Tajikistan Enterprise Surveys Data Set

1. Introduction

1. This document provides additional information on the data collected in Tajikistan during calendar year 2008 as part of the fourth round of the Business Environment and Enterprise Performance Survey, a joint initiative of the European Bank for Reconstruction and Development and the World Bank, in Tajikistan.

The objective of the survey is to obtain feedback from enterprises in client countries on the state of the private sector as well as to help in building a panel of enterprise data that will make it possible to track changes in the business environment over time, thus allowing, for example, impact assessments of reforms.

Through interviews with firms in the manufacturing and services sectors, the survey will assess the constraints to private sector growth and create statistically significant 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.

2. Sampling Structure

2. The sample for the Tajikistan 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.)

[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

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.

3. Three levels of stratification were used in this country: industry, establishment size, and oblast (region). The original sample designs with specific information of the industries and regions (oblasts) chosen are described in Appendix E

4. Industry stratification was designed in the way that follows: the universe was stratified into 20 manufacturing industries, 2 services industries -retail and IT-, and one residual sector as defined in the sampling manual. Each sector had a target of 120 interviews.

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).[4] 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 4 regions (oblasts). These regions are Capital, Sogdiskaya oblast, Khatlonskaya oblast,, and RRP.

3. Sampling implementation

7. Given the stratified design, sample frames containing a complete and updated list of establishments for the selected regions were required. 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. The source of the sample frame was the National Statistics Committee of Tajikistan (2008).

9. The quality of the frame was assessed at the onset of the project. The frame proved to be useful though it showed positive rates of non-eligibility, repetition, 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 percentage of confirmed non-eligible units as a proportion of the total number of contacts to complete the survey was 19% (126 out of 672 establishments).

[4] The panel firms from BEEPS 2005 with less than 5 employees are included in the 5 to 19 strata.

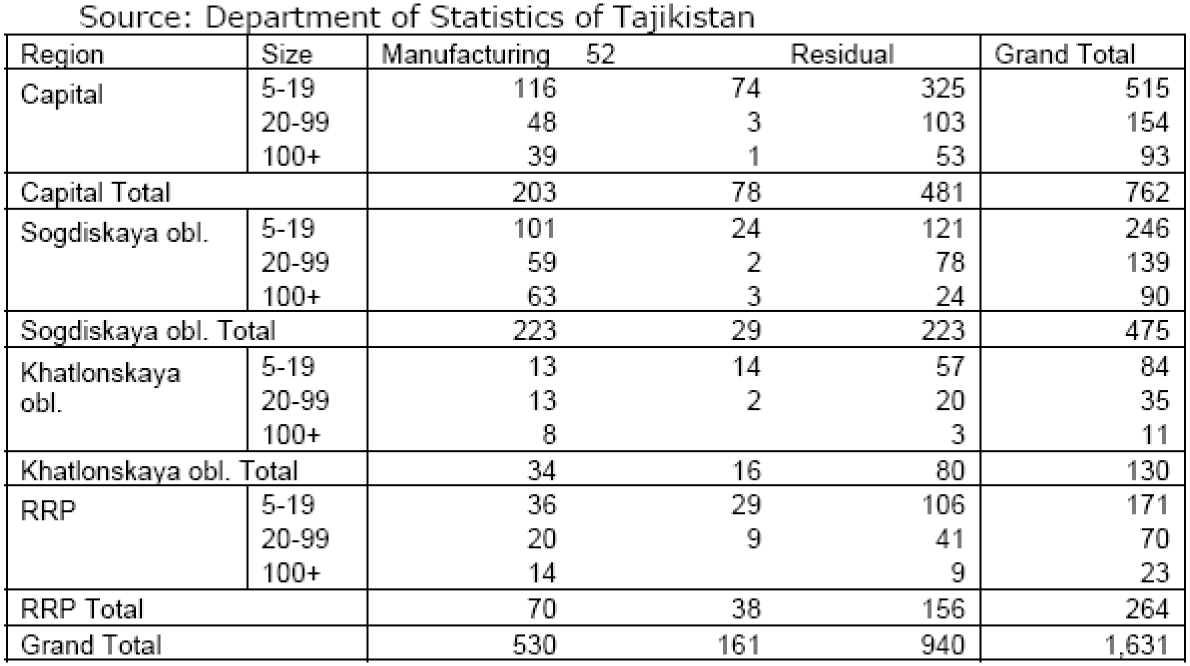
Local Agency team involved in the study:

Local Agency	Name: The Center of Sociological Research "Zerkalo" Country: Tajikistan Membership of international organization: No Activities since: 1999
Enumerators involved:	Enumerators: 37 Recruiters: 4 Some of the interviewers were involved in recruitment activities.
Other staff involved:	Fieldwork Coordinators: 5 Editing: 2 people Data Entry: 4 people Data Processing: N/A

Sample Frame:

Characteristic of sample frame used:	For the sample frame the Register of establishments of Tajikistan was used. The data was obtained from the National Statistics Committee of Tajikistan. The data base was issued in 2008, but the data is from 2007.
Source:	Register of establishments of Tajikistan, National Statistical Committee of Tajikistan
Year of publication:	2008
Comments on the quality of sample frame:	During the sample checking it turned out that only 11% of the addresses were valid. Whilst not ideal, this was still the only available frame and it was therefore used. The level of ineligible firms is dealt with in the universe estimation. It transpired during the survey that the database provided is based on information that was submitted by the businesses when they were established and no further follow-up information is available by the National Statistical Committee of Tajikistan. Due to this, several difficulties were encountered while finding businesses from the data base.

Sample Frame Tajikistan



Source: Department of Statistics of Tajikistan

Region	Size	Manufacturing	52	Residual	Grand Total
Capital	5-19	116	74	325	515
	20-99	48	3	103	154
	100+	39	1	53	93
Capital Total		203	78	481	762
Sogdiskaya obl.	5-19	101	24	121	246
	20-99	59	2	78	139
	100+	63	3	24	90
Sogdiskaya obl. Total		223	29	223	475
Khatlonskaya obl.	5-19	13	14	57	84
	20-99	13	2	20	35
	100+	8		3	11
Khatlonskaya obl. Total		34	16	80	130
RRP	5-19	36	29	106	171
	20-99	20	9	41	70
	100+	14		9	23
RRP Total		70	38	156	264
Grand Total		530	161	940	1,631

Sectors included in the Sample:

Original Sectors	Manufactures: *15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31,32,33, 35, 36* Services: *52* Residual: *45, 50, 51, 55, 60, 62, 63, 64*
Added Sectors	At the end of the fieldwork we used ISIC code 51 for the service sector to achieve the service target due to the shortage of addresses for services.

Sample:

Comments/ problems on sectors and regions selected in the sample:	*On sectors:* From the very beginning it was clear that for the services sector there were not enough addresses to achieve the required target. For example in the RRP region, there were no establishments that had 100 or more employees in the sample, but according to the requested target local institute was required to interview 8 establishments in this cell. In addition, in the remaining regions for the Services Sector there were establishments without enough preferences because many of the addresses were not valid and we also experienced several refusals. All of these factors prevented the timely completion of the fieldwork. *On regions:* The sample distribution according to the number of interviews among the regions was equal. However, not all of these regions are equally developed economically. This fact was not taken into consideration while designing the sample. This fact created problems for the fieldwork as interviewers lost much time searching for establishments in operation.
Comments on the response rate:	During the fieldwork, 678 establishments were contacted.
Comments on the sample design:	

Fieldwork:

Date of Fieldwork	May 1 - August 30 2008
Country	Tajikistan
Interview number	Manufactures: 116 Services: 151 Core:93
Problems found during fieldwork:	There were problems tracing the addresses from the sample. Also, there were problems getting appointments as the top managers were busy or had no desire to participate in the survey.
Other observations:	Due to the difficulties regarding the validity of the sample

	(addresses), some interviewers were dropped from the project.

4. Data Base Structure:

10. The structure of the data base reflects the fact that 3 different versions of the questionnaire were used. The basic questionnaire, the Core Module, includes all common questions asked to all establishments from all sectors (manufacturing, services and IT). The second expanded variation, the Manufacturing Questionnaire, is built upon the Core Module and adds some specific questions relevant to the sector. The third expanded variation, the Services Questionnaire, is also built upon the Core Module and adds to the core specific questions relevant to either retail or IT. Each variation of the questionnaire is identified by the index variable, *a0.*

11. 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 names preceded by a prefix "*ECA"* indicate questions used in the previous rollout (2005) 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.

12. 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 *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.

13. As noted above, there are 3 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 in the sample frame.

14. All of 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 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 (oblasts)

-*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 codes for the chosen industries for stratification. These codes include most manufacturing industries (15 to 36), and retail, and IT for services (52, and 72 respectively). All establishments within the residual stratum were coded with a4a=2.
-id2005: The variable contains the firm ids of the panel firms

15. 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*contain additional information and were also collected in the screening phase.

