COVID-19 High-Frequency Phone Survey (HFPS) in Brazil

Technical Note on Sampling Design, Weighting, and Estimation[1]

December 2021

1. Background of the Brazil HFPS

Starting in March 2020 the World Bank conducted a series of COVID-19 High-Frequency Phone Surveys (HFPS) to assess the impact of the coronavirus pandemic on the welfare of Latin American and Caribbean households. The HFPS was conducted initially in 13 Latin American countries: Argentina, Bolivia, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Paraguay and Peru. These surveys gathered information on food insecurity, changes in employment, income loss, access to health services and education, and coping mechanisms, as well as households' quarantine compliance and their knowledge of the disease. Eligible respondents for the HFPS were adults 18 years old and above. Only one respondent per household was interviewed, and he/she answered both individual and household-level questions. All national samples were based on a dual frame of cell and landline phones, which were selected as a one-stage probability sample, with geographic stratification of landline numbers. The samples were generated through a Random Digit Dialing (RDD) process covering all cell and landline telephone numbers active in each country at the time of the survey. The overall sample design and weighting procedures for this survey are described in the report "COVID-19 High-Frequency Phone Survey (HFPS) in Latin America: Technical Note on Sampling Design, Weighting and Estimation"[2] on survey methodology and sampling. The HFPS was generally conducted as a panel phone survey, with the same households and respondents interviewed at regular periods of time to allow for a longitudinal analysis.

In mid-2021 the World Bank launched a second Phase of the HFPS project, in collaboration with the United Nations Development Programme (UNDP). The first wave of the second phased included data collection in Brazil. The purpose of this note is to describe the sample design and weighting procedures used for the Brazil HFPS. This survey was designed to follow the same general sampling approach and weighting procedures as the regional HFPS Phase I. In the case of the weighting procedures, the same terminology from that Technical Note will be used for consistency, with a few new terms introduced based on the specific features of the Brazil HFPS design.

---

[1] Note prepared by David Megill, Sampling Consultant, World Bank. Ramiro Flores Cruz provided key inputs for Section 8. The author appreciates the comments from Gabriel Lara Ibarra, Anna Luisa Paffhausen and Ricardo Campante Cardoso Vale of the World Bank.

[2] Florez Cruz, Ramiro. 2021. High-Frequency Phone Survey (HFPS) In Latin America: HFPS Phase 1 Technical Note on Sampling, Weighting and Estimation.
https://openknowledge.worldbank.org/bitstream/handle/10986/36395/COVID-19-High-Frequency-Phone-Surveys-in-Latin-America-Technical-Note-on-Sampling-Design-Weighting-and-Estimation.pdf?sequence=1

2. Sampling Frame and Stratification for Brazil HFPS

Brazil is divided into 27 states. Both the landline and cell phone numbers in each state of Brazil have unique 2-digit area codes, which are ideal for stratifying each sampling frame by state. In most other countries where the HFPS was conducted only the landline phone numbers had unique area codes by state or province. This makes it possible to have a highly stratified sampling frame by type of phone and state for the Brazil HFPS. The geographic stratification of the sampling frame improves the statistical efficiency of the sample design based on the greater homogeneity of the households and population within each state, and the corresponding reduction in the sampling errors. Table 1 specifies the area codes by state for both the landline and cell phones of Brazil.

Table 1. Phone area codes by state

State	Area codes
RONDÔNIA	69
ACRE	68
AMAZONAS	92, 97
RORAIMA	95
PARÁ	91, 93, 94
AMAPÁ	96
TOCANTINS	63
MARANHÃO	98, 99
PIAUÍ	86, 89
CEARÁ	85, 88
RIO GRANDE DO NORTE	84
PARAÍBA	83
PERNAMBUCO	81, 87
ALAGOAS	82
SERGIPE	79
BAHIA	71, 73, 74, 75, 77
MINAS GERAIS	31, 32, 33, 34, 35, 37, 38
ESPÍRITO SANTO	27, 28
RIO DE JANEIRO	21, 22, 24
SÃO PAULO	11, 12, 13, 14, 15, 16, 17, 18, 19
PARANÁ	41, 42, 43, 44, 45, 46
SANTA CATARINA	47, 48, 49
RIO GRANDE DO SUL	51, 53, 54, 55
MATO GROSSO DO SUL	67
MATO GROSSO	65, 66
GOIÁS	62, 64
DISTRITO FEDERAL	61

Source: own compilation.

