The India 2014 Enterprise Survey Data Set

I. Introduction

1. This document provides additional information on the data collected in India between June 2013 and December 2014. The objective of the Enterprise Survey is to gain an understanding of what firms experience in the private sector.

The Enterprise Surveys, through interviews with firms in the manufacturing and service sectors, capture data covering measures of firm performance, firm structure as well as business perceptions on the biggest obstacles to enterprise growth, and the business environment in general. They are used to create statistically important business environment indicators that are comparable across countries.

The 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

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

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

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

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.

---

1 The complete text can be found at http://www.enterprisesurveys.org/documents/Implementation_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

3. 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 the accompanying Implementation Report Excel file (sheets "" and ""). These tables still need to be inserted into the accompany documentation.

4. Industry stratification was designed in the way that follows: the universe was stratified into 11 manufacturing industries (food, textiles, chemicals, rubber/plastics, non-metallic mineral products, basic metals, fabricated metal products, machinery and equipment, electrical machinery and communications equipment, motor vehicles, and other manufacturing), and 7 services industries (construction, sales and repair of motor vehicles, wholesale, retail, hotels and restaurants, transportation/storage/communications, and IT).

5. 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 purposes, the number of employees was defined on the basis of reported permanent full-time workers. This seems to be an appropriate definition of the labor force since seasonal/casual/part-time employment is not a common practice, except in the sectors of construction and agriculture.

6. Regional stratification was defined in 23 states. Delhi was included as a state for stratification purposes, whereas the 6 other UT's were excluded from the survey project (due to their very small size and low contribution to national GDP). The states of Sikkim and Mizoram were also excluded as their share of national GDP was less than 0.1%. At the onset of the survey project, Andhra Pradesh included Telangana (hence Telangana was not considered as its own state). Note that the states of Arunachal Pradesh, Nagaland, Manipur, Tripura, and Meghalaya were combined and considered one state (Arunachal Pradesh et al) for the purposes of stratification. Note the Enterprise Survey focuses on major cities and the surrounding business area.

III. Sampling implementation

7. 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*).

8. Nielsen India was hired to implement the India 2014 Enterprise Survey.

9. For manufacturing establishments, the sample frame used for the survey in India was from the Annual Survey of Industries (Central Statistics Office of the Ministry of

Statistics and Programme Implementation, Government of India) January 2013. The database contained the following information

- Business name;
- Business address;
- Total numbers of employees
- NIC 2 digit code (business sector classification)

For services establishments, industry associations lists were used to create lists from which to randomly sample from.

Counts from the sample frame are provided in the accompanying Implementation Report Excel file (sheet "Manuf Sample Frame").

10. For services establishments, industry association lists were used to create lists from which to randomly sample from.

11. 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.

12. 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 14.9% (566 out of 3789 establishments)[4]. Breaking down by stratified industries, the following sample targets were achieved (using a4a and a6a):

Counts of the achieved interviews are provided in the accompanying Implementation Report Excel file (sheets "Manuf Achieved Interviews" and "Services Achieved Interviews").

IV. Data Base Structure:

13. The structure of the data base reflects the fact that 2 different versions of the questionnaire were used for 3 categories of businesses (manufacturing, retail, and other services/non-retail). The Manufacturing Questionnaire includes all common questions asked to all establishments and some specific questions relevant to manufacturing firms. The Services Questionnaire, administered to retail and other services/non-retail establishments, includes all common questions asked to all establishments and some specific questions relevant retail and other services firms. Each variation of the questionnaire is identified by the index variable, *a0*.

14. 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*. Variable

4 Based on out of target contacts and impossible to contact establishments

names proceeded by a prefix *"SAR"* or "IND" indicate questions specific to the South Asia region or India only, 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.

15. 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.

16. 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, whereas the latter gives the actual establishment's industry classification (four digit code) in the sample frame.

17. 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. The code -*9* was used to indicate units for which size was undetermined in the sample frame.

