Weighting Report

MSE Survey Wave II (2018)

Rafael Pilliard Hellwig,

M&E Advisor, USAID Jordan Local Enterprise Support Project

## Introduction

This document provides information on the survey weights provided in the second wave of the USAID Jordan survey of micro- and small enterprises (2018). The survey consists of a median of 102 questions[[1]](#footnote-1) in an approximate stratified two-stage element sample.

## Study Population

The survey gathers representative information for all micro- and small-establishments (MSEs) operating for 3+ years in the governorates of Amman, Zarqa, Irbid, Karak, Tafilah, and Aqaba.[[2]](#footnote-2) Although the study is not intended to be national in scope, the target population of the six areas collectively capture 60% of the kingdom’s population and represents a diverse cross-section of the country.

### First Stage

As there is no readily-accessible frame of establishments covering the full population of formal and informal establishments in Jordan, USAID LENS created such a frame in 2014/15. The frame of establishments is itself a sample () from the wider population of enterprises, and fits into a large probability sampling design carried out in two phases. The enumeration, which was undertaken in 2014/2015, resulted in 977 geographic clusters (neighborhood blocks) being selected from the seven strata using a simple random sampling design without replacement (*srswor*).[[3]](#footnote-3) From these clusters, 97,347 households and commercial addresses were contacted through door-to-door screener interviews, of which 10,197 reported owning a business or income-generating project. The frame that resulted was used for subsequent sampling in each wave of the MSE survey.

The sampling probabilities for the primary sampling units (PSUs) are listed below. The table lists the stratum population size (), stratum sample sizes (), and first stage inclusion probabilities (). For example, 185 neighborhood blocks were sampled out of 2005 blocks in Irbid, for a first stage inclusion probability in Irbid of 9.2%.

| **stratum** |  |  |  |
| --- | --- | --- | --- |
| 1. IRB | 2005 | 185 | 0.09226933 |
| 1. ZAR | 1822 | 134 | 0.07354555 |
| 1. TAF | 231 | 136 | 0.58874459 |
| 1. AMM | 2545 | 237 | 0.09312377 |
| 1. AQB | 68 | 66 | 0.97058824 |
| 1. KAR | 583 | 176 | 0.30188679 |
| 1. NAR | 43 | 43 | 1.00000000 |

It is important to note that the first stage inclusion probabilities are not the same across different areas.

### Second Stage

For the 2018 wave of the survey, two samples were drawn. The core sample () consists of 3,629 elements sampled in a probabilistic design. A supplement was drawn of 1,875 additional cases using non-probabilistic methods that selected establishments closely matching a set of characteristics of USAID LENS beneficiaries. The supplement sample does not fit into the design-inference framework, and therefore is not assigned sampling weights.[[4]](#footnote-4) The subsequent paragraphs describes only the weighting procedure for the core sample.

For the second stage of the 2018 wave of sampling, all 977 out of 977 clusters from the frame were included for sub-sampling in the core sample. Selection was undertaken such that roughly 600 elements were drawn in each governorate, with at least 1 element sampled per cluster. Under these constraints, the sample size for a given cluster was set proportional to the size of the cluster () to obtain an approximately self-weighting design (by stratum).[[5]](#footnote-5)



The number of secondary sampling units (SSUs) are listed in the table below. For example, 602 establishments were sampled in Irbid for the core sample, with an additional 289 sampled for the (non-probabilistic) supplement.

| ***area*** | **Core**  ***nh*** | **Supplement** |
| --- | --- | --- |
| IRB | 602 | 289 |
| ZAR | 600 | 153 |
| TAF | 609 | 217 |
| AMM | 534 | 712 |
| AQB | 600 | 95 |
| KAR | 599 | 344 |
| NAR | 85 | 65 |
|  | 3629 | 1875 |

The second stage sampling probabilities and design variables are not listed here, but can be found in the dataset under the following headings:

## Weights

### Base Weight

The base weight is calculated using the design variables used in the sampling scheme, without any corrections for non-response. Had each in-sample business participated in the study, this weight would be design-unbiased. However, since the combined survey achieved a response rate (RR2) of 2503 / 5504 = 0.455, population estimates using the base weight will be shy of the true total. Other estimates such as means and proportions may also be biased due to the systematic patterns in the non-response.

The base weight is provided in the dataset as the basewt\_adj variable. Formally, it is defined as

where

is the probability that the th element is in the second-phase sample .

is the (realized) first-phase sample. This is the sampling frame of 10,197 enterprises.

is the probability of selecting the sample in which is realized. This is provided as the “pi1\_adj” variable in the dataset.

is the conditional probability of selecting the th element, given the first-phase sample. This is provided as the “pi2\_adj” variable in the dataset.

For simplicity, the added notation needed to account for the clustering and stratification in the MSE survey have been left out above. For the purposes of point estimation, the basic principles remain the same once the first phase and conditional second-phase inclusion probabilities are calculated for each.