16. Note that there are additional variables for location (*a3x*), industry (*d1a2*), and size (*l1*, *l6* and *l8*) that reflect more accurately the reality of each establishment. Advance users are advised to use these variables for analytical purposes.

17 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.

18. 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.

19. 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.

5. Universe Estimates

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

21. Appendix C shows the overall estimates of the numbers of establishments based on the strict, weak and median relative estimates.

6. 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.)[5]

23. Special care was given to the correct computation of the weights. Considering 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 (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, out of order, no tone in the phone line, answering machine, fax line, 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.

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:

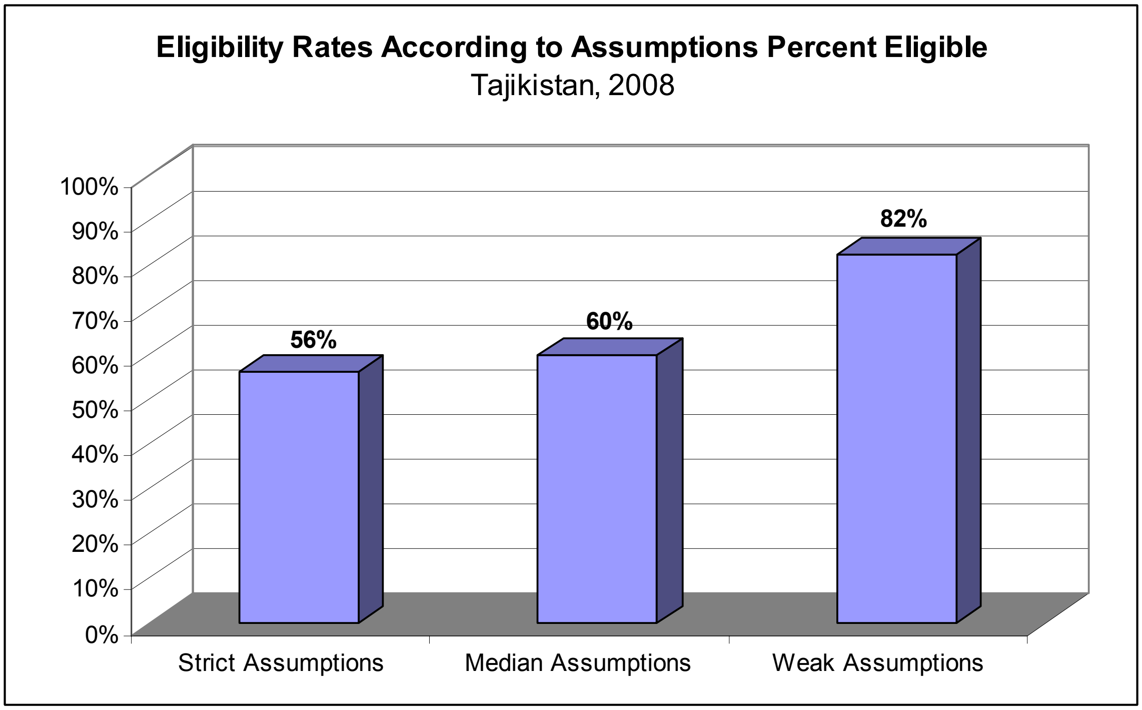
a- Strict assumption: eligible establishments are only those for which it was possible to directly determine eligibility. The resulting weights are included in the variable *w_strict*.

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 different eligibility rates under each set of assumptions.

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



25. Within each of these assumptions regarding eligibility a pair 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 (collapsed weights) 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.

Please note that for the purpose of the weights computations all panel firms were considered to be part of the current universe, although technically they are not randomly selected.

7. Appropriate use of the weights

26. 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.

27. 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.) [6]

28. 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[7]. 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.

8. 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.

