Latin America Enterprise Surveys Data Set

1. Introduction

1. The following document provides additional information on the data collected in Latin America during the calendar years 2006 and 2007. It describes the sampling design of the data, the data set structure and it provides additional information that may be useful when using the data such as information on non-response and the appropriate use of the weights.

2. Sampling Structure

2. The sample for each individual country was selected using stratified random sampling, following the methodology explained in the Sampling Manual. Stratified random sampling was preferred over simple random sampling for several reasons[1]:

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 (group D), construction (group F), services (groups G and H), and transport, storage, and communications (group I). Groups are defined following ISIC revision 3.1. Note that this definition excludes the following sectors: financial intermediation (group J), real estate and renting activities (group K, excluding sub-sector 72, IT, which was added to the population under study), and all public or utilities-sectors.

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.)

3. Three levels of stratification were used in every country: industry, establishment size, and region. The original sample designs with specific information of the industries and regions chosen for each country are included in the attached Excel file (Sampling Report.xls.)

4. Countries included in the project were classified according to the size of their economies into:

a- Small size: Guatemala, El Salvador, Honduras, Nicaragua, Panama, Peru, Ecuador, Bolivia, Paraguay, and Uruguay.

b- Middle size: Colombia, Venezuela, Argentina, and Chile.

c- Large size economy: Mexico.

5. Industry stratification was designed in the following way: In small economies the population was stratified into 3 manufacturing industries, one services industry – retail-, and one residual sector as defined in the sampling manual. Each industry had a target of 120 interviews. In middle size economies the population was stratified into 4 manufacturing industries, 2 services industries -retail and IT-, and one residual sector.

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

For the manufacturing industries sample sizes were inflated by 25% to account for potential non-response in the financing data. Mexico, due to its size, was stratified into 7 manufacturing industries, 2 services industries, retail and IT, and one residual stratum. The target number of interviews for manufacturing strata was also inflated by 25% to minimize the effect of item non-response.

6. Size stratification was defined following the standardized definition for the rollout: small (5 to 19 employees), medium (20 to 99 employees), and large (more than 99 employees). For stratification purposed, the number of employees was defined on the basis of reported permanent full-time workers. This resulted in some difficulties in certain countries where seasonal/casual/part-time labor is common.

7. Regional stratification was defined within country. In general, small economies included 2 to 3 regions, medium size economies included 4 regions, and in Mexico 8 regions were included. The actual selected regions for each country can be found in the attached Excel file (Sampling Report.xls.)

4. Sampling implementation

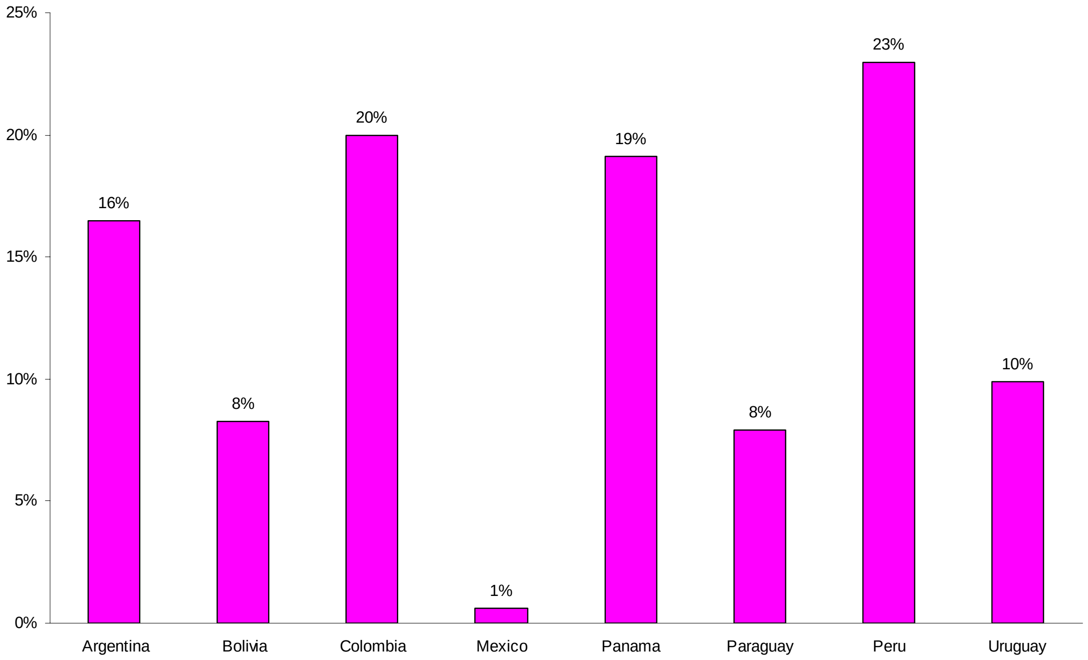
8. Given the stratified design sample frames containing a complete and updated list of establishments for the regions selected were required. For each country, great efforts were made to obtain the best source for these listings. However, the quality of the sample frames in most countries was not optimal and, therefore, adjustments were needed to correct for the presence of ineligible units. These adjustments are reflected in the weights computation (see below.)

