Digitalization in MENA region:

Sampling design and survey weights

This document describes sampling design and survey weighs of the survey of digitalization in MENA region. The two topics will be addressed for the three countries where the survey was collected: Egypt, Jordan and Morocco.

Sampling design

Target population and Sampling frame

The target population of the surveys was businesses in all sizes that started business operations before 2022. Businesses that started their operations during 2022 were not eligible for the survey. Different sampling frames were used across countries. An ideal sampling frame for a probability sample should covers all target population units, i.e., a list of all working businesses that started operations before 2022 with their telephone numbers. Unfortunately, finding such frames is not an easy task. Therefore, available lists, such as lists compiled by Yellow Pages, were used instead. Although the used lists do not necessarily cover all businesses in the study countries, they are large enough and cover many business sectors. Unfortunately, we could not assess the coverage of those lists especially with the lack of official numbers about our target population in the study countries. See Table 1 below for more details about the sampling frames.

Table 1: sampling frames by country

Country	Sampling frame /Description
Egypt	Yellow Pages (YP) - (https://www.yellowpages.com.eg/en) includes data about 300,000 businesses from more than 600 business sectors. The data is available on the website. List of all businesses was compiled on an excel sheet to facilitate the sampling design and selection. The list included business names, addresses, telephone numbers, and websites (if available).
Jordan	Kinz - (https://kinz.jo/) a website for a data mining Jordanian corporate, which had a larger list of businesses than the Yellow Pages in Jordan. The website grant access to subscribers. We had access to the complete list of about 102,000 businesses from 17 broad business sectors, such as Agricultural & Farming, Arts, Sports & Entertainments, Commerce, Education, …, etc. Lists of businesses are presented in about 10,200 pages (10 businesses per page). Unlike the case of the YP in Egypt, the sample selection was directly done on the Kinz website, where a random sample of pages within each business sector was selected for the survey.

Country	Sampling frame /Description
	The frame included business names, addresses and telephone numbers.
Morocco	Yellow Pages (Télécontact) - (https://www.telecontact.ma/) includes data about 91,000 businesses, 56,000 from the Casa-Rabat region, 20,000 from the North region, and 15,000 from the South region. The available digital copy of the list could not be readily used for sampling purposes. A pdf version of the frame was used instead. The pdf version lists different business contacts sorted by geographic region. Within each region, businesses were organized by business sector. Random samples of pages were selected within each region, and systematic random samples of businesses was selected from each for the survey. The frame included business names, addresses and telephone numbers.

Sample size

In all countries, samples were designed to represent E-firms and non-E-firms. *E-firms were defined as businesses used the internet during the last month before the interview date to conduct transactions, such as buying and/or selling of goods or services*. Samples were designed to complete 250 interviews with E-firms and 500 interviews with non-E-firms in each country. Since no data were available about the proportion of E-firms in any of the study countries, a two-phase sampling design was used. Data from the first phase was used to guide the sampling procedures in the second phase so the target sample size is achieved, especially for E-firms.

In all countries, the data collection was executed in two phases, with n_{de} denotes the target sample of completed interviews with E-firms, and n_{do} denotes the target sample of completed interviews with non-E-firms. According to the survey objectives, $n_{de} = 250$ *and* $n_{do} = 500$. In Phase 1, a stratified random sample of businesses was selected, their phone numbers were called, and eligible businesses were interviewed; eligible businesses include all businesses who started business operations before 2022. During the interviews, full interviews were collected from all responding businesses. Filter questions were used to identify E-firms and non-E-Firms.

After data collection in Phase 1, the un-weighted percentage of E-firms was calculated as $e_f = n_{1e}/n_1$, where n_{1e} and n_1 are total number of E-firms and all businesses interviewed in Phase 1, respectively. Sampling design and sample size for Phase 2 was then guided by e_f from Phase 1. In Phase 2, a stratified random sample of businesses was selected, their phone numbers were called, and eligible businesses were interviewed. Enough sample size was selected to achieve a total of n_2 interviews with eligible businesses, where

$$n_2 = \frac{n_{de}}{e_f} - n_1$$

During the data collection for Phase 2, interviews were collected with all responding E-firms. Whereas complete interviews were collected only from a subsample of non-E-firms, and the remaining non-E-firms were screened-out and not interviewed. The subsampling factor was calculated as follows

$$f_{2o} = \frac{n_{do} - n_{1o}}{n_2 - n_{de}}$$

Table 2 presents details about sample size by phase and country.