It can be seen in Table 1 that some larger states have more than one area code. In order to provide further implicit stratification within each state, the sampling frames of landline and cell phone numbers were sorted by area code within each state, and random systematic sampling was used for selecting the sample active phone numbers. This also provided a proportional allocation of the sample phone numbers by area code within each state.

In the case of random digit dialing (RDD), the first phase sampling frames for the landline and cell phones in each stratum (state) consist of all possible combinations of random digits within each area code and numbering system designated for that state. For the first phase sampling of landline and cell phone numbers, a Dutch firm was contracted to generate a large number of random digit combinations for each state, and test each number to determine whether it is working (active) or not working (not active). As expected, given the extremely large possible random digit combinations for phone numbers in the first phase, only a small portion of the random phone numbers are "active" in the first phase, especially for landline phone numbers. The Dutch firm continued screening all the random digit combinations for the landline and cell phone numbers for each state until a target number of "active" phone numbers was reached for the first phase sample for each state. This large database of "active" landline and cell phone numbers for each state is the frame for the second phase of sampling. A strict accounting is made of the size of the theoretical frame of random digit combinations, and the first phase sample of "active" and "not active" landline and cell phone numbers that are generated. These parameters are used later for the calculation of the weights, as described in the corresponding section of this report.

Prior to the first phase generation of the large database of RDD numbers, it was necessary to establish the target number of "active" landline and cell phone numbers that would be needed for each state. The first phase target sample of landline and cell phone numbers had to be a large multiple of the sample size of phone interviews allocated to each state. Because of the potential low response rate for phone surveys, it was necessary to select multiple batches of landline and cell phone numbers to complete each sample interview, as explained in the next section.

3. Sample size and allocation by state for Brazil HFPS

Based on the experience of the HFPS in other countries of the region and the expected level of precision for the national-level indicators, the initial target sample size for the Brazil HFPS was set at 3,000 phone interviews, with 15% (450) allocated to landline phones and 85% (2,550) to cell phones. For a national-level survey, it is statistically effective to allocate the sample to each stratum in proportion to the population, which is highly correlated with the total number of phones. For this purpose we used the Brazil population projections by state for 2021 from the *Instituto Brasileiro de Geografia e Estatística* (IBGE), shown in Table 2. This table also shows the percent population distribution by state and the approximate proportional allocation of 450 landline phone interviews and 2,550 mobile phone interviews.

Table 2. Allocation of total sample of active landline and mobile phone numbers by state for Brazil HFPS

State	2021 projected population	% population	Proportional allocation landline interviews*	Total active landline phone sample**	Proportional allocation cellphone interviews*	Total sample of active mobile phones**
RONDÔNIA	1,815,278	0.9%	4	200	22	660
ACRE	906,876	0.4%	2	100	12	360
AMAZONAS	4,269,995	2.0%	9	450	52	1,560
RORAIMA	652,713	0.3%	3	150	9	270
PARÁ	8,777,124	4.1%	19	950	108	3,240
AMAPÁ	877,613	0.4%	3	150	11	330
TOCANTINS	1,607,363	0.8%	3	150	20	600
MARANHÃO	7,153,262	3.4%	15	750	86	2,580
PIAUÍ	3,288,504	1.5%	7	350	39	1,170
CEARÁ	9,241,366	4.3%	19	950	110	3,300
RIO GRANDE DO NORTE	3,560,903	1.7%	7	350	42	1,260
PARAÍBA	4,059,905	1.9%	8	400	48	1,440
PERNAMBUCO	9,675,249	4.5%	20	1,000	114	3,420
ALAGOAS	3,364,895	1.6%	7	350	39	1,170
SERGIPE	2,338,688	1.1%	5	250	28	840
BAHIA	14,985,070	7.0%	31	1,550	178	5,340
MINAS GERAIS	21,411,923	10.0%	45	2,250	258	7,740
ESPÍRITO SANTO	4,108,508	1.9%	9	450	49	1,470
RIO DE JANEIRO	17,463,349	8.2%	36	1,800	204	6,120
SÃO PAULO	46,649,132	21.9%	98	4,900	556	16,680
PARANÁ	11,597,484	5.4%	25	1,250	139	4,170
SANTA CATARINA	7,338,473	3.4%	16	800	89	2,670
RIO GRANDE DO SUL	11,466,630	5.4%	24	1,200	136	4,080
MATO GROSSO DO SUL	2,839,188	1.3%	6	300	34	1,020
MATO GROSSO	3,567,234	1.7%	8	400	44	1,320
GOIÁS	7,209,247	3.4%	15	750	87	2,610
DISTRITO FEDERAL	3,091,667	1.4%	6	300	36	1,080
BRAZIL	**213,317,639**		450	22,500	2,550	76,500