9. The sources of the sample frame for each country were:

Country	Sample Frame Source	Date
Argentina	National Census (for totals per cell) and National Industry Register + Industry Guide + Telephone Directory of the Argentine Republic + National, Provincial and Local organisms + Private Commerce Chambers + TNS Gallup information (several steps in generation),	2004-2005-2006
Bolivia	Economic Establishments Census	2004 updated to 2006
Colombia	Comfecamaras	2004
Mexico	INEGI	2006
Panama	Industry and Commerce Census of Panama	1999
Peru	Base of Top 10000 Companies Peru, updated 2006 through studies conducted by DATUM International S.A.	2006
Paraguay	Paraguay's Census of Industry and List of contributor firms from 2000 to 2004	2004
Uruguay	Permanent Register of Economic Activities (companies, not establishments)	2004
Venezuela	NA (given the low quality of the original frame a new methodology was put in place)	
Ecuador		
Chile	Instituto Nacional de Estadísticas (INE)	2005

10. The quality of the frame was assessed at the onset of the project. The frames proved to be useful though they showed different rates of non-eligible, repetitions, non-existent units, etc. These problems are typical of establishment surveys but given the impact these inaccuracies may have on the results, adjustments were needed when computing the appropriate weights for individual observations. The following graph exhibits the percentage of confirmed non-eligible units found in each country as a proportion of the total number of contacts to complete the survey.

Observerved Rate of Non-Eligibility per Country



11. In Venezuela, due to the inaccuracy of the best sample frame available was found to be very inaccurate during the early stages of fieldwork. The decision was therefore taken to abort its use and employ more traditional area enumeration methods. The approach employed was as follows. Aerial maps of Caracas, Maracay and Valencia were obtained, divided into approximately equal blocks by size and classified using local knowledge into types of area- residential, retail and service, office, industrial, primary. The accuracy of this classification was checked in a small scale pilot of 31 blocks. A sample of 431 blocks was then fully enumerated and used as a second-stage sampling frame and also as the basis of projection to the eligible business establishment universe establishment. From within the enumerated eligible establishments a sample of establishments has been selected systematically within strata to provide 500 effective interviews using a shorter version of the primary questionnaire

3. Data Base Structure:

11. The structure of the data base reflects the fact that 3 different versions of the questionnaire were used. One basic variation, the Core Questionnaire, includes all common questions asked to all establishments. One expanded variation, the Manufacturing Questionnaire, adds some specific questions relevant for the sector. Another expanded variation, the Services Questionnaire, adds to the core specific questions relevant to either retail or IT. Each variation of the questionnaire is identified by the index variable, *a0*.

12. Since all countries used the same questionnaires, all data sets have been appended into a unique data set in which the country is identified by the index variable *a1*. There is only one country-specific question, the educational level of the labor force -*l9*-. Results

from the pilot of the questionnaire showed that it was very confusing to use the global scale for all countries as respondents are accustomed to the scales regularly used in household surveys in each country. The individual scales for each country were included at the end of the published manufacturing questionnaire. The data set contains each country-specific variable under the name *l9_country* as well as the equivalent match to the global education question *l9*. The criteria used for the matching was to take the largest number of years possible for any given category. For example, in most countries "primary incomplete" could be matched to either "0-3 years" or "3 to 6 years". The decision was to match it to the largest number "3 to 6 years". The original variable was included for users to make their own match according to their interests.

13. 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*. Variables preceded by a capital *L* are variables specific to Latin America and, therefore, they may not be found in the implementation of the rollout in other regions. All other variables are global. All variables are numeric with the exception of those variables with an "x" at the end of its name, which denotes that the variable is alpha-numeric.

