ADDENDUM:

Methology Note: COVID-19 High-Frequency Survey (HFS) in Latin American Countries Sampling Design and Weighing *

The COVID-19 High-Frequency Survey was conducted by phone in thirteen Latin American countries: Argentina, Bolivia, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Paraguay and Peru. Data collection spanned over three waves between May and August 2020, and collection periods lasted about ten days per wave on average. The survey interviewed one adult per household and asked both individual and household-level questions.

All national samples are based on a dual frame of cell phones and landlines and have a probability one-stage design with geographic stratification. The samples were generated through a Random Digit Dialing (RDD) process, thus ensuring coverage of all landline and cell phone numbers active at the time of the survey.

Survey estimates for each country represent households with a landline or in which at least one member has a cell phone, and individuals 18 years of age or above who have an active cell phone number or a landline at home.

Sampling design

Each country's sample is based on a dual frame of cell phone and landline numbers generated through an RDD process. The RDD methodology produces all *possible* phone numbers in the country under the national phone numbering plan and draws a random sample of numbers. This way, the process guarantees full coverage of phone numbers and eliminates any potential coverage bias with respect to the population with a phone[1].

First, in each country a large first-phase sample was selected in each of both number frames, with an allocation ranging from 0% landlines and 100% cell phones to 20% landlines and 80% cell phones (which can be distinguished based on their prefixes). Landlines were included with a small share in order to cover the landline-only households and persons, which have a low prevalence in most Latin American countries but yet exist, particularly among the senior population.

In all countries, the landline frame was geographically stratified by department, province or state, and the sample of landlines was selected with proportionate allocation among strata. Geographic proportionate stratification was also done for cell phones in Argentina, Bolivia and México[2]. It is underscored that the HFS sample design permits to obtain precise estimates at the country level only. Subnational estimates would have large sampling errors.

---

* This technical note was prepared by Ramiro Flores Cruz, partner member at Sistemas Integrales and World Bank consultant.

[1] In other words, given that the HFS used a sampling frame of phone numbers, its results represent only the population with a phone and exclude the population with no phone.

[2] Geographic stratification of cell phone numbers was feasible only in these three countries because only in them it is possible to link a cell phone number to the district (department, province or state) where it was issued.

The first-phase samples of landline and cell phone numbers were then screened through an automated process to identify the active numbers, and these were then cross-checked with business registries (based on yellow pages and websites) to identify business numbers, not eligible for this survey.

Second, a smaller second-phase sample[3] was selected from the *active residential* numbers identified in the first-phase sample and was delivered to the country team to be called by the interviewers. The reason for selecting a second-phase sample was that delivering a large first-phase sample of active numbers to the country teams at once could facilitate the "misuse" of the sample, raise nonresponse rates, and increase potential nonresponse biases.

Table 1 shows the final sample size per country and the allocation between both frames.[4]

Table 1. Sample size and allocation to cell phones and landlines in HFS Round 1

Country	Sample size	Cell phones	Landlines
Argentina	1000	85%	15%
Bolivia	1,000	100%	0%
Chile	1000	80%	20%
Colombia	1,000	85%	15%
Costa Rica	800	90%	10%
Domincan Rep	800	85%	15%
Ecuador	1,200	85%	15%
El Salvador	800	90%	10%
Guatemala	800	90%	10%
Honduras	800	100%	0%
Mexico	2,000	80%	20%
Paraguay	800	100%	0%
Peru	1,000	90%	10%

When an interview was obtained through a cell phone, the interviewer interviewed the person who answered the call (as long as he or she was 18 years of age or above) and asked all questions about the respondent and his or her household. When an interview was obtained through a landline, the interviewer requested to talk to any household member 18 years of age or older and asked all questions about the respondent and his or her household. Landlines are 10% - 15% of the sample in most countries, 20% in two of them, and 0% in three of them.

Respondents were recontacted in two additional rounds, thus producing panel data.

[3] Note that the selection of phone numbers involves two sampling *phases*, and not two sampling *stages*.

[4] The HFS samples have one-stage, so the design effects of all variables equal 1, or are even smaller than 1 due to stratification. As a result, the effective sample sizes are equal or larger than the nominal sizes, reducing the standard errors. This feature contrasts with multi-stage clustered samples, which typically have design effects significantly larger than 1 and, therefore, their effective sizes are smaller than their nominal size, increasing the standard errors.