Source: own elaboration using IBGE projected population estimates. Notes: * Allocations are approximate. ** Includes reserves

In order to ensure a sufficient second phase sample of "active" phone numbers for each state to accommodate a relatively low response rate, the specifications for the number of "active" phone numbers by state in the ToRs for the Dutch firm multiplied the target number of interviews by 50 for the landline phones and by 30 for the cell phones. Table 2 also shows the specified target number of "active" landline and cell phone numbers by state for the second phase sample.

When the Dutch firm delivered the first phase list of "active" phone numbers for each state, they increased the total landline and cell phone numbers for each state specified in Table 2 by 10 percent. Following the systematic selection and allocation of the sample active phone numbers for the different batches, the availability of the additional 10 percent of phone numbers in the first phase sample facilitated the selection of a separate sample for the Pilot Survey.

4. Selection of batches of sample landline and cell phone numbers for CATI operations

In order to increase the response rate for phone interviews as much as possible and reduce the nonresponse bias, strict protocols were established for data collection. First, enumerators would attempt at least 7 call-attempts for each phone number in the case of unanswered calls. Initially this included calls on the weekend, but later this protocol was relaxed when the low rate of completed calls was slowing down the data collection, and the number of required call-attempts was also reduced to 5[3]. Second, it was decided to only provide the survey firm (OPPEN) with reserve phone numbers for each interview in relatively small batches. For the landline phones each batch included 10 phone numbers per target phone interview, and for the cell phones each batch contained 5 phone numbers per target interview. Once the survey firm exhausted one batch of phone numbers, they requested another batch. A total of 5 batches were selected for the target 450 landline interviews, and 6 batches were selected for the target 2,550 cell phone interviews. This resulted in a total of 50 phone numbers for each landline phone interview, and a total of 30 phone numbers for each cell phone interview. That is, for each target landline phone interview, a total of 50 phone numbers were available to complete said interview. If the interview was completed before the 50 numbers were exhausted, the remaining numbers were not used for the rest of the project.

The sample of "active" landline and cell phone numbers by state for each batch were selected from the first phase sampling frames using stratified random systematic sampling. There were separate sampling frame databases for the "active" landline and cell phone numbers received from the Dutch firm. A 3-digit stratum code was assigned to all the first phase "active" phone numbers in each sampling frame database; the first digit of the stratum code was 1 for landline phone numbers and 2 for cell phone numbers, and the last two digits corresponded to a serial state code (from 1 to 27). The Complex Samples module of SPSS was used for the stratified random systematic sampling, after ordering the corresponding sampling frame of "active" phone

[3] From July 26 to August 20, there were required 7 call-attempts before discarding a phone number, being one per turn of the day (morning/afternoon/evening) for two days plus one in the weekend. From August 23 to September 24, 5 call-attempts were required, being in different turns in two different week working days and one in the weekend. After that, until the end of fieldwork, the call in the weekend was not required given the low response rates.

numbers by state (stratum) and area code. The total number of landline and cell phone numbers selected for each state, including reserves, is shown in Table 2.

All the batches from each sampling frame were selected at the same time, and then divided into systematic replicates for the individual batches. First the full sample of cell phone numbers were assigned unique serial interview numbers in groups of 30 in the same order in which they were selected. For example, the first 30 sample cell phones in the sample were assigned interview code 1, the second group of 30 numbers was assigned interview code 2, etc. A total of 2,550 interview codes were assigned to the full sample of cell phone numbers in groups of 30 numbers each. A similar procedure was used to assign interview codes to the full landline phone numbers in groups of 50, starting with interview code 2551 (following the last interview number for the cell phone sample), and ending with code 3000. One advantage of assigning a consecutive group of sample phone numbers to one interview is that given the ordering of the frame and the sample by area code within a state, it is more likely that the replacement phones will be from the same area code within a state.

