Source of the Data and Accuracy of the Estimates for the 2020 Household Pulse Survey – Phase 2

Interagency Federal Statistical Rapid Response Survey to Measure Household Experiences during the Coronavirus (COVID-19) Pandemic

SOURCE OF THE DATA

The 2020 Household Pulse Survey (HPS), an experimental data product, is an Interagency Federal Statistical Rapid Response Survey to Measure Household Experiences during the Coronavirus (COVID-19) Pandemic, conducted by the United States Census Bureau in partnership with seven other agencies from the Federal Statistical System:

- Bureau of Labor Statistics (BLS)
- National Center for Health Statistics (NCHS)
- Department of Agriculture Economic Research Service (ERS)
- National Center for Education Statistics (NCES)
- Department of Housing and Urban Development (HUD)
- Social Security Administration (SSA)
- Bureau of Transportation Statistics (BTS)

These agencies collaborated on the design and provided content for the HPS, which was also reviewed and approved by the Office of Management and Budget (OMB). (OMB # 0607-1013; expires 10/31/2020.)

The HPS asks individuals about their experiences regarding employment status, spending patterns, food security, housing, physical and mental health, access to health care, program receipt, and educational disruption. The ability to understand how individuals are experiencing this period is critical to governmental and non-governmental response in light of business curtailment and closures, stay-at-home and safer-at-home orders, school closures, changes in consumer patterns and the availability of consumer goods, and other abrupt and significant changes to American life.

The HPS is designed to produce estimates at three different geographical levels. The first level, the lowest geographical area, is for the 15 largest Metropolitan Statistical Areas (MSAs). The second level of geography is state-level estimates for each of the 50 states plus the District of Columbia, and the final level of geography is national-level estimates.

The U.S. Census Bureau conducted Phase 1 of the 2020 HPS every week starting April 23, 2020. For details of Phase 1 see the Source and Accuracy Statements at: https://www.census.gov/programs-surveys/household-pulse-survey/technical-documentation.html. Phase 1 of the Household Pulse Survey was collected and disseminated on a weekly basis. Phase 1 collection ended July 21, 2020.

The first data collection of Phase 2 of the 2020 HPS was conducted over 13 days starting August 19, 2020 and ending August 31, 2020. Despite going to a two-week collection period, the Household Pulse Survey continues to call these collection periods "weeks" for continuity with Phase 1. The table below provides the data

collection periods. This document will be updated after each data collection period until the end of the survey.

Table 1. Data Collection Periods for Phase 2 of the 2020 Household Pulse Survey

Data Collection Period	Start Date	Finish Date
Week* 13	August 19, 2020	August 31, 2020

* For Phase 2 the Household Pulse Survey continues to call these collection periods "weeks" for continuity with Phase 1.

Sample Design

The HPS utilizes the Census Bureau's Master Address File (MAF) as the source of sampled housing units (HUs). The sample design was a systematic sample of all eligible HUs, with adjustments applied to the sampling intervals to select a large enough sample to create state level estimates[1] and estimates for the top 15 MSAs. Sixty-six independent sample areas were defined. For each data collection period, independent samples were selected and each sampled HU was interviewed once, unlike the Phase 1 of the 2020 HPS.

Sample sizes were determined such that a three percentage coefficient of variation (CV) for an estimate of 40 percent of the population would be achieved for all sample areas with the exception of the 11 smallest states. In these smaller states, the sample size was reduced to produce a 3.5 percent CV. The overall sample sizes within the sampling areas were adjusted for an anticipated response rate of nine percent. For those counties in one of the top MSAs, the sampling interval was adjusted to select the higher of the sampling rate for either the state or MSA.

To enable the use of a rapid deployment internet response system, we added email and mobile telephone numbers from the Census Bureau Contact Frame to the MAF. Since 2013, the Census Bureau has maintained contact frames to allow appended contact information onto sample units within household sample frames to aid in contacting respondents at those households. The primary motivation for creating this contact frame was to support research on potential contact strategies for the 2020 Census.

The Contact Frame information is maintained in two separate files – one containing phone numbers (both landline and cell phones) and the other containing email addresses. Information is obtained primarily from commercial sources, with additions from respondents to the American Community Survey (ACS) and Census tests, as well as participants in food and other assistance programs from a few states, as well as from the Alaska Permanent Fund Division. Commercial sources were evaluated against respondent reported phone numbers to determine which sources would be acquired, after determining which vendors provided the best value for the government.

Commercial, survey, and administrative record data providers link phone numbers and email addresses to physical addresses before providing them for the Contact Frame.