-*a4a*: coded using ISIC Rev 3.1 codes for the chosen industries for stratification. These codes include most manufacturing industries (15 to 37), retail (52), and (45, 50, 51, 55, 60-64, 72) for other services.

18. 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. 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.

19. Note that there are additional variables for location (*a3x*) and 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.

20. 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.

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.

22. Variables *a17x* gives interviewer comments, including problems that occurred during an interview and extraordinary circumstances which could influence results. Please note that sometimes this variable is removed due to privacy issues.

V. Universe Estimates

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

24. Overall estimates of the numbers of establishments in India are based on 1) the manufacturing sample frame and 2) for services, the 5th Economic Census. Refer to the accompanying Implementation Report Excel file (sheets "Manuf Universe…" and "Services Universe …").

25. 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.

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

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

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

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

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

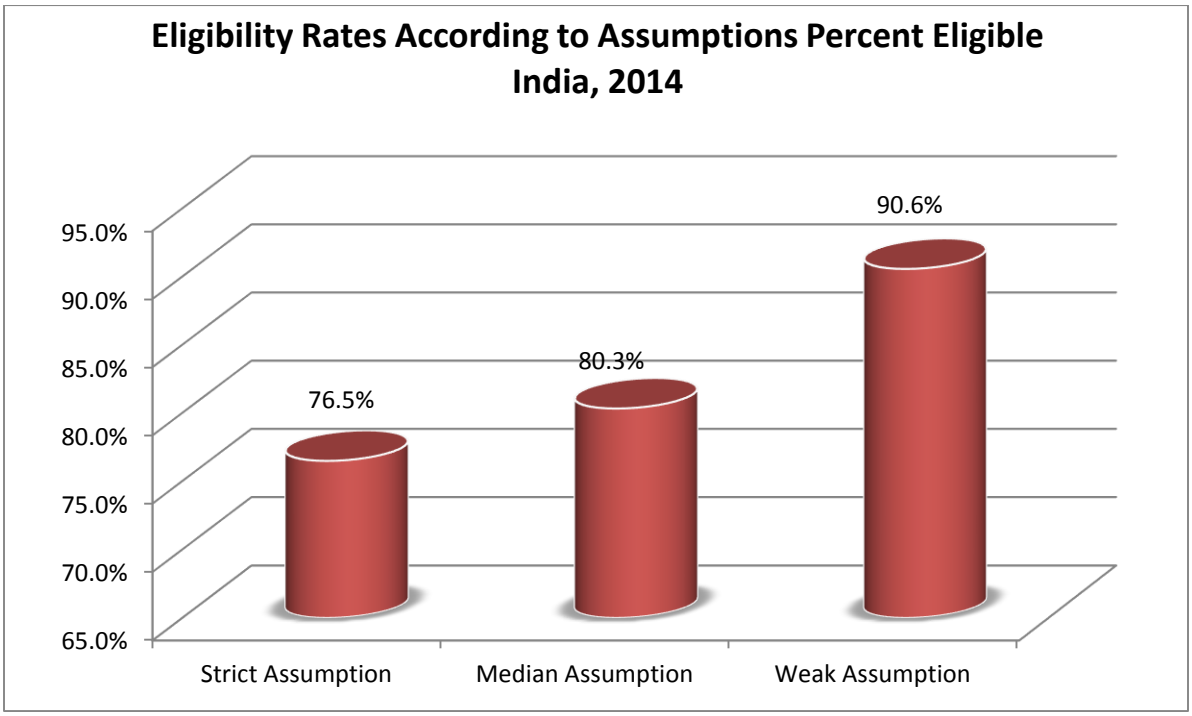
29. Weak assumption: in addition to the establishments included in points a and b, all establishments for which it was not possible to contact or that refused the screening questionnaire are assumed eligible. This definition includes as eligible establishments

with dead or out of service phone lines, establishments that never answered the phone, and establishments with incorrect addresses for which it was impossible to find a new address. Under the weak assumption only observed non-eligible units are excluded from universe projections. The resulting weights are included in the variable *wweak*.