Weighting

The HFS has two sample units: households and individuals. Sampling weights were computed for each unit and should be used according to the estimate of interest. The weighting process involves four steps:

1. Calculation of the inclusion probabilities of landline and cell phone numbers.
2. Computation of base weights for households and individuals.
3. Nonresponse weighting adjustment.
4. Calibration of individual and household weights, using external data from official sources (adjusted for the national phone coverage).

In the second and third HFS waves, household and individual weights were adjusted for attrition.

Step 1: Inclusion probabilities of landline and cell phone numbers

A first-phase sample was selected in each of the two frames (cell phone numbers and landline numbers) with simple random selection without replacement, and the automated screening classified the selected numbers into active and inactive. The first-phase inclusion probabilities of cell phones and landlines are[5]

$$\pi^{C}_{(1)i} = \frac{n^{C}_{(1)}}{N^{C}_{(1)}} = \frac{n^{C}_{(1)A} + n^{C}_{(1)IN}}{N^{C}_{(1)}}$$

$$\pi^{L}_{(1)hi} = \frac{n^{L}_{(1)h}}{N^{L}_{(1)h}} = \frac{n^{L}_{(1)hA} + n^{L}_{(1)hIN}}{N^{L}_{(1)h}}$$

where

$\pi^{C}_{(1)i}$ is the first-phase inclusion probability of the i-th cell phone number

$n^{C}_{(1)}$ is the size of the first-phase sample of cell phones, composed of $n^{C}_{(1)A}$ active cell phones and $n^{C}_{(1)IN}$ inactive cell phones

$N^{C}_{(1)}$ is the cell phone frame size (all possible cell phone numbers according to the national numbering plan)

$\pi^{L}_{(1)hi}$ is the first-phase inclusion probability of the i-th landline number in stratum h

$n^{L}_{(1)h}$ is the size of the first-phase sample of landlines in stratum h, composed of $n^{L}_{(1)hA}$ active landlines and $n^{l}_{(1)hIN}$ inactive landlines

$N^{L}_{(1)h}$ is the landline frame size in stratum h (all possible landline numbers according to the national numbering plan)

[5] Inclusion probabilities of cell phone numbers do not show a stratum index since most cell phone samples were not stratified because of the reasons stated above. Only the cell phone samples for Argentina, Bolivia and Mexico were stratified.

Next, two second-phase samples were selected independently out of the first-phase samples of active cell phones and landlines. The second-phase inclusion probabilities of cell phones and landlines are

$$\pi^{C}_{(2)i|(1)i} = \frac{n^{C}_{(2)A}}{n^{C}_{(1)A}}$$

$$\pi^{L}_{(2)hi|(1)hi} = \frac{n^{L}_{(2)hA}}{n^{L}_{(1)hA}}$$

where

$\pi^{C}_{(2)i|(1)i}$ is the second-phase inclusion probability of the i-th active cell phone number conditional on being selected in the first phase

$n^{C}_{(2)A}$ is the size of the second-phase sample of active cell phones

$\pi^{L}_{(2)hi|(1)hi}$ is the second-phase inclusion probability of the i-th active landline number in stratum h conditional on being selected in the first phase

$n^{L}_{(2)hA}$ is the size of the second-phase sample of active landlines in stratum h

Then, the unconditional inclusion probabilities of the second-phase active cell phones and landlines are

$$\pi^{C}_{i} = \pi^{C}_{(1)i}\,\pi^{C}_{(2)i|(1)i} = \frac{n^{C}_{(1)A} + n^{C}_{(1)IN}}{N^{C}_{(1)}}\,\frac{n^{C}_{(2)A}}{n^{C}_{(1)A}} = \frac{n^{C}_{(1)A} + n^{C}_{(1)IN}}{n^{C}_{(1)A}}\,\frac{n^{C}_{(2)A}}{N^{C}_{(1)}} = \frac{n^{C}_{(2)A}}{\widehat{RA}^{C}_{(1)}N^{C}_{(1)}} = \frac{n^{C}_{(2)A}}{\hat{A}^{C}_{(1)}}$$

$$\pi^{L}_{hi} = \pi^{L}_{(1)hi}\,\pi^{L}_{(2)hi|(1)hi} = \frac{n^{L}_{(1)hA} + n^{L}_{(1)hIN}}{N^{L}_{(1)h}}\,\frac{n^{L}_{(2)hA}}{n^{L}_{(1)hA}} = \frac{n^{L}_{(1)hA} + n^{L}_{(1)hIN}}{n^{L}_{(1)hA}}\,\frac{n^{L}_{(2)hA}}{N^{L}_{(1)h}} =$$

$$= \frac{n^{L}_{(2)hA}}{\widehat{RA}^{L}_{(1)h}\,N^{L}_{(1)h}} = \frac{n^{L}_{(2)hA}}{\hat{A}^{L}_{(1)h}}$$

Where $\widehat{RA}_{(1)}$ is the rate of active phones estimated in the first phase[6]. Hence, the unconditional inclusion probabilities of the second-phase active numbers π^{C}_{i} and π^{L}_{hi} can be expressed as the ratio between the second-phase selected active numbers and an estimate of the total active numbers in the frame $\hat{A}_{(1)}$.