Table 2: Number of selected and interviewed businesses by phase and country

Phase	Egypt	Jordan	Morocco
Phase 1			
Selected numbers	9087	3134	9099
Interviewed E-firms	82	126	67
Interviewed non-E-firms	447	657	447
Phase 2			
Selected numbers	18901	3160	25120
Interviewed E-firms	162	133	185
Interviewed non-E-firms	115	0	108
Screened-out non-E-firms	1839	313	4149
Total			
Interviewed E-firms	224	259	252
Interviewed non-E-firms	562	657	555

Sample design and selection

Due to the different structures and available data across countries, different sample designs were adopted as follows:

1. Egypt:
   a systematic sample of 9,087 businesses were selected in Phase 1. To align the sample distribution across business sectors with the distribution in the frame distribution, the selection was done after sorting the frame according to the business sectors. The sample of Phase 1 composed of two samples: main sample of 5,066 businesses and a supplement sample of 4,021 businesses. Businesses with website available on YP were oversampled to increase the chance of finding E-firms. This has been accounted for during the weight calculation to retrieve the actual distribution in YP frame. In Phase 2, a systematic sample of 18,901 businesses were selected from businesses with website available on YP.

2. Jordan:
   a stratified sample of 3,134 businesses were selected in Phase 1. The sample was stratified according to 17 business sectors. The sample was selected in two stages; in the first stage, 316 pages were selected from the Kinz website, where the selected pages were proportionally distributed to the distribution of all pages by business sector. In the second stage, all businesses in selected pages were contacted. In average, there are 10 businesses per page. In Phase 2, a stratified sample of 3,160 businesses were selected using the same

approach used in Phase 1. See Table 3 for sample allocation of pages by business sector and design phase.

3. Morocco:
a stratified sample of 9,099 businesses were selected in Phase 1. The sample was stratified according to three main geographic regions of Morocco, Casa-Rabat, North, and South. The sample of Phase 1 composed of two samples: main sample of 4,388 businesses and a supplement sample of 4,711 businesses. In the two samples, 158 pages were selected for each sample, and 28-30 businesses were contacted from each page. In Phase 2, a stratified sample of 25,120 businesses were selected in two stages, where 314 pages were selected in the first stage, and about 80 businesses were contacted per each selected page. See Table 4 for sample allocation of pages and businesses by region and design phase.

Table 3: Sample allocation of pages by business sector and design phase: Jordan

Business sector	Pages in sampling frame	Pages selected	
		Phase 1	Phase 2
Agricultural & Farming	43	2	2
Arts, Sports & Entertainments	99	4	4
Commerce	4161	124	124
Education	697	22	22
Engineering, Contracting and Real estate	582	18	18
Financing and Insurance	70	4	4
Health & Social work Activities	1087	32	32
Hospitality, Travel and Tourism	881	26	26
Industry	760	24	24
IT	131	4	4
Law Firms	268	8	8
Marketing	119	4	4
Mining & Quarrying	13	2	2
Printing & Publishing	96	4	4
Professionals, Scientific & Technical Activities	596	18	18
Support Services activities	340	12	12
Transportation and Shipping	253	8	8
Total	10196	316	316

Table 4: Sample allocation of pages and businesses by region and design phase: Morocco

Regions	Casa-Rabat	North	South	Total
Pages in sampling frame	355	149	126	630
Phase 1: Main sample				
Selected pages	89	37	32	158
Selected businesses	2482	1028	878	4388
Phase 1: Supplement sample				
Selected pages	89	37	32	158
Selected businesses	2655	1098	958	4711
Phase 2				
Selected pages	177	75	62	314
Selected businesses	14160	6000	4960	25120

Survey implementation

Up to three calls were attempted to contact phone numbers that did not answer or busy lines. Table 5 presents the distribution of the selected businesses according to the final status after the three attempts.