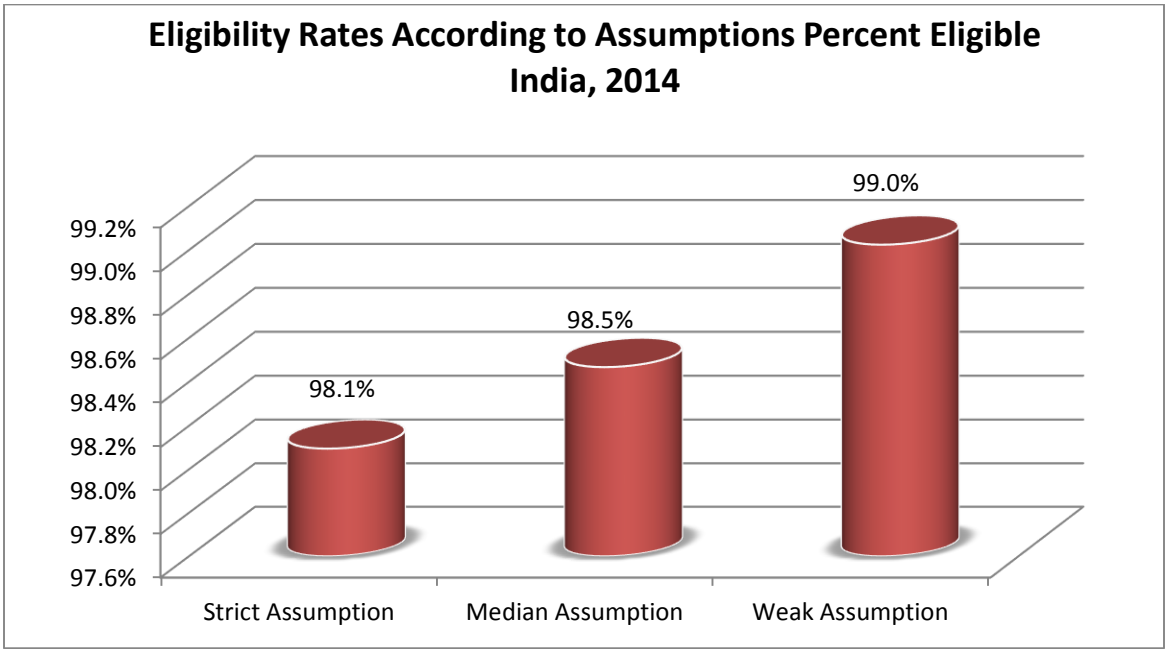
Weak eligibility= (Sum of the firms with codes 1,2,3,4,16,91,92,93,10,11,12,&13) / Total

30. The indicators computed for the Enterprise Survey 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.

Manufacturing firms:



Services firms:



31. Universe estimates for the number of establishments in each industry-region-size cell in India were produced for the strict, weak and median eligibility definitions. Refer to the accompanying Implementation Report Excel file (sheets "Manuf Universe Strict" and "Manuf Universe Medium" and "Manuf Universe Weak")

32. 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

33. 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]

34. 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 on the phone line, answering machine, or fax line[6], wrong address or moved away and could not get the new references). The information required for the

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

6 For the surveys that implemented a screener over the phone.

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.

35. Refer to the accompanying Implementation Report Excel file (sheets "Manuf Weights…" and "Services Weights") for the cell weights for registered establishments in India.

VII. Appropriate use of the weights

36. 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.

37. 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. 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.)[7]

38. 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

39. 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.

