Technical Note on L2Taj 2023 Baseline Survey Sampling Design and Weight Calculation

Data Inputs. This sampling design uses data from Census 2020. It provides information on census enumeration areas including their unique IDs, number of households, and population.

Methodology. The sampling design follows a stratified 2-stage proportional-to-size (PPS) clustered sampling approach. This approach combines stratification, PPS sampling, and cluster sampling in the following steps:

1. The population is stratified into 9 strata. These are the city of Dushanbe, the rural and urban regions of the 4 oblasts (Sughd; Khatlon; DRS; GBAO). Within each stratum, the first sampling stage follows a PPS approach, where each cluster (census enumeration area) is used as the primary sampling unit and sampled with probability proportional to cluster size in terms of population per cluster. This step results in a total of 200 clusters sampled.
2. In the second stage, households will be randomly sampled from each chosen cluster following a "random walk" procedure. In urban clusters, 30 households must be interviewed and in rural clusters 18 households must be interviewed to achieve strata-level representativeness. An exception is GBAO region where the urban cluster is allocated to have 20 households, rural clusters 16 households. This will result in a total of 4000 households at the national level. Table 1 below shows the final distribution of clusters and households.

Table 1. Distribution of clusters and households across strata for L2Taj Baseline Survey

	Total number of PSUs (clusters)	Cluster size	# of HHs
National	200		4000
Dushanbe	13	30	390
Sughd – Urban	9	30	270
Sughd – Rural	48	18	864
Khatlon – Urban	8	30	240
Khatlon – Rural	66	18	1188
RRP – Urban	4	30	120
RRP – Rural	46	18	828
GBAO – Urban	1	20	20
GBAO – Rural	5	16	80

Weights. The approach results in a self-weighted sample. This is because both the stratification and the 2-stage cluster sampling are proportional to its size. The probability of a cluster being selected within each stratum (pr_1) in the first stage of sampling and the probability of a household being selected within cluster (pr_2) in the second stage are calculated as follows:

$$pr_1 = \frac{\#\ of\ selected\ clusters \times\ population\ per\ cluster}{population\ per\ stratum}$$

$$pr_2 = \frac{1}{total\ \#\ of\ households\ per\ cluster}$$

Based on pr_1 and pr_2, an initial sampling weight ($origsw$) is calculated as:

$$origsw = \frac{1}{pr_1 \times pr_2}$$

It is further rescaled by the strata population number. This weight can simply be interpreted as "*the population number that each cluster represents within each stratum*".

Once L2Taj Baseline data collection is finalized, population weight for each household ($popw$) is calculated by dividing the sampling weight ($origsw$) by the actual number of households interviewed per cluster.[1] This weight will show "*the population number that each household represents within each stratum*". Finally, individual weights ($indw$) are calculated by dividing the population weight ($popw$) by household size, which would show "*the population number that each individual in the survey represents within each stratum*".

[1] Ideally, the number of households interviewed per cluster should correspond to the numbers reported in Table 1. However, if there are any slight deviations (e.g. if more households are interviewed per cluster than planned), then using the actual number of households interviewed per cluster generates correct population weights, i.e. by redistributing weight if more households are interviewed per cluster, or by increasing weight if fewer households are interviewed per cluster.