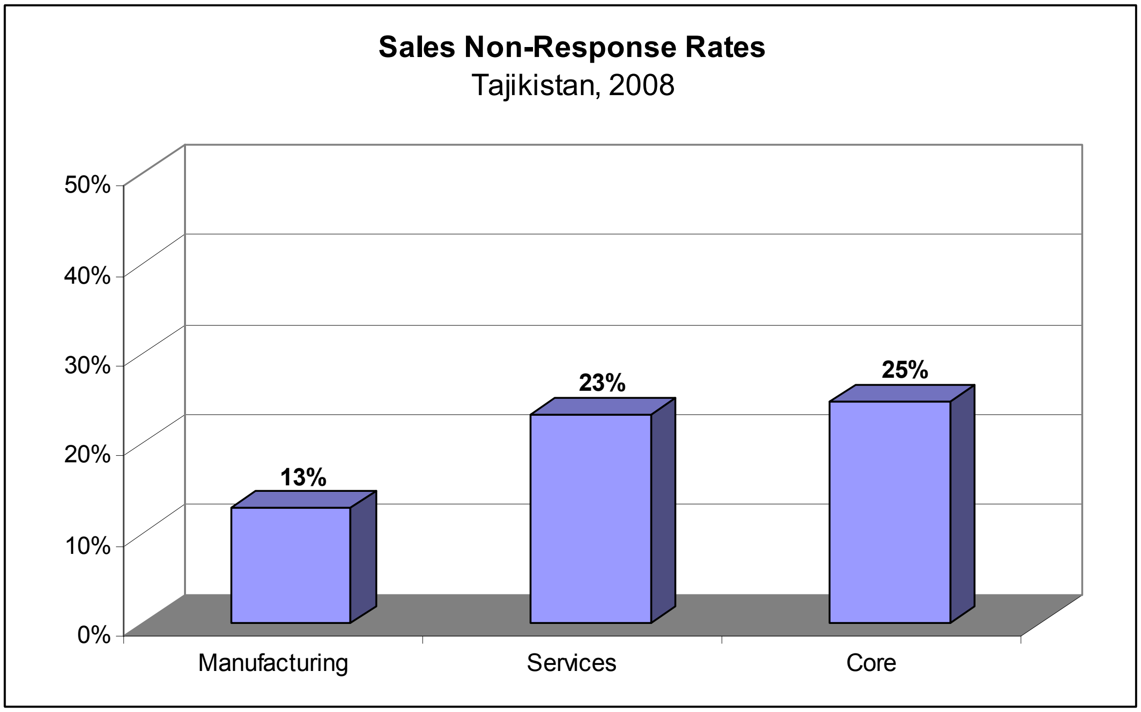
30. Item non-response was addressed by two strategies:

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 (-7).

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 type of questionnaire. Please, note that the coding utilized in this dataset does not allow us to differentiated between "Don't know" and "refuse to answer", thus the non-response in the table below reflects both categories (DKs and NAs).

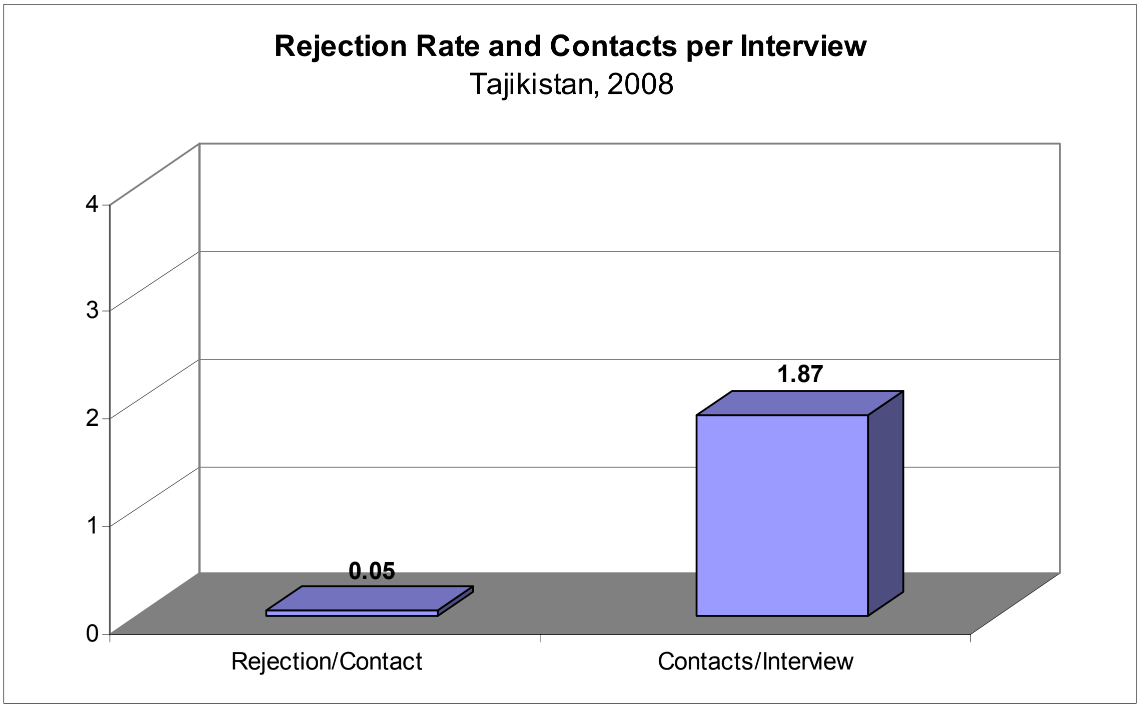
[6] Note that weighted OLS in Stata using the command regress with the option of weights will estimate wrong standard errors. Using the Stata survey specific commands svy will provide appropriate standard errors.