[6] $\widehat{RA}_{(1)}$ estimates are highly precise due to the very large size of the first-phase samples.

Step 2: Base weights for households and individuals

The selection probabilities of households and individuals 18+ are based on the inclusion probabilities of the cell phone and landline numbers through which they are reached. Therefore, the computation of household and individual weights should account for multiplicity and for the specific overlapping pattern between the frames of cell phones and landlines. Otherwise, household and individual-level estimates will be biased.

Multiplicity adjustment

There is multiplicity when a household has a larger selection probability because it can be selected through different sample elements. Thus, if a household has more than one cell phone or more than one landline, the household selection probability needs to be adjusted to account for the increased chance of selection. The multiplicity-adjusted household selection probabilities in each frame are computed as

$$\pi^{C}_{mj} = m_{cj}\pi^{C}_{i}$$

$$\pi^{L}_{mhj} = m_{lj}\pi^{L}_{hi}$$

where

π^{C}_{mj} is selection probability of the j-th household when contacted through a cell phone, adjusted for multiplicity of working cell phones in the household

m_{cj} is the number of working cell phones in the j-th household

π^{L}_{mhj} is the selection probability of the j-th household in stratum h when contacted through a landline, adjusted for multiplicity of working landlines in the household

m_{lj} is the number of working landlines in the j-th household

Therefore, if a household has m_c cell phones, its chance of being selected through a cell phone is m_c higher than a household where there is only one cell phone. The same applies to landlines, in which case the multiplicity factor is m_l. Since the number of cell phones and landlines in a household is unknown at the time of the sample design, it needs to be asked during the interview as part of the questionnaire.

The probability of an individual being selected through a cell phone equals the inclusion probability of his or her cell phone number. On the other hand, the probability of an individual being selected through a landline equals the selection probability of his or her household, conditional on the number of working landlines in the household, over the number of adult individuals in the household:

$$\pi_k^C = \pi_i^C$$

$$\pi_{hjk}^L = \pi_{mhj}^L / \sum_j k$$

where

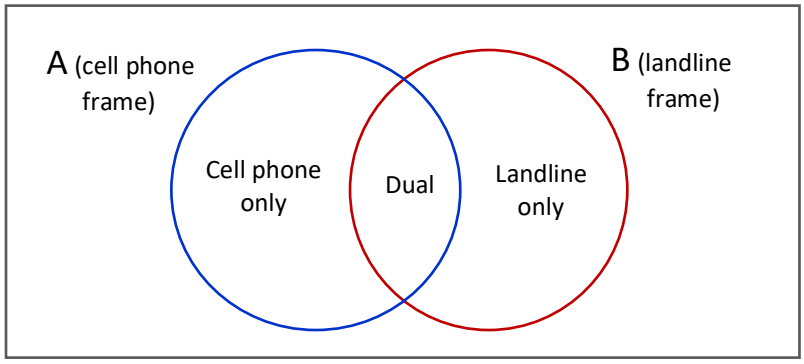
π_k^C is the selection probability of the k-th individual when contacted through a cell phone

π_{hjk}^L is the selection probability of the k-th individual in stratum h when contacted through a landline in the j-th household

Overlapping sampling frames

Households and individuals that have both cell phone and landline (dual cases) have a higher probability of being selected than those which have only cell phones or only landlines. The following diagram displays the overlapping pattern of cell phone and landline sampling frames.

Figure 2. Partially overlapping frames



In order to adjust the selection probabilities for overlapping frames and multiplicity, it is essential to collect some related information during the interview. It is necessary to know the domain ownership of the sample households and individuals, as well as the number of cell phones and landlines in the sample households. To this purpose, the HFS questionnaire included the following three questions:

1. *How many working cell phones in total are owned by the persons in your household, including you?*
2. *Is there any working landline in your household?*
3. *How many working landlines are there in your household currently?*

By knowing the domain of ownership, the selection probability for each sample unit can be calculated based on the following probability property

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

where $P(A \cap B) = P(A) \times P(B)$, given that A and B are independent

- *In general*, in a dual-frame telephone sample design

$$\pi = \begin{cases} \pi^C & \text{if the sample unit is cell phone only} \\ \pi^L & \text{if the sample unit is landline only} \\ \pi^C + \pi^L - \pi^C \pi^L & \text{if the sample unit is dual} \end{cases}$$

where π^C y π^L are the selection probabilities of the sample units (households or individuals) in each domain (cell phone only, landline only, dual).