Table 5: Distribution of selected samples according to the final contact result by country and design phase

Final contact result	Egypt			Jordan			Morocco		
	Phase1	Phase2	All	Phase1	Phase2	All	Phase1	Phase2	All
1. Phone disconnected/busy	504	874	1378	81	129	210	242	884	1126
2. Not in service	2426	6371	8797	485	224	709	2474	7104	9578
3. Did not answer	950	1235	2185	178	331	509	736	1832	2568
4. Picked up & refused	2219	3838	6057	989	978	1967	2151	5889	8040
5. Incomplete & refused	1277	2588	3865	134	751	885	1341	3871	5212
6. Incomplete & call returned	56	128	184	2	0	2	3	6	9
7. Complete	529	277	806	783	133	916	514	293	807
8. Ineligible	7	4	11	29	12	41	4	0	4
9. Government	0	3	3	0	2	2	0	1	1
10.Cancel - respondent didn't know enough	0	2	2	0	3	3	0	3	3
11.Cancel - interviewers mistakes	0	3	3	0	4	4	0	8	8
12.Unable to reach eligible person	1106	1739	2845	453	280	733	1634	1080	2714
13.Screened-out non-E-firms	0	1839	1839	0	313	313	0	4149	4149
Total	9074	18901	27975	3134	3160	6294	9099	25120	34219

Survey weights

Due to the different sampling designs adopted across countries, different procedures were used to calculate survey weights as follows:

1. Egypt:
   The weight calculations started by calculating design weights that reflect the selection probabilities of selecting the businesses from the sampling frame. The design weights accounted for the oversampling of businesses with websites and for the subsampling of non-E-firms in Phase 2. Selection probabilities of selected businesses were calculated as:

   $$p_{hij} = \frac{n_{hi}}{N_h} f_{2o}$$

   where p_{hj} is the selection probability of a business j selected in phase i from stratum h (stratum 1: businesses with website on YP; stratum 2: businesses without website on YP), n_{hi} is the number of businesses selected in phase i from stratum h, N_h is the total number of businesses from stratum h in the YP frame, and f_{2o} is the subsampling fraction of non-E-firms in Phase 2, as defined earlier under the sample size section. Note that f_{2o} = 1 where i = 1 or if j is an E-firm. The inverse of the selection probability is the design weight as follows:

   $$W_{hij}^{0} = \frac{1}{p_{hij}}$$

   The design weights were adjusted for nonresponse among eligible phone numbers, including numbers without known eligibility. Eligible cases are defined in Table 6. A nonresponse adjustment factor was calculated based on data from Phase 1 and was used to adjust all data from phases 1 and 2. The nonresponse factor was calculated as the inverse of the weighted response rates by business sectors as follows:

   $$A_c = \frac{\sum_{c=1}^{E_c} W_{cj}^{0}}{\sum_{c=1}^{E_c} W_{cj}^{0} R_{cj}}$$

   where E_c is the number of eligible businesses in business sector c, R_{cj} identifies the completed businesses among eligible businesses, where R_{cj} = 1 for businesses who completed the survey and R_{cj} = 1 otherwise. R_{cj} by final status are defined in Table 6. The adjusted weight for nonresponse was then calculated as:

   $$W_{hij}^{1} = W_{hij}^{0} A_c$$

   The survey weight was then calculated as a normalized version of the adjusted weight for nonresponse as follows:

   $$W_{hij}^{2} = \frac{W_{hij}^{1} n_{comp}}{\sum_{j=1}^{n_{comp}} W_{cj}^{1}}$$

where n_{comp} is the total number of businesses completed the survey.

Table 6: Final contact results by eligibility status

Final contact results	Eligibility	R_{cj}
1. Phone disconnected/busy	Eligible non-respondent	0
2. Not in service	Ineligible	NA
3. Did not answer	Eligible non-respondent	0
4. Picked up & refused	Eligible non-respondent	0
5. Incomplete & refused	Eligible non-respondent	0
6. Incomplete & call returned	Eligible non-respondent	0
7. Complete	Eligible respondent	1
8. Ineligible	Ineligible	NA
9. Government	Ineligible	NA
10.Cancel - respondent didn't know enough	Eligible non-respondent	0
11.Cancel - interviewers mistakes	Eligible non-respondent	0
12.Unable to reach eligible person	Eligible non-respondent	0
13.Screened-out non-E-firms	Ineligible	NA