After adjustment for non-response (see sections below), the expanded values can be used with the -estimator to estimate population totals from the sample ,

The weights can be used to estimate weighted means and proportions:

The -estimator of the population total is unbiased and measurable, in that an unbiased variance estimate can be calculated. The weighted mean is non-linear and approximately unbiased. Interested readers are encouraged to consult Särndal, Swensson, and Wretman pp. 343-379 for detailed information on the properties of the estimators. The lesser-known reweighted expansion estimate (REE) described by Kott and Stukel may also be appropriate and will produce smaller variances than .

### Survey Weight

The simplest of survey weight adjusts for non-response by treating missing observations as missing completely at random within each stratum (area). This assumption is a strong one—it is left to the analyst to choose whether he or she believes this to be an acceptable assumption. Under this approach, weights of respondents are rescaled such that they sum to the original size of the intended sample in each stratum. The area-level response rates[[6]](#footnote-6) are

| ***Stratum (area)*** | **RR4 for Core Sample ()** | **Estimated Eligibility ()** |
| --- | --- | --- |
| TAF | 0.670 | 0.791 |
| KAR | 0.650 | 0.852 |
| IRB | 0.576 | 0.748 |
| ZAR | 0.566 | 0.865 |
| AMM | 0.553 | 0.827 |
| NAR | 0.547 | 0.645 |
| AQB | 0.489 | 0.712 |
| All | 0.585 | 0.797 |

The survey weight (“svywt” variable) is calculated for the th element based on the observed probability of response in the element’s weighting class, :

The above implementation represents a small departure from the true MCAR assumption, as the approach are MAR given the weighting class rather than having a constant propensity of response across all governorates/sectors. A survey weight that truly assumes a MCAR process would be constructed as .

### Post-Stratification Weight (District Level)

Post-stratification generally improves the accuracy of survey estimates by adjusting weights to match known population totals. Because the MSE survey uses an area sample (with exact district population totals unknown), post-stratification is applied relative to estimates from stage I. This is sometimes referred to as a *weighting class* approach to non-response adjustment.

Weighting classes are constructed at the district level to adjust basewt\_adj for non-response. Weights are post-stratified to match the estimated population totals for each geographic district. The resulting pswt weight variable washes out some of the random variability in the same from the second stage and are generally preferable to the basewt\_adj for analysis that requires point estimation at the district level. However, it does not make strong adjustments for non-response. Furthermore, as the post-stratification is calibrated against population totals that themselves carry uncertainty, estimating variance of estimators relying on this weight remains a difficult problem.

| **District ID** |  |
| --- | --- |
| Df4c437c6 | 11719 |
| De6ecffb1 | 11201 |
| D370b6704 | 11172 |
| Deb30fe81 | 10954 |
| D8631d099 | 9225 |
| D3a5774d4 | 3453 |
| D4d64e670 | 2295 |
| Dc4ca3122 | 2276 |
| De47fd735 | 2102 |
| Dce483c2d | 1712 |
| Df9d7ae3d | 1696 |
| D382ad443 | 1686 |
| Ddf39fd18 | 1560 |
| D632cec06 | 1051 |
| Dffcf5456 | 715 |
| D78b5bbd8 | 629 |
| D7d40a4c0 | 586 |
| D8d42bfa1 | 571 |
| Db1da243f | 563 |
| D24f68bb1 | 490 |
| Dda6e41cc | 410 |
| Dcc426589 | 406 |
| D31fc4c60 | 379 |
| D66c37442 | 290 |
| Dda9abc03 | 238 |
| D6ad13412 | 224 |
| De3250c32 | 215 |
| D7fafb171 | 185 |
| Df80c7503 | 161 |
| Dd2cf4606 | 150 |
| D434a7559 | 83 |
| Da889b0d3 | 76 |

### Propensity Weights

The propensity score approach to non-response adjustment attempts to model and predict the non-response probability for each in-sample unit, given several auxiliary covariates.

The assumptions with this approach are that the response propensities can be known based on several external features, and that this information can be used to ‘adjust out’ the non-response bias. The propensity weight is obtained by dividing the base weight by the estimated response propensities for the *k*th element. These propensities are captured by the “propensity” variable in the data.

For the 2018 MSE survey, the response propensities are calculated using regularized boosted trees that rely on random forests as a base learner. The model is trained on the intended sample drawn from the stage I frame, which includes auxiliary variables such as sector, municipality, sex of the owner, type of building (commercial/residential), and the number of employees. The final model was chosen using 10-fold cross validation, using the AUC measure from the receiver operating characteristic (ROC) as a measure of fit. The final model achieves an estimated AUC of 0.597, indicating that the auxiliary variables are generally weak in their ability to predict response. The best model accurately predicts if a business will be a responder (specificity of 0.704) but has more difficulty identifying non-responders (sensitivity of 0.421).

The propensity weight is provided as the “propwt” variable in the dataset, and is defined for the th element as

Although the sampling mechanism is rooted in design-based theory, the propensity score approach follows an inverse probability of treatment weighting (IPTW) philosophy for non-response adjustment. This is the weight variable recommended for general use.

## Discussion of the Sampling Design

Due to the proportional mechanism in stage two, the design results in approximately self-weighting data within each stratum. This can be verified by producing histograms of weights by stratum:



For the purposes of estimating governorate-by-governorate means and proportions, it will generally suffice to use the sample mean.