14. There are 3 establishment identifiers, *idstd*, *idu*, and *id*. The first is a global unique identifier. The second is a regional unique identifier, and *the* third one is a country unique identifier. The variables *region_sample*, *size_sample,* and *ind_sample* 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 and adjusted according to the available information (for ex. a stratum for "unknown size" had to be created because some establishments lack this information in the sample frame.)[2]

15. As noted above, these are 3 levels of stratification within each country: industry, size and region. Different combinations of these variables generate the strata cells for each industry/region/size combination. The variable *strata* identifies each cell. A distinction should be made between the variable *ind_sample* and *isic*. The former gives the establishment's classification into one of the chosen industry-strata in a given country whereas the latter gives the actual establishment's industry classification in the sample frame.

16. All the following variables contain information from the sampling frame and were defined with the sampling design. They may not coincide with the reality of individual establishments as sample frames are inaccurate. These variables that contain sample frame information is 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. Note that no previous data set generated at The World Bank includes comparable information and users are advised not to use these variables for analytical purposes.

[2] In the previous version of the data, 3 variables that were supposed to capture the information from the sample frame (a2, a4a, and a6a) were incorrectly used by the implementing firms. Consequently, they were dropped form the last version of the survey in exchange for the exact variables taken from the sample-control lists generated when selecting of the sample.

-region_sample: coded following the codes in the attached spreadsheet “Sampling Report.xls”, worksheet “Region”

-size_sample: coded using the same standard for small, medium, and large establishments as defined above. The code *-9* was used to indicate units for which size was undetermined in the sample frame.

-ind_sample: coded using ISIC codes for the chosen industries for stratification in each country. These codes include most manufacturing industries (15 to 36), and wholesale, retail, and IT for services (51, 52, and 72 respectively). All establishments within the residual stratum were coded with ind_sample=2.

-strata: unique stratum identifier. This variable is important in Stata when setting the data set as a survey data set.

-isic: original ISIC classification from the sample frame. Note that a few cases lack this classification and were assigned to the residual.

17. The surveys were implemented following a 2 stage procedure. In the first stage a screener questionnaire was applied over the phone to determine eligibility and to make appointments; in the second stage, a face-to-face interview took place with the Manager/Owner/Director of each establishment. The variables *a4b* and *a6b* contain the industry and size of the establishment from the screener questionnaire. Variables *a8* to *a11* were also collected in the screening phase.

18. Note that there are additional variables for region, industry, and size that reflect more accurately the reality of each establishment: *a3x, d1a2*, and *l1*, *l6* and *l8*. Users are advised to use these variables for analytical purposes.

19 Variable *a3x* indicates the actual location of the establishment. There may be divergences between the location in the sampling frame and the actual location, as establishments may be listed in one place but the actual physical location is in another place.

20. Variable *d1a2* indicates the actual ISIC code of the main output of the establishment as answered by the interviewee. This is probably the most accurate variable to classify establishments by activity. However, question *d1a2* was only asked from manufacturing establishments and, therefore, establishments in the services and residual strata must be classified using sampling information (*isic.*)

21. Variables *l1*, *l6* and *l8* 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.

3. Weights

22. 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 *pa* in Stata.)[3]

23. Special care was given to the correct computation of the weights. Given the varying quality of the sample frames, it was imperative to accurately adjust the totals within each region/industry/size stratum to account for the presence of ineligible units (non-existing units, public establishments, establishments with less than 5 employees, and non-business units). 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.

24. For some units it was impossible to determine eligibility because the contact was not successfully completed. Consequently, different assumptions as to their eligibility result in different universe cells' adjustments and in different sampling weights. Three sets of assumptions were considered:

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

b- 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 *w_median*.