In the case of the cell phones, serial numbers from 1 to 30 were assigned to the phone numbers corresponding to each interview code. This procedure was used for facilitating the selection of 6 systematic replicates corresponding to the 6 batches of sample cell phone numbers. A similar procedure was used for the full sample of landline phone numbers; in this case serial numbers from 1 to 50 were assigned to all the phone numbers for a particular interview code. The next step involved dividing the sample cell and landline phones into systematic replicates corresponding to the individual batches. In the case of the sample of cell phone numbers, the phones with serial numbers from 1 to 5 for each interview code were assigned to batch 1, those with serial numbers from 6 to 10 were assigned to batch 2, etc., and finally the cell phone numbers with serial numbers 26 to 30 were assigned to batch 6. For the full sample of landline phones, serial numbers from 1 to 10 were assigned to batch 1, those with serial numbers 21 to 30 were assigned to batch 2, etc., and finally those with serial numbers from 41 to 50 were assigned to batch 5. In each case the serial numbers for the phones corresponding to each interview code represent the order in which replacement phone numbers will be selected, starting with batch 1. This approach provided more operational control and facilitated the logistics of the replacements for the CATI operation. As an example, in the case of the cell phone sample, when for a particular interview it is necessary to replace the phone number with serial number 5, the second batch would be used to access the phone number with serial number 6 for that same interview code.

An Excel file was generated for each batch of the sample landline and cell phone numbers, specifying the batch number, the interview codes and the phone serial numbers. The World Bank initially provided the survey firm with one batch each for the sample landline and cell phone numbers, and only released the next batches as needed.

5. Results of Brazil HFPS implementation

The CATI data collection for the Brazil HFPS proceeded at a slower pace than expected mainly due to the relatively low response rate, and the natural spacing of time required to implement the quality control protocols in terms of the number of call-attempts, the initial requirement for call-attempts on weekends, etc. The data collection period was originally intended to be a little over a month, but it had to be extended. As the progress of the completed call interviews by state was being monitored, at a certain point it was decided to relax some of the protocols, such as removing the requirement for weekend call-attempts, and reducing the total number of callbacks to 5. A review of the CATI records indicated that following the fifth call-attempt the remaining calls had a very low completion rate, so this adjustment of the protocol did not have much effect on quality of the results.

A deadline of 1 October was set to end the data collection (totaling 10 weeks of fieldwork), following the completion of 1836 cell phone interviews and 330 landline phone interviews, for a total of 2,166 completed phone interviews. In this case the total number of completed cell and landline phone interviews in each state were considered the final sample size for the calculation of the weights. In the case of the states of Amapá and Roraima, no landline phone interview was completed during the CATI operation. Hence, for the purposes of calculating the weights it was necessary to "collapse" or combine any state without any landline phone numbers with the landline stratum of a neighboring state. Therefore, the landline stratum of Amapá was combined with that of Pará, and the landline stratum of Roraima was combined with that of Amazonas. In this case the landline phone respondents of Pará will also represent Amapá, and those of Amazonas will also represent Roraima, in order to complete the national estimates for Brazil.

Table 3 presents a summary of the completed number of landline and cell phone interviews by state, and the corresponding landline and cell phone response rates. The response rate for each stratum is calculated as the number of completed interviews divided by the corresponding number of eligible sample phone numbers that were called.