[7] 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.



31. Survey non-response was addressed by maximizing efforts to contact establishments that were initially selected for interview. Up to 4 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.

32. As the following graph shows, the number of contacted establishments per realized interview was 1.87. 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 relatively low ratio of contacted establishments per realized interview (1.87) suggests that the main source of error in estimates in the Tajikistan may be selection bias and not frame inaccuracy.



33. Details on rejections rates, eligibility rates, 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 the Tajikistan. 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.

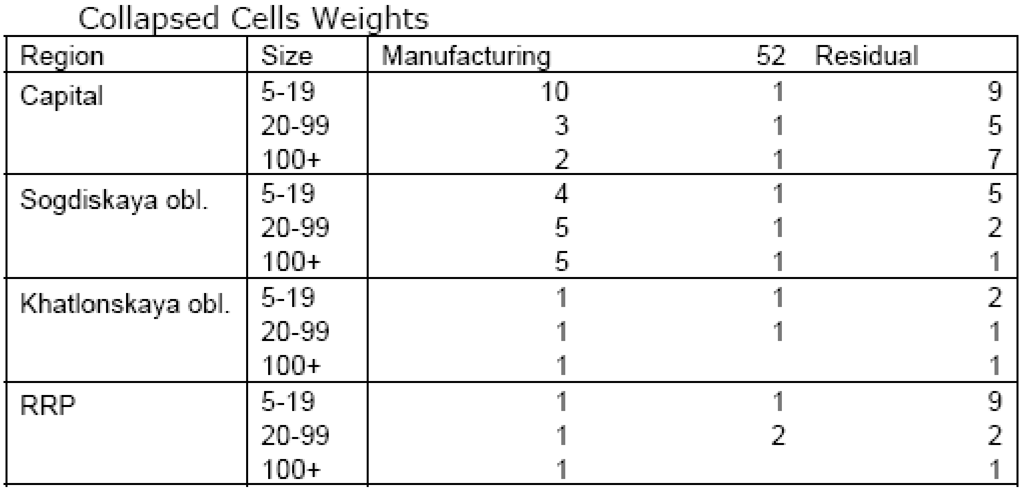
Levy, Paul S. and Stanley Lemeshow, Sampling of Populations: Methods and Applications, 1999.

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

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

Appendix A

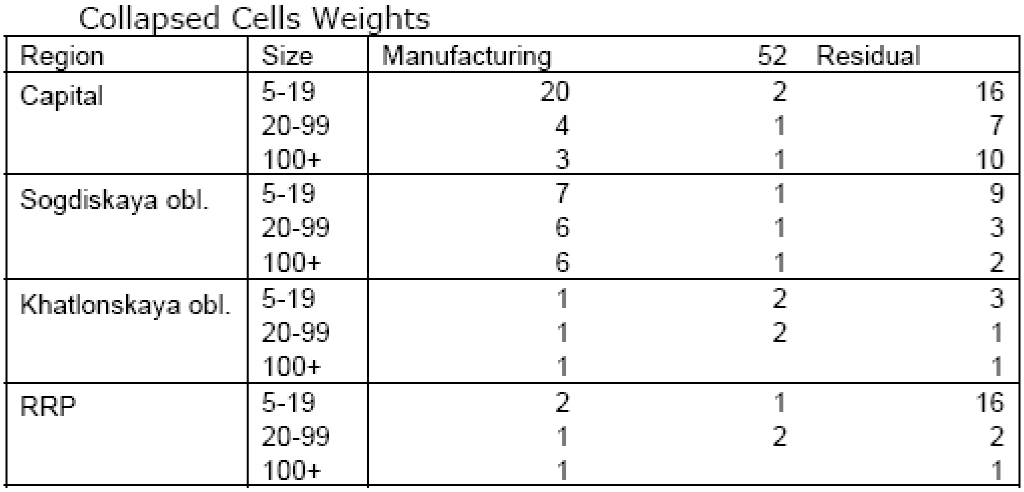
Tajikistan Strict Weights



Collapsed Cells Weights

Region	Size	Manufacturing	52	Residual
Capital	5-19	10	1	9
	20-99	3	1	5
	100+	2	1	7
Sogdiskaya obl.	5-19	4	1	5
	20-99	5	1	2
	100+	5	1	1
Khatlonskaya obl.	5-19	1	1	2
	20-99	1	1	1
	100+	1		1
RRP	5-19	1	1	9
	20-99	1	2	2
	100+	1		1