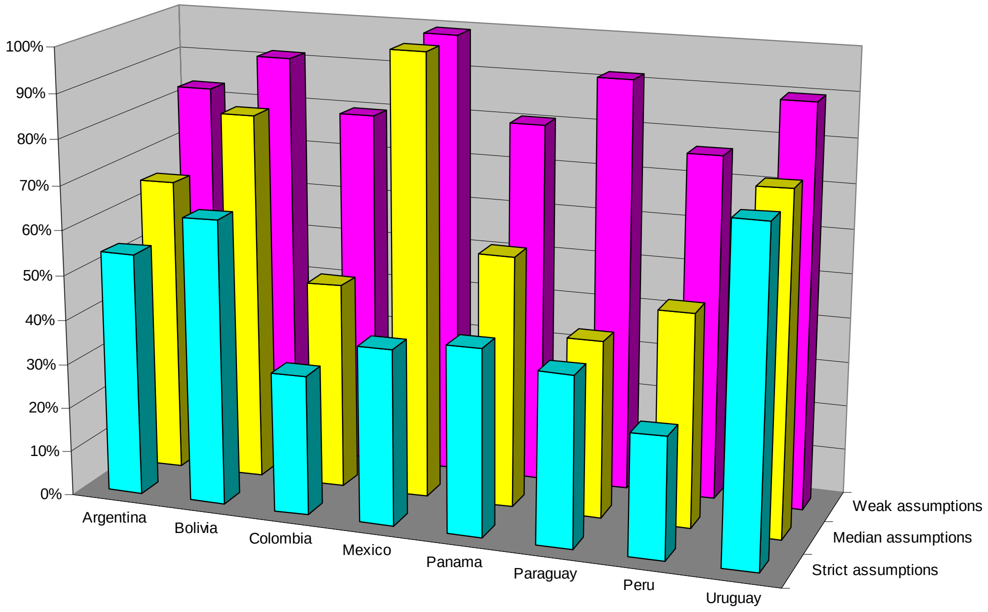
c- Weak assumption: in addition to the establishments included in points a and b, all establishments for which it was not possible to finalize a contact are assumed eligible. This includes establishments with dead or out of service phone lines, establishments that never answered the phone, and establishments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

The following graph exhibits the differences in eligibility rates under each set of assumptions. The sharp increase in eligibility for the case of Mexico between strict and median assumption is the result of an implementation variation: in this country the screener questionnaire was implemented along with the main questionnaire and, consequently, all rejections to participate were included as rejections to the screener. Consequently, for cross-country comparisons it is recommended that *w_median* be used, as they will include the same set of eligible establishments across countries[4].

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

[4] Using w_strict will penalize universe projections for Mexico vis a vis the other countries as all firms that rejected the survey in Mexico would be considered as non eligible, just because rejection took place during the actual appointment for the interview.

Differences in Eligibility Rate According to Assumptions



23. Within each of these assumptions regarding eligibility a pair (two) of weight sets was calculated. The first set of estimates calculated proportions using the raw sample count for each cell. However, the achieved sample numbers in many cells were small. Hence, those eligibility rates, and the adjusted universe cells projections, are subject to relatively large sampling variations. Therefore a second set of more robust estimates was also produced. These estimates made use of the multiples of the relative eligibility rates for each industry, size, and region. Those relative rates were based on much larger samples than the individual cells and thus produced values with smaller sampling variations. The data sets include only these robust weights.

4. Appropriate use of the weights

24. As discussed above, 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.

25. 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 has the advantage of providing an estimate that is independent of the sample design. This latter point may be quite relevant for the Enterprise Surveys 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.) [5]

26. 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. 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[6].

28. Since one disadvantage of stratified random sampling is that estimates for subpopulations may not be unbiased (Levy, Lemeshow, 1999, pp.150), and for comparative purposes because many other firm-level data sets in the World Bank were collected drawing exclusively from the manufacturing sector, a new set of weights is being computed to compute estimates for the manufacturing sector. They will be available, on request, in the near future.

5. Non-response

29. 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.

30. Item non-response was addressed by two strategies:

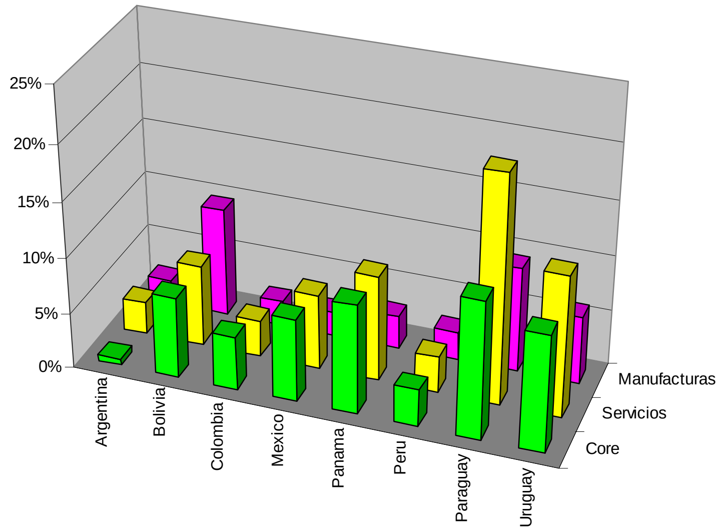
a- For sensitive questions that may generate negative reactions from the respondent, such as corruption or tax evasion, enumerators were instructed to collect the refusal to respond as a different option from don't know.