Table 3. Summary of results from Brazil HFPS implementation by state

State	Total number of completed phone interviews	Completed landline phone interviews	Landline phone response rate	Completed cell phone interviews	Cell phone response rate
Acre	9	1	8.3%	8	8.1%
Alagoas	34	5	11.1%	29	10.7%
Amapá	6	0	0.0%	6	7.1%
Amazonas	43	3	5.3%	40	10.8%
Bahia	134	17	8.1%	117	7.1%
Ceará	92	12	9.7%	80	9.5%
Distrito Federal	39	6	17.6%	33	14.8%
Espírito Santo	46	8	27.6%	38	12.6%
Goiás	71	9	7.5%	62	8.5%
Maranhão	71	10	10.9%	61	8.8%
Mato Grosso	37	5	11.4%	32	8.7%
Mato Grosso do Sul	25	2	5.7%	23	8.4%
Minas Gerais	231	38	14.2%	193	9.3%
Pará	84	10	7.5%	74	8.9%
Paraíba	46	6	11.1%	40	11.6%
Paraná	117	19	12.8%	98	8.5%
Pernambuco	95	14	9.7%	81	9.3%
Piauí	36	7	21.2%	29	7.5%
Rio de Janeiro	186	31	13.7%	155	9.2%
Rio Grande do Norte	35	4	6.5%	31	8.5%
Rio Grande do Sul	98	17	10.8%	81	5.4%
Rondônia	19	3	8.1%	16	8.7%
Roraima	9	0	0.0%	9	16.4%
Santa Catarina	70	11	10.3%	59	7.4%
São Paulo	493	87	14.1%	406	8.1%
Sergipe	22	3	6.3%	19	7.3%
Tocantins	18	2	12.5%	16	14.7%
Total	2,166	330	11.3%	1,836	8.5%

Source: Own compilation based on fieldwork data from Brazil Phone Survey.

Table 4 shows the eligibility of each category of the final call status. In the case of the "Others" category, the specific reason for each call had been recorded and was manually coded as either eligible or not eligible.

Table 4. Final call status categories for cell and landline phone calls, and corresponding eligibility

Status category	Eligibility for survey
Concluded Interviews	Eligible
Incomplete and refused	Eligible
Incomplete, call again	Eligible
Underage	Eligible for landline, not eligible for cell phones
The responsible is not available	Eligible
Refusal	Eligible
Rescheduled	Eligible
Busy or no answers	Eligible
Invalid, Inexistent, Inactive	Not eligible
Commercial	Not eligible
Others - eligible[4]	Eligible
Others – not eligible	Not eligible

Cellphone numbers where the owner was underage were considered ineligible. In the case of the landline phones, the underage category was considered eligible, since it is highly likely that there are also adults living in the household where the phone was answered from.

6. Calculation of phone-level probabilities and weights for Brazil HFPS

As mentioned in the introduction, the sample design and weighting procedures for the Brazil HFPS are designed to be consistent with those used for the regional HFPS in various countries of Latin America and the Caribbean. The report "COVID-19 High-Frequency Phone Survey (HFPS) in Latin America: Technical Note on Sampling Design, Weighting and Estimation"[5], by Ramiro Flores Cruz, is used as a reference. Many of the terms in the weighting formulas defined here are the same as those specified in that report.

The overall objective of weighting procedures is to expand the data in each stratum to represent the distribution of the frame. In general, the weights for each sampling unit (phone, household or individual) are calculated as the inverse of the corresponding probabilities of selection, taking into account each sampling stage. Since the sampling units in different strata and groups may be selected with different probabilities, the weights will compensate for any differential sampling rates. Given the nature of the RDD sampling process, the sampling was carried out in two different phases.

[4] "Others - not eligible" include numbers whose final status was coded as "Others" by the surveyor, but in which the respective comments contained some information that characterized non-eligibility. For example, when it was reported that the number was commercial or that a child answered to the mobile phone. While "Others – eligible" are the cases in which the comments report that the number would be valid despite being coded as "Others", e.g. the responsible for the line was away.

[5] Florez Cruz, Ramiro (2021).

Since the sampling involves the selection of landline and cell phone numbers that are called for the interviews, we first calculate the probabilities and corresponding weights at the phone level (separately for landline and cell phones). These weights are then used with the relevant survey data to calculate the individual and household weights.

As indicated previously, the first phase of the RDD sampling process involves selecting a very large number of possible phone numbers as random combinations of digits within the area codes defined for each state, separately for the landline and cell phone numbers. Most of these random phone numbers are not working, but the proportion of these phone numbers that are working establishes the level of the actual frame. The first phase is designed to obtain a large sampling frame of "active" phone numbers for the selection of the second phase sample of landline and cell phone numbers. The probabilities and weights take into account both phases of the sampling process.