[1] Including the District of Columbia as a state.

Addresses were matched to the MAF. For addresses matched with confidence, the contact information was added to the frame along with the unique identifier from the MAF. Approximately 140,000,000 housing units are represented in the MAF and were considered valid for sampling. The phone frame contains over a billion phone/address pairs, and the email frame contains over 686 million well-formed email/address pairs. The phone frame contains phone/address pairs for over 88 percent of addresses in the country, and over three quarters of those phone numbers were acquired in the past two years. The email frame contains email/address links for almost 80 percent of addresses in the country, and two-thirds of those emails were acquired in past two years. Unique phone numbers and email addresses were identified and assigned to only one HU. The HUs on MAF were then limited to these addresses on the Contact Frame as the final eligible HUs for the HPS. Table 2 shows the number of addresses with contact information.

Table 2. Number of Addresses on the Master Address File with Contact Information

Total Addresses	144,136,000
Addresses with any contact information:	116,533,000
Addresses with cell phone	87,883,000
Addresses with email	107,075,000

Source: U.S. Census Bureau Master Address File Extracts and Contact Frame

Sampled households were contacted by both email if an email was available, and by text if a phone number was available. Contact email addresses and phone numbers were only sent on weekdays and reminders were sent to nonrespondents.

The Census Bureau conducted the HPS online using Qualtrics as the data collection platform. Qualtrics is currently used at the Census Bureau for research and development surveys and provides the necessary agility to deploy the HPS quickly and securely. It operates in the Gov Cloud, is FedRAMP authorized at the moderate level, and has an Authority to Operate from the Census Bureau to collect personally-identifiable and Title-protected data.

Approximately 1,033,000 housing units were selected from the sampling frame for the first collection period of Phase 2. Approximately 109,000 respondents answered the online questionnaire. Table 3 shows the sample sizes and the number of responses by collection period for Phase 2 of the HPS.

Table 3. Sample Size and Number of Respondents at the National Level

Data Collection Period	Sample Size	Number of Respondents
Week* 13	1,032,959	109,051

Source: U.S. Census Bureau, 2020 Household Pulse Survey
* For Phase 2 the Household Pulse Survey continues to call these collection periods "weeks" for continuity with Phase 1.

State-level sample sizes and number of responses can be found in Table A1 on the Appendix A1 tab in the State-level Quality Measures spreadsheet at

https://www.census.gov/programs-surveys/household-pulse-survey/technical-documentation.html under the Source and Accuracy Statements section.

Estimation Procedure

The final HPS weights are designed to produce biweekly estimates for the total persons age 18 and older living within HUs. These weights were created by adjusting the household-level sampling base weights by various factors to account for nonresponse, adults per household, and coverage.

The sampling base weights for each incoming sample in each of the 66 sample areas are calculated as the total eligible HUs in the sampling frame divided by the number of eligible HUs selected for interviews each week. Therefore, the base weights for all sampled HUs sum to the total number of HUs for which contact information is known.

The final HPS person weights are created by applying the following adjustments to the sampling base weights:

1. Nonresponse adjustment – the weight of all sample units that did not respond to the HPS are evenly allocated to the units that did respond within the same sample collection period, sample area (MSA or balance of state) and state. After this step, the weights of all respondents sum to the total HUs with contact information in the sampling frame.
2. Occupied HU ratio adjustment – this adjustment corrects for undercoverage in the sampling frame by inflating the HU weights after the nonresponse adjustment to match independent controls for the number of occupied HUs within each state. Each sampled respondent was assigned to the state where they reported their current address, which may be different from the selected state. For this adjustment, the independent controls are the 2018 American Community Survey (ACS) one-year, state-level estimates available at www.census.gov[2].
3. Person adjustment – this adjustment converts the HU weights into person weights by multiplying them by the number of persons age 18 and older that were reported to live within the household. The number of adults is based on the subtracting the number of children under 18 in the household from the number of total persons in the household. This number was capped at 10 adults. If the number of total persons and number of children was not reported, then it is imputed.
4. Iterative Ranking Ratio to Population Estimates – this procedure controls the person weights to independent population controls by various demographics within each state. The ratio adjustment is done through an iterative raking procedure to simultaneously control the sample estimates to two sets of population controls -- Educational attainment estimates from the 2018 1-year ACS estimates (Table B115001)[3] by age and sex, and the July 1, 2020 Hispanic origin/race by age and sex estimates from the Census Bureau's Population Estimates Program (PEP). PEP