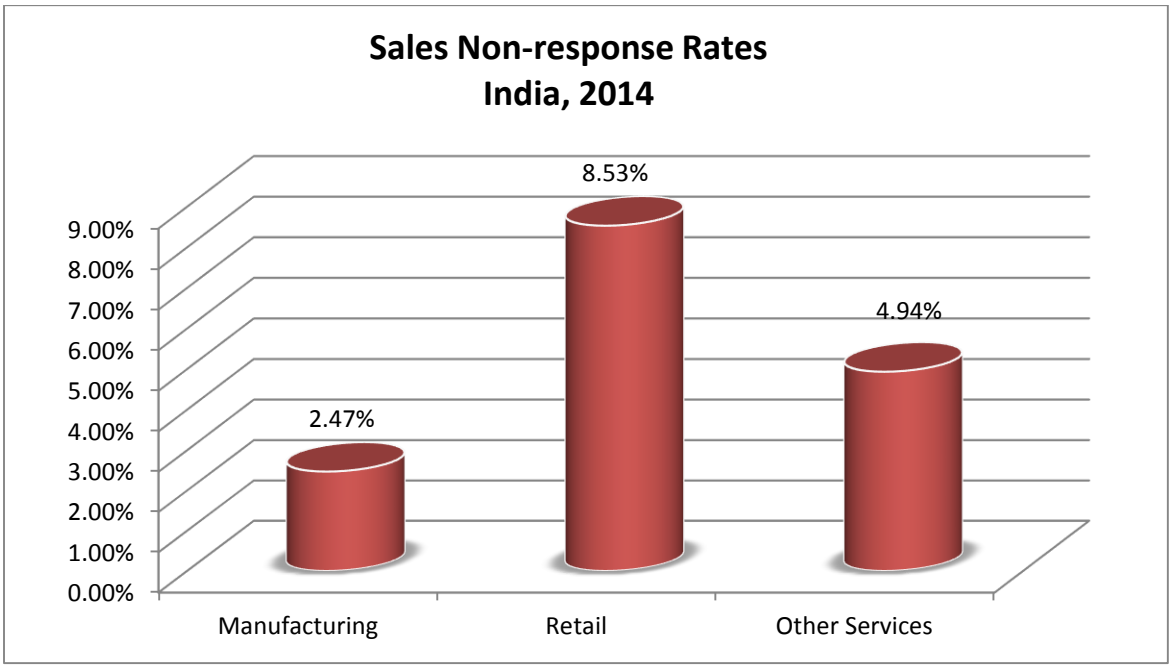
40. Item non-response was addressed by two strategies:

---

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 of 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.

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 (-8).
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 the coding utilized in this dataset does not allow us to differentiate between "Don't know" and "refuse to answer", thus the non-response in the chart below reflects both categories (DKs and NAs).



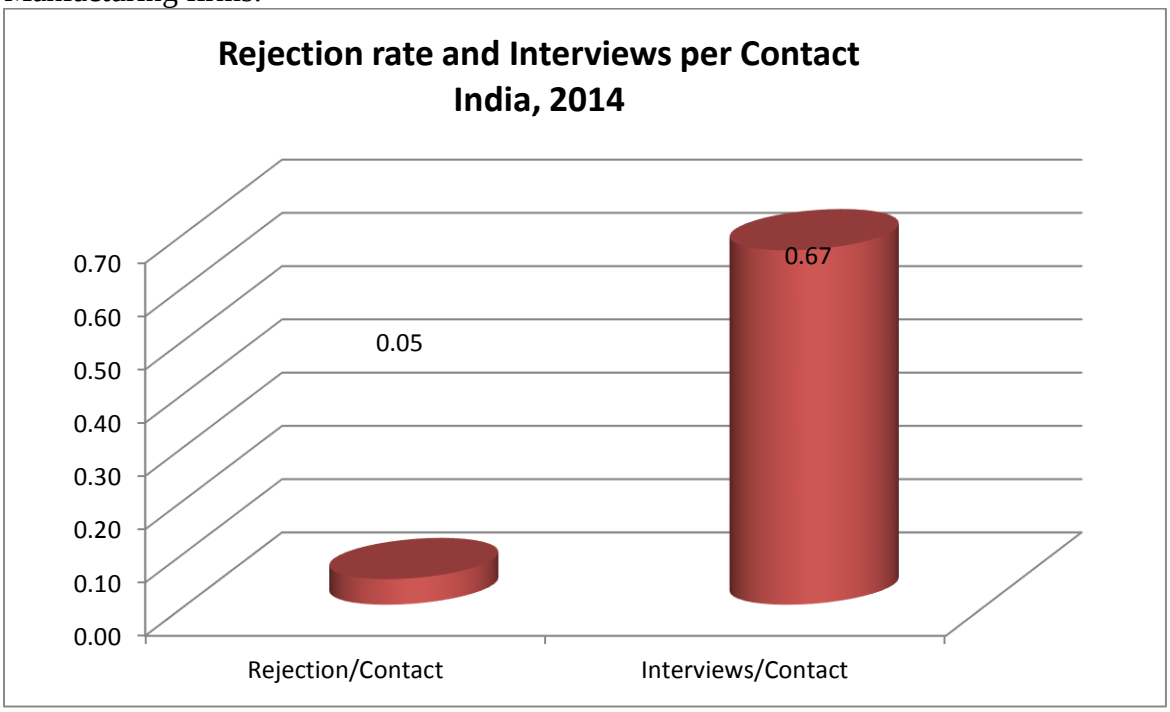
41. Survey non-response was addressed by maximizing efforts to contact establishments that were initially selected for interview. Attempts were made to contact the establishment for interview at different times/days of the week before a replacement establishment (with similar strata characteristics)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

