3	1	
	3:Large	3	1	1	1	3	1	1	0	
Ndola	Small	4	2	1	3	3	17	5	3	**70**
	Medium	3	3	3	3	2	2	6	1	
	3:Large	2	0	1	1	2	1	2	0	
Lusaka	Small	12	7	12	11	15	20	8	16	**169**
	Medium	9	4	5	6	10	8	6	3	
	3:Large	4	0	1	1	5	2	3	1	
Livingstone	Small	4	1	1	0	0	19	1	0	**40**
	Medium	2	0	0	0	0	9	1	0	
	3:Large	0	0	0	0	0	1	1	0	
		50	**20**	**30**	**30**	**50**	**100**	**40**	**30**	**350**

Appendix D

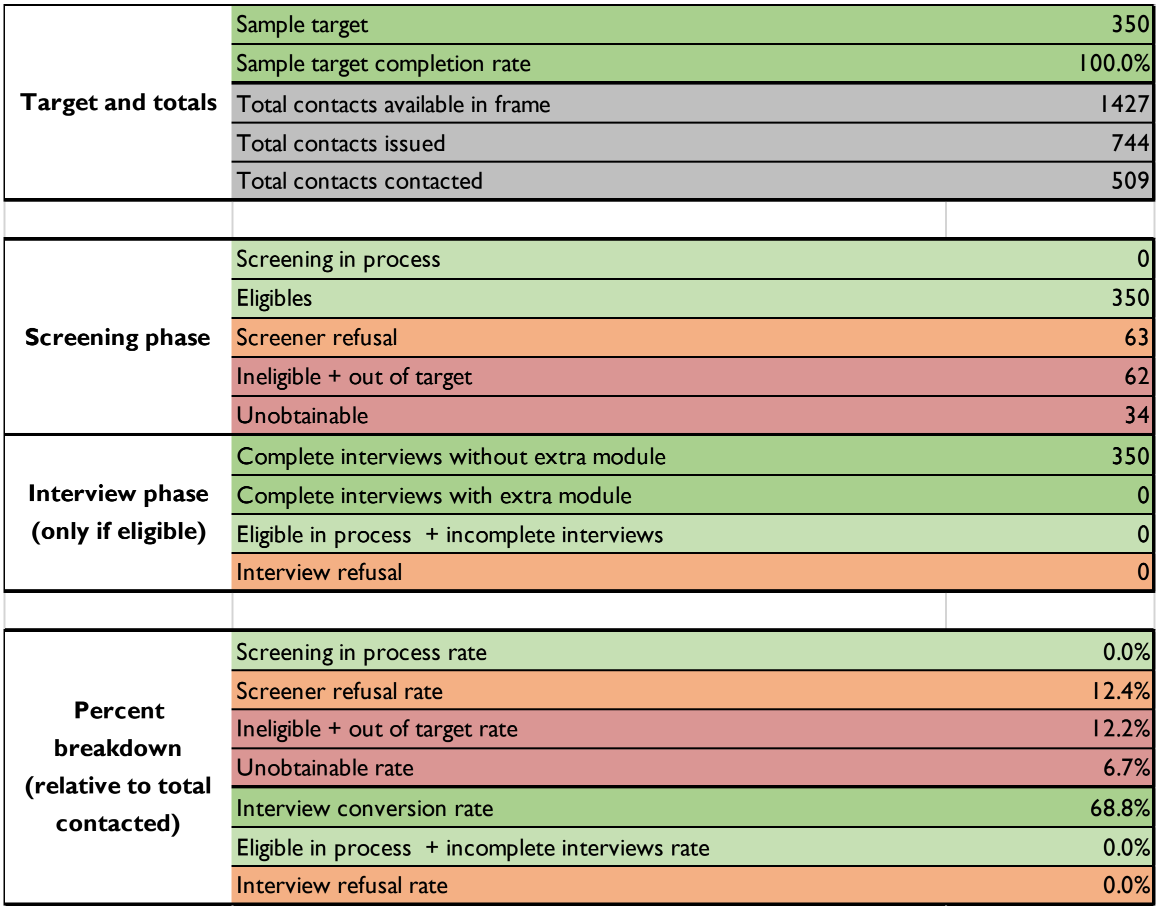
Completed Interviews, Zambia Firm-Level Skills Survey

		Food	Textile and Garments	Fabr Metal	Funiture	Construction	Hotel	Transport	IT	Grand Total
Kitwe	Small	4	2	5	3	6	18	3	6	**71**
	Medium	3	0	1	1	4	2	3	0	
	Large	3	1	0	1	3	1	1	0	
Ndola	Small	4	2	1	3	3	17	5	3	**70**
	Medium	3	3	3	3	2	2	6	1	
	Large	2	0	1	1	2	1	2	0	
Lusaka	Small	12	7	12	11	15	20	8	18	**169**
	Medium	9	4	5	6	10	8	6	2	
	Large	4	0	1	1	5	2	3	0	
Livingstone	Small	4	1	1	0	0	19	1	0	**40**
	Medium	2	0	0	0	0	9	2	0	
	Large	0	0	0	0	0	1	0	0	
		50	**20**	**30**	**30**	**50**	**100**	**40**	**30**	**350**