[2] The one-year estimates are at this URL: https://www2.census.gov/programs-surveys/acs/summary_file/2018/data/
[3] The1-year state-level detailed table B15001 is located at this URL: https://www.census.gov/programs-surveys/acs/data/summary-file.html, Then clicking 1-year summary file -> 1_year_detailed_tables -> state -> B15001.csv

provided July 1, 2020 household population estimates by single year of age (0-84, 85+), sex, race (31 groups), and Hispanic origin for states from the Vintage 2019 estimates series[4]. The ACS 2018 estimates were adjusted to match the 2020 pop controls within states by sex, and the five age categories in the ACS educational attainment estimates. Tables 4 and 5 show the demographic groups formed.

Before the raking procedure was applied, cells containing too few responses were collapsed to ensure all cells met the minimum response count requirement. The cells after collapsing remained the same throughout the raking. These collapsed cells were used in the calculation of replicate weights.

Table 4: Educational Attainment Population Adjustment Cells within State

Age	No HS diploma Male	No HS diploma Female	HS diploma Male	HS diploma Female	Some college or Associate's degree Male	Some college or Associate's degree Female	Bachelor's degree or higher Male	Bachelor's degree or higher Female
18-24								
25-34								
35-44								
45-64								
65+								

Table 5: Race/Ethnicity Population Adjustment Cells within State

Age	Hispanic Any Race Male	Hispanic Any Race Female	Non-Hispanic White-Alone Male	Non-Hispanic White-Alone Female	Non-Hispanic Black-Alone Male	Non-Hispanic Black-Alone Female	Non-Hispanic Other Races Male	Non-Hispanic Other Races Female
18-24								
25-29								
30-34								
35-39								
40-44								
45-49								
50-54								
55-64								
65+								

The detailed tables released for this experimental Household Pulse Survey show frequency counts rather than percentages. Showing the frequency counts allows data users to see the count of cases for each topic and variable that are in each response category and in the 'Did Not Report' category. This 'Did Not Report' category is not a commonly used data category

[4] The Vintage 2019 estimates methodology statement is available at this URL: https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2010-2019/natstcopr-methv2.pdf The Modified Race Summary File methodology statement is available at this URL: https://www2.census.gov/programs-surveys/popest/technical-documentation/methodology/modified-race-summary-file-method/mrsf2010.pdf

in U.S. Census Bureau tables. Most survey programs review these missing data and statistically assign them to one of the other response categories based on numerous characteristics.

In these tables, the Census Bureau recommends choosing the numerators and denominators for percentages carefully, so that missing data are deliberately included or excluded in these counts. In the absence of external information, the percentage based on only the responding cases will most closely match a percentage that would result from statistical imputation. Including the missing data in the denominator for percentages will lower the percentages that are calculated.

Microdata will be available by FTP in the future. Users may develop statistical imputations for the missing data but should ensure that they continue to be deliberate and transparent with their handling of these data.

ACCURACY OF THE ESTIMATES

A sample survey estimate has two types of error: sampling and nonsampling. The accuracy of an estimate depends on both types of error. The nature of the sampling error is known given the survey design; the full extent of the nonsampling error is unknown.

Sampling Error

Since the HPS estimates come from a sample, they may differ from figures from an enumeration of the entire population using the same questionnaires, instructions, and enumeration methods. For a given estimator, the difference between an estimate based on a sample and the estimate that would result if the sample were to include the entire population is known as sampling error. Standard errors, as calculated by methods described below in "Standard Errors and Their Use," are primarily measures of the magnitude of sampling error. However, the estimation of standard errors may include some nonsampling error.

Nonsampling Error

For a given estimator, the difference between the estimate that would result if the sample were to include the entire population and the true population value being estimated is known as nonsampling error. There are several sources of nonsampling error that may occur during the development or execution of the survey. It can occur because of circumstances created by the respondent, the survey instrument, or the way the data are collected and processed. Some nonsampling errors, and examples of each, include:

- Measurement error: The respondent provides incorrect information, the respondent estimates the requested information, or an unclear survey question is misunderstood by the respondent.
- Coverage error: Some individuals who should have been included in the survey frame were missed.
- Nonresponse error: Responses are not collected from all those in the sample or the respondent is unwilling to provide information.

- Imputation error: Values are estimated imprecisely for missing data.

To minimize these errors, the Census Bureau applies quality control procedures during all stages of the production process including the design of the survey, the wording of questions, and the statistical review of reports.

Two types of nonsampling error that can be examined to a limited extent are nonresponse and undercoverage.