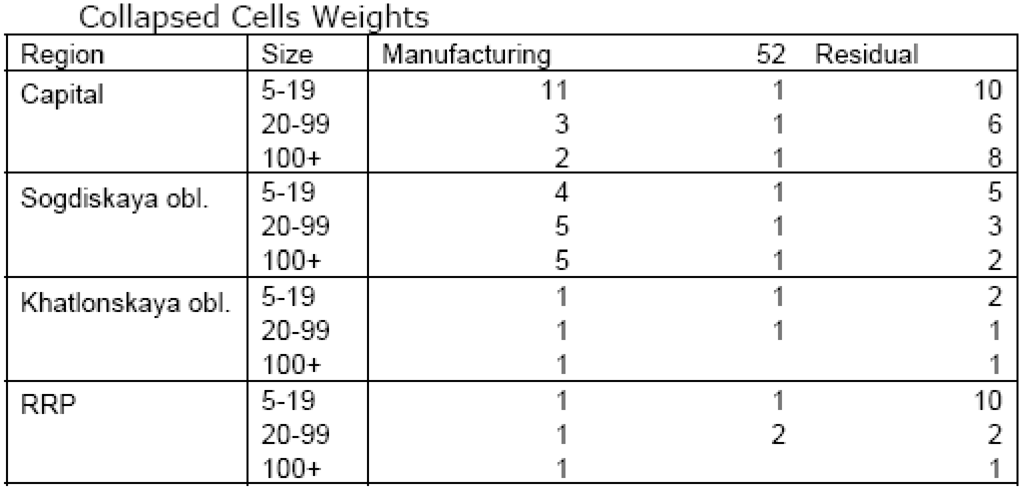
Tajikistan Weak Weights



Collapsed Cells Weights

Region	Size	Manufacturing	52	Residual
Capital	5-19	20	2	16
	20-99	4	1	7
	100+	3	1	10
Sogdiskaya obl.	5-19	7	1	9
	20-99	6	1	3
	100+	6	1	2
Khatlonskaya obl.	5-19	1	2	3
	20-99	1	2	1
	100+	1		1
RRP	5-19	2	1	16
	20-99	1	2	2
	100+	1		1