Appendix E: Eligibility Code, Zambia Enterprise Skills Survey

0	Screening in process	14. In process (the establishment is being called/ is being contacted - previous to ask the screener)	0
350	Eligible	1. Eligible establishment (Correct name and address)	348
		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)	2
		16. Eligible establishment (Panel Firm - now less than five employees; this code applies only to panel firms.)	0
63	Screener refusal	13. Refuses to answer the screener	63
57	Ineligible	5. The establishment has less than 5 permanent full time employees	6
		616. The firm discontinued businesses - (Establishment went bankrupt)	5
		.	0
		618. The firm discontinued businesses - (Original establishment disappeared and is now a different firm)	2
		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)	38
		621. The firm discontinued businesses - (Other)	0
		71. Ineligible legal status: not a business, but private household	0
		72. Ineligible legal status: cooperatives, non-profit organizations, etc.	0
		8. Ineligible activity: Education, Agriculture, Finances, Government, etc.	6
5	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	1
		156. Duplicated firm within the sample	2
34	Unobtainable	91. No reply after having called in different days of the week and in different business hours	2
		92. Line out of order	1
		93. No tone	0
		94. Phone number does not exist	0
		10. Answering machine	0
		11. Fax line- data line	0
		12. Wrong address/ moved away and could not get the new references	31
509	Total contacted		

Appendix F: Interview Conversion rates, Further Detail



Target and totals	Sample target	350
	Sample target completion rate	100.0%
	Total contacts available in frame	1427
	Total contacts issued	744
	Total contacts contacted	509
Screening phase	Screening in process	0
	Eligibles	350
	Screener refusal	63
	Ineligible + out of target	62
	Unobtainable	34
Interview phase (only if eligible)	Complete interviews without extra module	350
	Complete interviews with extra module	0
	Eligible in process + incomplete interviews	0
	Interview refusal	0
Percent breakdown (relative to total contacted)	Screening in process rate	0.0%
	Screener refusal rate	12.4%
	Ineligible + out of target rate	12.2%
	Unobtainable rate	6.7%
	Interview conversion rate	68.8%
	Eligible in process + incomplete interviews rate	0.0%
	Interview refusal rate	0.0%

Appendix G

Local Agency team involved in the study:

Local Agencies	**Name:** Lusaka Probe Market Research. **Country:** Zambia **Activities since:** 2003
Enumerators involved:	**Enumerators:** 25 **Supervisors**: 5
Other staff involved:	**Fieldwork Coordinators:** 2 **Data Processing:**2

Sample Frame:

Characteristic of sample frame used:	Zambia Census of Business Establishments
Source:	Zambia Central Statistics Office
Year:	2010

Sectors included in the Sample:

Original Sectors	Eight selected economic activities, viz., food processing (ISIC[8] 15), textile and garments (ISIC 17 & 18), fabricated metal products (ISIC 28), furniture (ISIC 36), construction (ISIC 45), hotel and restaurant (ISIC 55), transport (ISIC 60-64) and Information technology (ISIC 72)).
Added (top up) Sectors	None

[8] ISIC refers to the ISIC code revision 3.1.

Fieldwork and Country Situation:

Date of Fieldwork	9 May 2016 to 19th September 2016.
Country	Zambia
Use of CAPI	• Computer-assisted personal interviewing (CAPI) was used to collected data. The CSPro software was used for the survey.
Problems found during fieldwork:	• Some of the contact numbers provided on the list were no longer in use or could not go through. To overcome this challenge, both supervisors and enumerators were encourage to establish the location of such firms by asking people who are familiar with the location. • Some respondents refused to participate in the survey, mostly noting that they have been over researched. Some in fact, requested whether there is any monetary benefits for participating in the survey. In those cases, supervisors and interviewers were encouraged to explain the survey produces aggregated national benefits and not individual monetary benefits as was stressed during field team training. To aid the supervisors and enumerators to explain the benefits of the survey, printed documents including the introduction letter and other materials were provided to potential participants to read. • The questionnaire was a bit long for managers to complete. Consequently, in many cases managers either asked enumerators to return on another date, or referred them to another person who did not have enough information about the company as the managers do. However, making follow ups to partially completed interviews proved to be expensive in some cases as enumerators would have to visit the firm more than once or twice. • Some respondents were less transparent, particularly on financial information. This necessitated repeated callbacks to convince respondents to answer these questions. Enumerators were encouraged to guarantee the respondent of confidentiality of the information provided by the respondent. • Obtaining information on some of the questions in the labor section was also a challenge for some respondents, as they thought the information may be used by the government for inspection and compliance on labor standards. Enumerators did all their best to guarantee the respondent of confidentiality of the information provided by the respondent. • There was a national election on 11th of August, and this slowed down the final stages of the fieldwork, particularly callbacks to finalize partially complete interviews. Fieldwork was also suspended for few days before and immediately after the election date. However, this was of limited consequence since the fieldwork was almost in the final stages at that point.