Using the terminology in the reference report, the first-phase inclusion probabilities of cell phone and landline numbers can be expressed as follows:

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

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

where:

$p^{C}_{(1)hi}$ = first-phase inclusion probability of the *i*-th active cell phone number in stratum (state) *h*

$n^{C}_{(1)h}$ = total number of cell phone numbers selected in the first-phase sample of cell phones, composed of $n^{C}_{(1)hA}$ active cell phones and $n^{C}_{(1)hIN}$ inactive cell phones in stratum *h*

$N^{C}_{(1)h}$ = total number of all possible cell phone numbers (frame size) according to the national phone numbering plan in stratum *h*

$p^{L}_{(1)hi}$ = first-phase inclusion probability of the *i*-th active landline phone number in stratum *h*

$n^{L}_{(1)h}$ = total number of landline phone numbers selected in the first-phase sample of landline phones in stratum *h,* composed of $n^{L}_{(1)hA}$ active landline phones and $n^{l}_{(1)hIN}$ inactive landline phones

$N^{L}_{(1)h}$ = total number of all possible landline phone numbers (frame size) in stratum *h* according to the national phone numbering plan

The firm that provided the first phase sample of cell and landline phone numbers also provided a list of cell and landline business phone numbers that were excluded from the RDD selection. A count of these phone numbers by stratum was subtracted from the stratum frame size ($N^{L}_{(1)h}$, $N^{C}_{(1)h}$) for the calculation of the probabilities.

The second-phase sample of landline and cell phones were selected from the corresponding first-phase samples of active cell and landline phone numbers. The conditional second-phase inclusion probabilities of cell and landline phones can be expressed as follows:

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

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

where

$p^{C}_{(2)hi|(1)hi}$ = second-phase inclusion probability of the *i*-th sample active cell phone number in stratum *h*, conditional on being selected in the first phase

$n^{C}_{(2)hA}$ = number of sample active cell phones selected for the second phase in stratum h

$p^{L}_{(2)hi|(1)hi}$ = second-phase inclusion probability of the *i*-th sample active landline phone number in stratum *h*, conditional on being selected in the first phase

$n^{L}_{(2)hA}$ = number of sample active landline phones selected for the second phase in stratum h

Since it was found that many of the second phase sample phone numbers classified as active at the first phase were either not working or not eligible during the CATI data collection operations, it was necessary to adjust the second phase inclusion probabilities to represent the eligible phone numbers. In this case it is necessary to adjust the denominators of the second stage cell and landline probabilities ($n^{C}_{(1)hA}$ and $n^{L}_{(1)hA}$) by the proportion of these sample phone numbers that are estimated to be eligible, based on the results of the CATI data collection operation. For the calculation of the weights the final count of completed cell and landline phone interviews is used as the final sample size for cell phones ($n^{C}_{(2)hAc}$) and landline phones ($n^{L}_{(2)hAc}$). In this way, the weights are automatically adjusted for nonresponse. The completed interviews for each type of phone in each state will represent all the eligible phone numbers in this stratum that were called, including the non-respondents. This nonresponse adjustment is based on the assumption that

the characteristics of the non-respondents are similar to those of the persons and households interviewed in the same class (state and type of phone). However, there may be a slight bias given the relatively low response rate and potentially different characteristics of non-respondents. The adjusted conditional second stage probabilities were calculated as follows:

$$p'^{C}_{(2)hi|(1)hi} = \frac{n^{C}_{(2)hA}}{n^{C}_{(1)A} \times \frac{n^{C}_{(2)hA}}{n'^{C}_{(2)hA}}} = \frac{n^{C}_{(2)hAc}}{n^{C}_{(1)hA}} \times \frac{n'^{C}_{(2)hA}}{n^{C}_{(2)hAe}}$$

$$p'^{L}_{(2)hi|(1)hi} = \frac{n^{L}_{(2)hAc}}{n^{L}_{(1)A} \times \frac{n^{L}_{(2)hA}}{n'^{L}_{(2)hA}}} = \frac{n^{L}_{(2)hAc}}{n^{L}_{(1)hA}} \times \frac{n'^{L}_{(2)hA}}{n^{L}_{(2)hA}}$$

where:

$p'^{C}_{(2)hi|(1)hi}$ = adjusted second-phase inclusion probability of the *i*-th sample active cell phone number in stratum *h*, conditional on being selected in the first phase

$n^{C}_{(2)hA}$ = number of completed cell phone interviews in stratum h

$n^{C}_{(2)hAe}$ = number of eligible cell phone numbers called in stratum h

$n'^{C}_{(2)hA}$ = total number of cell phone numbers called in stratum h, including calls that are not eligible