b- For information that establishments may consider too private to share such as financial information sample sizes were inflated by 25% to account for a margin of non-response. Each country was treated separately when considering non-response rates and special attention was paid to the variables needed to assess performance at the establishment level. In almost every country establishments with incomplete information were recalled to complete this information, whenever necessary. However, there were clear cases of low response. The following graph and table shows non-response rates for the sales variable, *d2,* by type of questionnaire for all 8 countries. With the exception of Paraguay and the services sector in Uruguay in all cases, non-response was kept below a 10% threshold.

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

[6] 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.

Sales Non-Response Rates

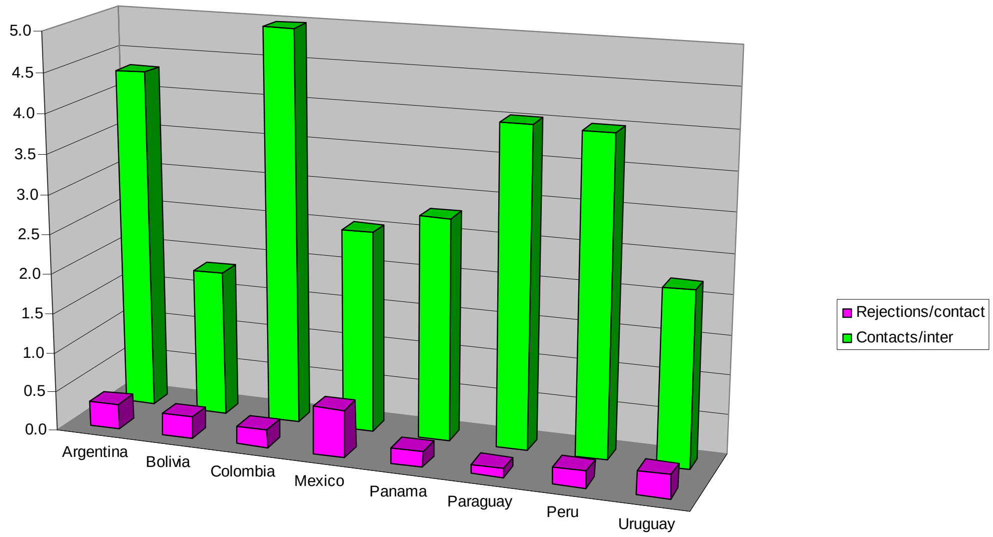


	Argentina	Bolivia	Colombia	Mexico	Panama	Peru	Paraguay	Uruguay
🟩 Core	0.60%	7.32%	4.80%	7.44%	9.84%	3.36%	12.38%	10.53%
🟨 Servicios	2.89%	7.32%	3.32%	6.75%	9.50%	3.25%	20.47%	12.70%
🟪 Manufacturas	2.14%	9.81%	2.21%	2.14%	2.92%	2.50%	9.45%	6.08%

31. Survey non-response was addressed by maximizing efforts to contact establishments that were first selected in the sample and by trying to keep a tight control over the process of substitutions. However, non-response of the complete survey was faced in every country to different degrees.

32. As the following graph and table show, the number of contacted establishments per realized interview varied from 1.8 in Bolivia to 5 in Colombia. 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. Consequently, it is not surprising that Mexico shows the highest rate of rejection and one of the lowest numbers of contacts per interview (the sample frame in Mexico was one of the most accurate frames in the whole study). The main source of error in estimates in Mexico may be selection bias and not frame inaccuracy. Colombia, on the other hand, shows an average rate of rejection and the highest number of contacts needed to obtain an interview. For Colombia, estimates should be qualified by the fact that the deficiencies of the sample frame are compounded by selection bias.

Rejections Rates and Contacts per Interview



	Argentina	Bolivia	Colombia	Mexico	Panama	Paraguay	Peru	Uruguay
🟪 Rejections/contact	31.8%	27.3%	22.9%	59.7%	19.6%	11.6%	21.6%	29.8%
🟩 Contacts/inter	4.3	1.8	5.0	2.5	2.8	4.0	4.0	2.2

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

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.