42. As the following graph shows, the number of realized interviews per contacted establishment was 0.67[9] for both manufacturing and services firms. 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

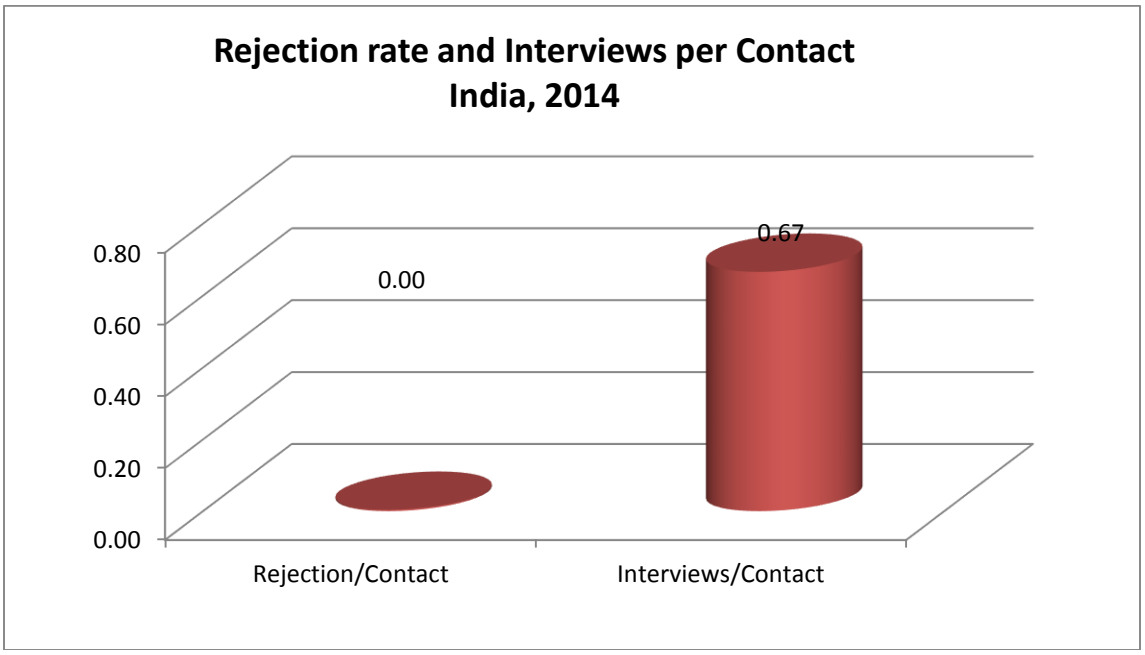
9 The estimate is based on the total number of firms contacted including ineligible establishments.

number of rejections per contact was 0.05 for manufacturing firms and 0 for services firms.

Manfacturing firms:



Services firms:



43. Details on the rejection rate, eligibility rate, and item non-response are available at the strata level. 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 India. 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.

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