$p'^{L}_{(2)hi|(1)hi}$ = adjusted second-phase inclusion probability of the *i*-th sample active landline phone number in stratum *h*, conditional on being selected in the first phase

$n^{L}_{(2)hAc}$ = number of completed landline phone interviews in stratum h

$n^{L}_{(2)hAe}$ = number of eligible landline phone numbers called in stratum h

$n'^{L}_{(2)hA}$ = total number of landline phone numbers called in stratum h, including calls that are not eligible

The overall adjusted probabilities of the final sample of completed cell and landline phone interviews can be expressed as follows:

$$p'^{C}_{hi} = p^{C}_{(1)hi} \times p'^{C}_{(2)hi|(1)hi} = \frac{n^{C}_{(1)hA}+n^{C}_{(1)hIN}}{N^{C}_{(1)h}} \times \frac{n^{C}_{(2)hAc}}{n'^{C}_{(1)hA}} \times \frac{n'^{C}_{(2)hA}}{n^{C}_{(2)hAe}}$$

$$p'^{L}_{hi} = p^{L}_{(1)hi} \times p'^{L}_{(2)hi|(1)hi} = \frac{n^{L}_{(1)hA}+n^{L}_{(1)hIN}}{N^{L}_{(1)h}} \times \frac{n^{L}_{(2)hAc}}{n^{L}_{(1)hA}} \times \frac{n'^{L}_{(2)hA}}{n'^{L}_{(2)hAe}}$$

The adjusted overall phone-level probabilities of selection for the eligible cell and landline phones were calculated at the stratum (state) level in an Excel file, with separate spreadsheets for the cell and landline phone samples. Each spreadsheet included information from the corresponding sampling frames for each component of the formulas in the probabilities by stratum and type of phone expressed above, and from the summary of all the landline and cell phone calls by final status category and state from the survey CATI operation. In the case of the combined landline strata for Amapá-Pará and Roraima-Amazonas, each count for the frame and sample were summed across the corresponding combined states.

7. Calculation of household and individual weights for Brazil HFPS

The selection probabilities of households and individuals 18 years of age and older are based on the inclusion probabilities of the cell and landline phones through which they can be reached. Therefore, the computation of household and individual weights should account for multiple chances of selection and for the overlapping between the cell phone and landline frames. This multiplicity weighting adjusts estimates to eliminate the over-representation of households and individuals in the sample that can be reached through more phone numbers than other households and individuals, and eliminates the chance for multiplicity sampling bias.

There is multiplicity probability when a household has a larger selection probability because it can be selected through different phone numbers. Households with more than one cell phone or more than one landline phone number have higher selection probabilities. Therefore, it is necessary to adjust these probabilities to account for this increased chance of selection. The number of cell and landline phones is asked during the interview in the questionnaire. The multiplicity-adjusted *household* selection probabilities in each frame are calculated as follows:

$p^{C}_{mhj} = m_{chj} \times p'^{C}_{hi}$, if the household only has cell phones

$p^{L}_{mhj} = m_{lhj} \times p'^{L}_{hi}$, if the household only has landline phones

where:

p^{C}_{mhj} = selection probability of the *j*-th household in stratum *h* when contacted through a cell phone, adjusted for multiplicity of working cell phones in the household

m_{chj} = number of working cell phones in the *j*-th household in stratum *h*

p^{L}_{mhj} = 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_{lhj} = number of working landlines in the *j*-th household in stratum *h*

In the case of households with both cell and landline phones (dual cases) it is necessary to adjust the selection probabilities for multiplicity. For these cases the household probability is calculated as follows:

$$m_{chj} \times p'^{C}_{hi} + m_{lhj} \times p'^{L}_{hi} - m_{chj} \times p'^{C}_{hi} \times m_{lhj} \times p'^{L}_{hi}$$

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 phone equals the selection probability of his or her household, conditional on the number of working landline phones in the household, over the number of individuals 18 years of age and older in the household. These individual probabilities are calculated as follows:

$p^{C}_{hk} = p'^{C}_{hi}$, if the individual only has a cell phone

$p^{L}_{hj} = \frac{m_{lhj} \times p'^{L}_{hi}}{a_{hj}}$, if the individual lives in a household with only landline phones

where:

p^{C}_{hk} = selection probability of the *k*-th individual in stratum h when contacted through a cell phone

p^{L}_{hjk} = selection probability of the *k*-th individual in stratum *h* when contacted through a landline phone in the *j*-th household

a_{hj} = number of eligible adults (18 years of age or older) in the *j-th* household in stratum h

Individuals with a cell phone who live in a household that also has a landline phone have a higher probability of being selected than those with only cell phones or only landline phones. In this case the individual probability is calculated as follows:

$$p'^{C}_{hi} + \frac{m_{lhj} \times p'^{L}_{hi}}{a_{hj}} - p'^{C}_{hi} \times \frac{m_{lhj} \times p'^{L}_{hi}}{a_{hj}}$$

The household and individual design weights are calculated as the inverse of the corresponding probabilities specified above.

The 2021 Brazil HFPS included a child module that was administered for one random eligible child selected in each household. In this case the child weight is equal to the household weight multiplied by the number of eligible children in the household.

A database with the survey data for all the completed interviews was generated with information on the number of cell phones, landline phones, individuals 18 years old, and eligible children under 18 in each household. Then the phone-level probabilities and weights for the cell and landline phones for the corresponding state were merged in this database for the calculation of the household, individual and child weights. The SPSS software was used for the calculation of

the household, individual and child weights, using the formulas specified above. The SPSS syntax used for calculating these weights is presented in Annex A.

8. Calibration and trimming of the weights for Brazil HFPS

As described previously, the design (probability-based) weights specified above have an implicit adjustment for nonresponse at the level of the geographic stratum and type of phone. However, given differential nonresponse by some characteristics such as sex and age group, the weighted distribution of the survey data by these characteristics may be different from the actual distribution shown in demographic projections or other sources of data. For this reason, the household and individual design weights were further adjusted based on calibration by demographic and socioeconomic characteristics and trimming of extreme values.

Calibration allowed reflecting the total population with phone by region, sex, age and educational attainment available from external national official sources. In Brazil, the calibration totals by region, sex and age were based on the 2021 official population projections, whereas the education distribution was taken from the *Pesquisa Nacional por Amostra de Domicilios – Continua* 2019. The calibration method was raking, which was most suitable given that all available auxiliary variables (region, sex, age groups and educational attainment) were categorical and with multiple categories. It used a logit distance function since it generally fitted a more exact adjustment on the four calibration auxiliaries. The procedures followed to calibrate and trim the weights are consistent with those used for the regional HFPS.

ANNEX A - SPSS syntax for generating household, individual and child weights for the Brazil HFPS data

i) Calculate probabilities for household weights

```
IF  (tipo_tel = 12 & mobile = 0) p_hh=landline * p_l.
EXECUTE.
IF  (tipo_tel = 12 & mobile > 0) p_hh=landline * p_l + mobile * p_m - landline * p_l * mobile * p_m.
EXECUTE.
IF  (tipo_tel = 13 & landline = 0) p_hh=mobile * p_m.
EXECUTE.
IF  (tipo_tel = 13 & landline > 0) p_hh=landline * p_l + mobile * p_m - landline * p_l * mobile * p_m.
EXECUTE.

*Calculate probabilities for individual weights

IF  (tipo_tel = 12 & mobile = 0) p_indiv=landline * p_l / adults.
EXECUTE.
IF  (tipo_tel = 12 & mobile > 0) p_indiv=p_m + landline * p_l / adults  - p_m *  landline * p_l / adults.
EXECUTE.
IF  (tipo_tel = 13 & landline = 0) p_indiv=p_m.
EXECUTE.
IF  (tipo_tel = 13 & landline > 0) p_indiv=p_m + landline * p_l / adults  - p_m *  landline * p_l / adults.
EXECUTE.
```

ii) Calculate household and individual weights

```
COMPUTE wt_hh=1 / p_hh.
EXECUTE.
COMPUTE wt_indiv=1 / p_indiv.
EXECUTE.
```

iii) Calculate child weight

```
COMPUTE wt_child=wt_hh * cri_elegiveis.
EXECUTE.
```

**Note:
p_l = adjusted probability for landline phone in state
p_m = adjusted probability for mobile phone in state
p_hh = household probability
p_indiv = individual probability
wt_hh = household weight
wt_indiv = individual weight
wt_child = child weight