Tajikistan Median Weights

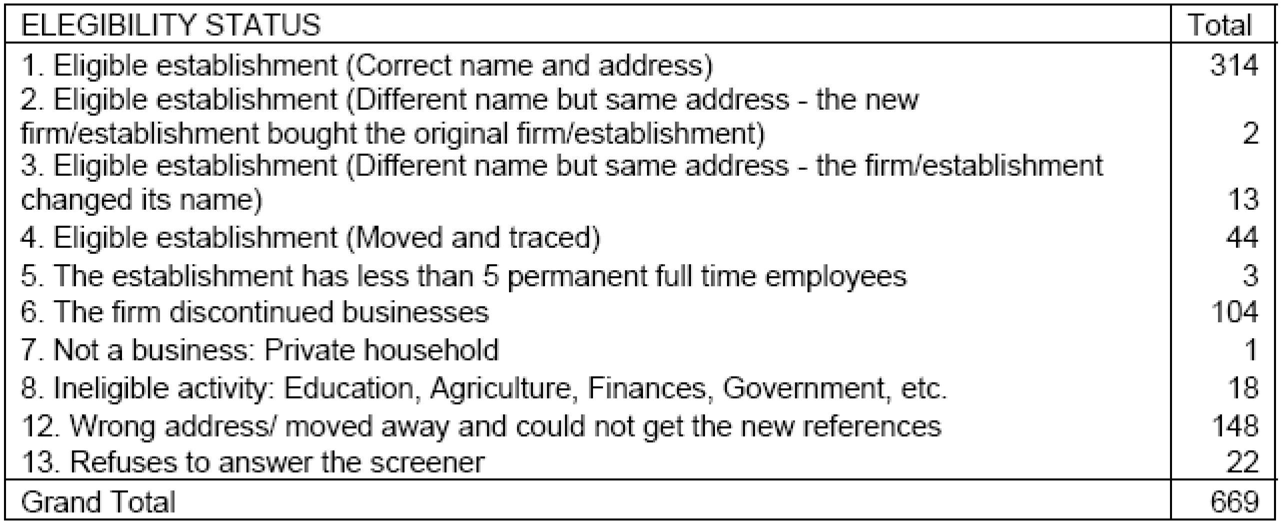


Collapsed Cells Weights

Region	Size	Manufacturing	52	Residual
Capital	5-19	11	1	10
	20-99	3	1	6
	100+	2	1	8
Sogdiskaya obl.	5-19	4	1	5
	20-99	5	1	3
	100+	5	1	2
Khatlonskaya obl.	5-19	1	1	2
	20-99	1	1	1
	100+	1		1
RRP	5-19	1	1	10
	20-99	1	2	2
	100+	1		1

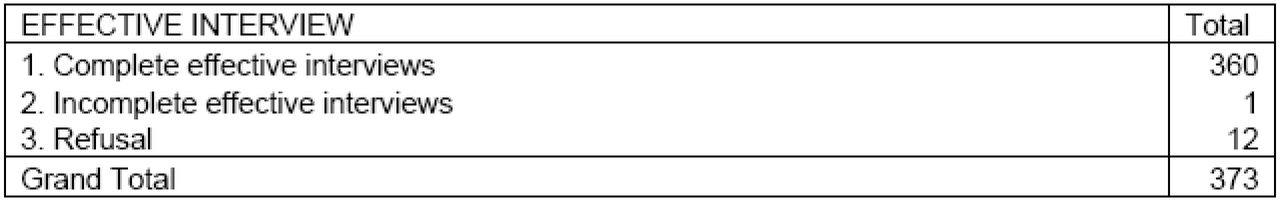
Appendix B

Status Codes



ELEGIBILITY STATUS	Total
1. Eligible establishment (Correct name and address)	314
2. Eligible establishment (Different name but same address - the new firm/establishment bought the original firm/establishment)	2
3. Eligible establishment (Different name but same address - the firm/establishment changed its name)	13
4. Eligible establishment (Moved and traced)	44
5. The establishment has less than 5 permanent full time employees	3
6. The firm discontinued businesses	104
7. Not a business: Private household	1
8. Ineligible activity: Education, Agriculture, Finances, Government, etc.	18
12. Wrong address/ moved away and could not get the new references	148
13. Refuses to answer the screener	22
Grand Total	669

Response Outcomes



EFFECTIVE INTERVIEW	Total
1. Complete effective interviews	360
2. Incomplete effective interviews	1
3. Refusal	12
Grand Total	373

Appendix C

Eligibility Rules

Status Code	Eligibility Criteria		
	Strict	Weak	Median
1. Eligible establishment (Correct name and address)	1	1	1
2. Eligible establishment (Different name but same address - the new firm/establishment bought the original firm/establishment)	1	1	1
3. Eligible establishment (Different name but same address - the firm/establishment changed its name)	1	1	1
4. Eligible establishment (Wrong address - the firm/establishment has changed address and the address could be found)	1	1	1
5. The establishment has less than 5 employees	0	0	0
6. The firm discontinued businesses/ unattainable	0	0	0
7. Not a business: Private	0	0	0
8. Not a business: Education or Government	0	0	0
9. No reply (after having called in different days of the week and in different business hours) out of order, no tone	0	0	1
10. Answering machine	0	1	1
11. Fax line	0	1	1
12. Wrong address/ moved away and could not get the new references	0	0	1
13. Refuses to answer the screener	0	1	1
14. In process (the establishment is being called/ is being contacted - previous to ask the screener)	0	0	0
15. Out of target - cooperative, outside the covered regions	0	0	0

Tajikistan Establishment Estimates

Cells	Strict	Weak	Median
Un-collapsed Cells	910	1357	970
Collapsed Cells	916	1342	976

Appendix D

Questionnaires:

Problems for the understanding of questions	Response rate for the questions n6 and n7 were low, because respondents didn't want to answer these questions Services questionnaire TJTJ: the term inventory in d17 was not always perceived correctly, because this word has another meaning also.
Problems found in the navigability of – questionnaires (for example, skip patterns).	No special problems encountered
Comments on questionnaires length:	The questionnaire is too long; the average duration of the interview is 70 minutes. Respondents were tired during the interview and became irritated.
Suggestions or other comments on the questionnaire:	