2. Jordan:

The weight calculations started by calculating design weights that reflect the selection probabilities of selecting the businesses from the sampling frame. As described earlier, frame pages were selected within each business sector, and all businesses within selected pages were called. In Phase 2, only E-firms were interviewed because the required sample size of non-E-firms was achieved in Phase 1. Therefore, the design weights were calculated as the inverse of the overall selection probability of businesses as follows:

$$W_{hij}^{0} = \frac{M_h}{m_{hi}}$$

where M_h is the total number of pages from stratum *h* in the Kinz frame (strata are business sectors), m_{hi} is the number of pages selected in phase *i* from stratum *h*. The design weights were then adjusted for nonresponse among eligible phone numbers using the same approach used in Egypt, yielding the adjusted weight for nonresponse as:

$$W_{hij}^{1} = W_{hij}^{0} A_c$$

Because non-E-firms were not eligible for interviews in Phase 2, the weight was post-stratified to retrieve the actual percentage distribution of E-firms and non-E-firms in the population. We used the weighted distribution from Phase 1 for post-stratification. The post-stratified weight was calculated as follows:

$$W_{hij}^{2} = \begin{cases} W_{hij}^{1}\dfrac{p_1}{p_{1+2}} & j \in E - firms \\ W_{hij}^{1}\dfrac{(1-p_1)}{p_{1+2}} & j \in non - E - firms \end{cases}$$

where p_1 is the weighted proportion of E-firms from Phase 1, weighted by W_{hij}^{1}, and p_{1+2} is the weighted proportion of E-firms from phases 1 and 2, weighted by W_{hij}^{1}.

The survey weight was then calculated as a normalized version of W_{hij}^{2} as follows:

$$W_{hij}^{3} = \frac{W_{hij}^{2} n_{comp}}{\sum_{j=1}^{n_{comp}} W_{cj}^{2}}$$

where n_{comp} is the total number of businesses completed the survey.

3. Morocco:
The weight calculations started by calculating design weights that reflect the selection probabilities of selecting the businesses from the sampling frame. As described earlier, frame pages were selected within each region, and a sample of businesses was then selected from each selected page. The design weights accounted for the multi-stage sampling and for the subsampling of non-E-firms in Phase 2. The design weights were calculated as the inverse of the overall selection probability of businesses as follows:

$$W_{hij}^{0} = \frac{M_h}{m_{hi}} \frac{N_{hij}}{n_{hij}} \frac{1}{f_{2o}}$$

where M_h is the total number of pages from stratum h in the Télécontact frame (strata are the geographic regions), m_{hi} is the number of pages selected in phase i from stratum h, N_{hij} is the total number of businesses listed in page j, n_{hij} is the number of businesses selected from page j, and f_{2o} is the subsampling fraction of non-E-firms in Phase 2 where $f_{2o} = 1$ where $i = 1$ or if j is an E-firm. The design weights were then adjusted for nonresponse among eligible phone numbers using the same approach used in Egypt, but with adjustment done by regions, yielding the adjusted weight for nonresponse as:

$$W_{hij}^{1} = W_{hij}^{0} A_c$$

Similar to the situation in Jordan, the weight was post-stratified as follows:

$$W_{hij}^{2} = \begin{cases} W_{hij}^{1}\dfrac{p_1}{p_{1+2}} & j \in E - firms \\ W_{hij}^{1}\dfrac{(1-p_1)}{p_{1+2}} & j \in non - E - firms \end{cases}$$

where p_1 is the weighted proportion of E-firms from Phase 1, weighted by W_{hij}^{1}, and p_{1+2} is the weighted proportion of E-firms from phases 1 and 2, weighted by W_{hij}^{1}. After normalization the weight was trimmed to avoid extreme outliers; weights were capped at 3.5 times the weight median. The survey weight was then calculated as a normalized version of W_{hij}^{2} as follows:

$$W_{hij}^{3} = \frac{W_{hij}^{2} n_{comp}}{\sum_{j=1}^{n_{comp}} W_{cj}^{2}}$$

where n_{comp} is the total number of businesses completed the survey.