- *In the specific HFS setting* (with overlapping frames and multiplicity)

Selection probabilities for households are

$$\pi_j = \begin{cases} m_{cj}\pi_i^C & \text{if the household is cell phone only} \\ m_{lj}\pi_{hi}^L & \text{if the household is landline only} \\ m_{cj}\pi_i^C + m_{lj}\pi_{hi}^L - m_{cj}\pi_i^C \; m_{lj}\pi_{hi}^L & \text{if the household is dual} \end{cases}$$

And selection probabilities for individuals are

$$\pi_k = \begin{cases} \pi_i^C & \text{if the individual is cell phone only} \\ m_{lj}\pi_{hi}^L / \sum_j k & \text{if the individual is landline only} \\ \pi_i^C + m_{lj}\pi_{hi}^L / \sum_j k - \pi_i^C \; m_{lj}\pi_{hi}^L / \sum_j k & \text{if the individual is dual} \end{cases}$$

Household and individual base weights, w_{0j} and w_{0k} respectively, are the inverse of the above selection probabilities

$$w_{0j} = \pi_j^{-1}$$

$$w_{0k} = \pi_k^{-1}$$

Step 3: Nonresponse adjustment

When a phone number is called, it is not always possible to carry out an interview. Nonresponse occurs either because nobody answers the call (no contact), because the respondent is unwilling to cooperate (refusal), or because of other barriers such as language.

The HFS put in place four main strategies to minimize nonresponse:

a. The survey central management team sent SMSs to the sample cell phone numbers before calling to inform the user who was about to call and to persuade him or her to answer the call.

b. In some countries, the sample was released to the country teams over successive replicates in order to keep nonresponse properly monitored and under the control of the central management team.

c. Stringent calling protocols were put in place and monitored to ensure a minimum number of attempts on different days and times (5 to 10 attempts depending on the country).

d. The survey offered monetary and nonmonetary incentives to those who cooperated (gift cards and phone credit).

e. In some countries, the most experienced interviewers recontacted the numbers classified as a "Refusal" to convert them into a "Complete interview".

These actions allowed reaching response rates higher than similar studies based on RDD samples. Final nonresponse rates varied across the HFS countries, with the lowest levels in Bolivia and Ecuador, and the highest in Argentina and Mexico.

The base weights of the responding households and individuals were adjusted to compensate for nonresponse and thus reduce the potential bias it may cause on the survey estimates. For this purpose, a class-based adjustment was used. This approach consists of forming classes by crossing all categories of auxiliary variables that are both known to be correlated with the likelihood of responding and are available for respondents and nonrespondents. Given that the survey used an RDD sample, the information in the sampling frame was limited and the only variables known for both respondents and nonrespondents were the type of phone number (landline or cell phone) and the corresponding geographic region (known for landlines in all countries, and for cell phones only in Argentina, Bolivia and Mexico).

One type of weighting class nonresponse adjustment is based on the inverse of the weighted response rate estimate in each class. This is the ratio of the sum of the base weights for all units (respondents and nonrespondents) in class c to the sum of the base weights for the respondents in that class.

$$a_{jc} = \frac{\sum_{j \in c,R} w_{0j} + \sum_{j \in c,NR} w_{0j}}{\sum_{j \in c,R} w_{0j}} \quad ; \quad a_{kc} = \frac{\sum_{k \in c,R} w_{0k} + \sum_{k \in c,NR} w_{0k}}{\sum_{k \in c,R} w_{0k}}$$

where a_{jc} is the nonresponse adjustment factor that should be applied to responding households in class c, and a_{kc} is the nonresponse adjustment factor for responding individuals in that class. R and NR indicate the responding and nonresponding units, respectively.