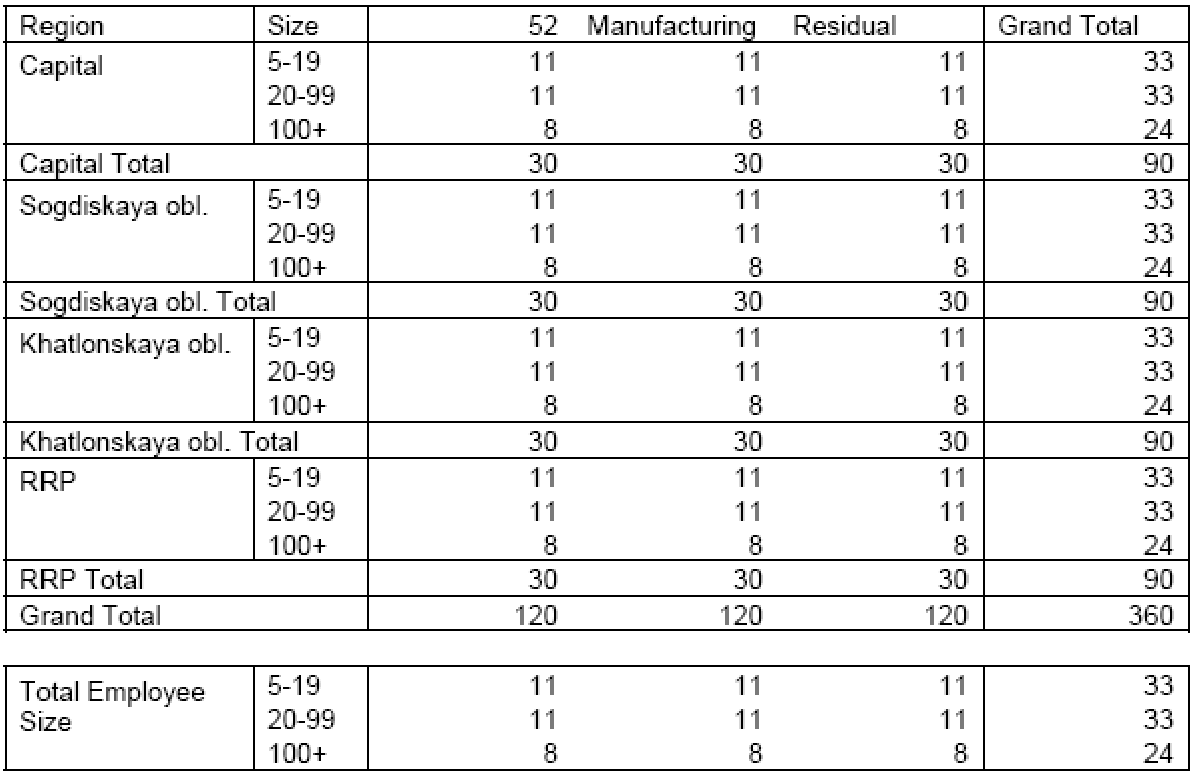
Database

Comments on the data map	None
Comments on the data processing	Data entry program chosen: **CONFIRMIT**

Country situation

General aspects of economic, political or social situation of the country that could affect the results of the survey:	The economic situation is characterized by high administrative regulation, state intervention in business affairs, high level of corruption and the 'shadow' economy. All these factors are causing negative attitudes towards survey research among the business community. Businessmen are negatively disposed to any questions from third parties and are not willing to share information and find it difficult to recognize the benefit of partaking in the survey. Thus, conducting B2B survey among Tajik businesses is considerably difficult.
Relevant country events occurred during fieldwork:	In June, 2008 the president announced a moratorium for two years to inspect privately-owned businesses in Tajikistan by the tax authorities.
Other aspects:	

Appendix E
Original Sample Design



Region	Size	52	Manufacturing	Residual	Grand Total
Capital	5-19	11	11	11	33
	20-99	11	11	11	33
	100+	8	8	8	24
Capital Total		30	30	30	90
Sogdiskaya obl.	5-19	11	11	11	33
	20-99	11	11	11	33
	100+	8	8	8	24
Sogdiskaya obl. Total		30	30	30	90
Khatlonskaya obl.	5-19	11	11	11	33
	20-99	11	11	11	33
	100+	8	8	8	24
Khatlonskaya obl. Total		30	30	30	90
RRP	5-19	11	11	11	33
	20-99	11	11	11	33
	100+	8	8	8	24
RRP Total		30	30	30	90
Grand Total		120	120	120	360

Total Employee Size	5-19	11	11	11	33
	20-99	11	11	11	33
	100+	8	8	8	24