Strictly speaking, the 2018 wave of sampling follows a two-phase approach rather than a two-stage design. The reason for this is that the second-stage probabilities are not invariant to the first stage sample, making the sample size depend on (itself a random quantity). Another way of conceptualizing this is that unlike in a true multistage design, the second stage sample size is not specified *a priori*. To complicate things further, certain PSUs that had few elements in the intended sample resulted in a non-interview (e.g. unreachable businesses). These idiosyncratic PSUs are missing from the set of responder data, and consequently make the theory of two-phase (rather than two-stage) sampling more apposite. Nevertheless, analyzing the data as though it were a true two-stage design is likely a reasonable approximation.

### Variance Estimation

Empirical testing by the research team reveals that the actual design effects on marginal estimates typically[[7]](#footnote-7) range between 2.1 and 3.0. The (weighted) median across marginal estimates is 2.6, implying a typical effective sample size of 660 units. If survey software for complex surveys is unavailable, a crude but simple approach is to multiply standard errors (confidence intervals, margins of error) from simple statistical methods by a factor of . As an upper bound, the design-based standard errors on marginal proportions/means will seldom exceed x the naïve estimate. As the sample was optimized for governorate-level estimation, design effects for statistics at the stratum level will be better, typically resulting in errors only 1.2x larger than the naïve stratum-level estimates.

Analysts wishing to obtain estimates of error (e.g. confidence intervals) may use the 1000 replicate weights provided along with the main data (“replicate\_weights.csv”). Software such as R, SAS, and Stata will handle such design specifications (though SPSS will not). Exemplar code is provided for specifying the such a design in R.

### Coverage Error

As a general rule, the 2018 survey data only captures “persisters.”[[8]](#footnote-8) It does not make adjustments for under-coverage due to so-called births of new establishments in the period between the frame creation and the interviews. Though some of the effects of market exit will have been ‘corrected’ via the propensity weights, some variables will still require careful interpretation. Examples include the year that businesses were established, owner’s age, whether or not businesses are registered, etc.

## References

The American Association for Public Opinion Research. 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 9th edition*. AAPOR.

Lumley, Thomas. *Complex Surveys: A Guide to Analysis Using R*. Hoboken, NJ: John Wiley & Sons, 2010.

Särndal, Carl-Erik, and Lundström, Sixten. *Estimation in Surveys with Nonresponse*. Chichester: Wiley, 2005.

Särndal, Carl-Erik, Swensson, Bengt, and Wretman, Jan. *Model Assisted Survey Sampling*. New York: Springer-Verlag, 2003.

Valliant, Richard, Jill A. Dever, and Frauke Kreuter. *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer, 2013.

1. As the survey is administered through software that uses conditional filtering, the number of questions varies per respondent. The minimum and maximum number of questions for any respondent were 86 and 142, respectively. [↑](#footnote-ref-1)
2. In the survey, the governorates of Amman and Aqaba exclude West Amman and the ASEZA free zone. As USAID LENS aims at working with lesser advantaged Jordanians, these wealthier areas do not comprise the population of interest for the study. [↑](#footnote-ref-2)
3. At the onset, six geographic were chosen for inclusion in the survey. Naour—which is a district within Amman governorate but is no longer a part of the Greater Amman Municipality— was added to the study after the initial fieldwork had begun. [↑](#footnote-ref-3)
4. The supplement cases have been included, however, in models for non-response weight adjustment applied to the core sample [↑](#footnote-ref-4)
5. The exact mechanism for determining inclusion probabilities under these constraints is determined using non-linear programming. The procedure is applied separately for each stratum (governorate) as follows:

   1. Define a lower bound of 1 on the cluster sample size such that all ≥ 1. Let denote the population size for the th cluster, let denote the stratum population size, and let denote the desired stratum sample size.
   2. Calculate the relative quantities Zhi = ()/ for each cluster in stratum . In the first iteration of the procedure, this represents the expected cluster size under exact proportionality (note that some values fall below 1).
   3. Define to be the set of clusters where . Set the expected sample size for clusters with less than the lower bound to 1. That is, let

   Where is a stratum scaling factor -  *+ ) /*  which ensures sums to the desired nh

   1. Repeat steps (2)–(3) until no cluster sample sizes fail the bounds check.

   [↑](#footnote-ref-5)
6. The response rate used in the adjustment correspond to the American Association of Public Opinion Research (AAPOR) “Response Rate 4.” This makes downward adjustments on the original frame to account for estimated ineligibles in each area, such that the sum total estimates the population of MSE ‘persistors’ who have existed for three or more years. [↑](#footnote-ref-6)
7. Here, the range is given as the area between the 25th and 75th weighted percentiles. Design effects are calculated on all marginal estimates of the survey and weighted so that each question receives equal weight, regardless of the number of response choices for a given item. [↑](#footnote-ref-7)
8. Carl-Erik Särndal and Sixten Lundström, Estimation in Surveys with Nonresponse, Chichester: Wiley, 2005. [↑](#footnote-ref-8)