Thus, the nonresponse adjusted weights for responding households and individuals are

$$w'_j = w_{0j}\ a_{jc}$$

$$w'_k = w_{0k}\ a_{kc}$$

Step 4: Calibration if individual and household weights

As the last step, the weights for the responding households and individuals were calibrated using external data from official sources. This last adjustment has two objectives:

- Use auxiliary variables from external sources to further reduce potential nonresponse biases that were not addressed by the auxiliary variables used in Step 3. This can be achieved as long as the calibration auxiliaries are correlated with nonresponse.
- Improve the precision of estimators (i.e. reduce the sampling variances), as long as the auxiliaries are correlated with the analysis variables of interest.[7]

The goal of calibration is to find a set of weights *that are close to the input weights* (nonresponse adjusted weights in this case), and when used to estimate totals of the auxiliaries, reproduce the population totals exactly. Put formally, calibration minimizes a measure of the distance[8] between the input weights and the calibrated weights, under the constraint that the sum of the calibrated weights equals the sum of the totals of the auxiliaries from an external source. Unlike the nonresponse adjustment carried out in the previous step, calibration requires that the auxiliary variables be available for respondents only, and not for both respondents and nonrespondents.

Among the several existing calibration techniques, the HFS used the raking method. This method was most suitable given that the available auxiliary variables (region, sex and age groups) were all categorical, that region had many categories in most countries, and that the HFS samples are rather small.

The final weights for responding households and individuals can then be expressed as

$$w_j = w_j' \, g_j = w_{0j} \, a_{jc} \, g_j$$

$$w_k = w_k' \, g_k = w_{0k} \, a_{kc} \, g_k$$

where

w_{0j} is the base weight for the j-th household

a_{jc} is the nonresponse adjustment factor for households in class c

g_j is the calibration factor for the j-th household

w_{0k} is the base weight for the k-th individual

a_{kc} is the nonresponse adjustment factor for individuals in class c

g_k is the calibration factor for the k-th individual

[7] This objective was not addressed in this survey since it would have entailed computing a large set of replicate weights (with bootstrap or jackknife replication methods), which could be confusing for the final user and lead to error when estimating.

[8] The HFS weight calibration applies the raking calibration method, using the logit distance function.

Table 2 shows the data sources used for calibrating the weights in each country. All the population counts taken from these sources were adjusted for telephone coverage, using the national phone coverage rates published by the International Telecommunication Union (ITU) from the United Nations.

Table 2. Data sources for the auxiliary data used for weight calibration

Country	Data source used for weight calibration
Argentina	Instituto Nacional de Estadística y Censos. Proyecciones Elaboradas en base al Censo Nacional de Población, Hogares y Viviendas 2010.
Bolivia	Instituto Nacional de Estadística. Proyecciones de Población. 2020.
Chile	Instituto Nacional de Estadística. Estimaciones y Proyecciones de la Población de Chile 1992-2050.
Colombia	Departamento Administrativo Nacional de Estadística. Proyecciones de Población Nacional para el Periodo 2018-2070.
Costa Rica	Centro Centroamericano de Población. Proyecciones Distritales de Población de Costa Rica 2000-2050.
Domincan Rep.	Oficina Nacional de Estadistica. Población Estimada y Proyectada para el Período 1950 – 2100.
Ecuador	World Bank. Ecuador Sociodemographic and Labor Force Survey for Oopulation in Human Mobility - EPEC (2019).
El Salvador	Centro Centroamericano de Población. Proyecciones de Población de El Salvador. 2000-2050.
Guatemala	Instituto Nacional de Estadística. Proyecciones Nacionales 1950-2050.
Honduras	Instituto Nacional de Estadística. Proyecciones de Población 2013-2015.
Mexico	Consejo Nacional de Población. Proyecciones de la Población de México y de las Entidades Federativas, 2016-2050.
Paraguay	Dirección General de Estadística, Encuestas y Censos. Proyección de la población nacional por sexo y edad, 2000-2025. Revisión 2015.
Peru	Instituto Nacional de Estadística e Informática. Estimaciones y Proyecciones de Población. Boletín Especial Nº 21 y 22.

Reference literature

Lohr, S., RAO, J., (2006). Estimation in Multiple-Frame Surveys, Journal of the American Statistical Association, 101, 1019−1030.

Lohr, S., (2011). Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames, Survey Methodology, 37, 197−213. Statistics Canada.

Skinner, C., Rao, J., (1996). Estimation in Dual-Frame Surveys with Complex Designs, Journal of the American Statistical Association, 91, 349−356.

Thompson, S. (2012). Chapter 15: Network Sampling and Link-Tracing Designs, in Sampling. New York, Wiley.

Valliant, R., Dever J., and Kreuter F., (2016). Practical Tools for Designing and Weighting Sample Surveys. New York, Springer.