Nonresponse

The effect of nonresponse cannot be measured directly, but one indication of its potential effect is the nonresponse rate. Table 6 shows the unit response rates by collection period.

Table 6. National Level Weighted Response Rates by Collection Period for the 2020 Household Pulse Survey

Data Collection Period	Response Rate (Percent)
Week* 13	10.3

Source: U.S. Census Bureau, 2020 Household Pulse Survey
* For Phase 2 the Household Pulse Survey continues to call these collection periods "weeks" for continuity with Phase 1.

State-level response rates can be found in Table A1 on the Appendix A1 tab in the State-level Quality Measures spreadsheet at https://www.census.gov/programs-surveys/household-pulse-survey/technical-documentation.html under the Source and Accuracy Statements section.

In accordance with Census Bureau and Office of Management and Budget Quality Standards, the Census Bureau will conduct a nonresponse bias analysis to assess nonresponse bias in the HPS.

Responses are made up of complete interviews and sufficient partial interviews. A sufficient partial interview is an incomplete interview in which the household or person answered enough of the questionnaire to be considered a complete interview. Some remaining questions may have been edited or imputed to fill in missing values. Insufficient partial interviews are considered to be nonrespondents.

Undercoverage

The concept of coverage with a survey sampling process is defined as the extent to which the total population that could be selected for sample "covers" the survey's target population. Missed housing units and missed people within sample households create undercoverage in the HPS. A common measure of survey coverage is the coverage ratio, calculated as the estimated population before poststratification divided by the independent population control. The national household-level coverage ratio is 0.96. State household-level coverage ratios can be found in Table A1 on the Appendix A1 tab in the State-level Quality Measures spreadsheet at https://www.census.gov/programs-surveys/household-

pulse-survey/technical-documentation.html under the Source and Accuracy Statements section.

HPS person coverage varies with age, sex, Hispanic origin/race, and educational attainment. Generally, coverage is higher for females than for males and higher for non-Blacks than for Blacks. This differential coverage is a general problem for most household-based surveys. The HPS weighting procedure tries to mitigate the bias from undercoverage within the raking procedure. However, due to small sample sizes, some demographic cells needed collapsing to increase sample counts within the raking cells. In this case convergence to both sets of the population controls was not attained. Therefore, the final coverage ratios are not perfect for some demographic groups. Table 7 shows the coverage ratios for the person demographics of age, sex, Hispanic origin/race, and educational attainment before and after the raking procedure is run.

Table 7. Person-Level Coverage Ratios at the National Level for 2020 Household Pulse Survey Before and After Raking for Collection Week* 13: August 19 – August 31, 2020

Demographic Characteristic	Before Raking	After Raking
Total Population	1.04	1.00
Male	0.89	1.00
Female	1.17	1.00
Age 18-24	0.51	0.93
Age 25-29	0.68	0.97
Age 30-34	0.90	1.05
Age 35-39	1.16	1.02
Age 40-44	1.29	1.03
Age 45-49	1.27	0.99
Age 50-54	1.37	1.02
Age 55-64	1.23	1.01
Age 65+	1.00	0.99
Hispanic	0.74	1.01
Non-Hispanic white-only	1.16	1.00
Non-Hispanic black-only	0.77	0.99
Non-Hispanic other races	1.08	1.00
No high-school diploma	0.22	0.72
High-school diploma	0.48	1.12
Some college or associate's degree	1.12	1.00
Bachelor's degree or higher	1.77	1.00

Source: U.S. Census Bureau, 2020 Household Pulse Survey
* For Phase 2 the Household Pulse Survey continues to call these collection periods "weeks" for continuity with Phase 1.

The previous data collection's national person-level coverage ratios and state person-level coverage ratios can be found in Table A2 on the Appendix A2 tab in the State-level Quality Measures spreadsheet at https://www.census.gov/programs-surveys/household-pulse-survey/technical-documentation.html under the Source and Accuracy Statements section.

Biases may also be present when people who are missed by the survey differ from those interviewed in ways other than age, sex, Hispanic origin/race, educational attainment, and

state of residence. How this weighting procedure affects other variables in the survey is not precisely known. All of these considerations affect comparisons across different surveys or data sources.

Comparability of Data

Data obtained from the HPS and other sources are not entirely comparable. This is due to differences in data collection processes, as well as different editing procedures of the data, within this survey and others. These differences are examples of nonsampling variability not reflected in the standard errors. Therefore, caution should be used when comparing results from different sources.

A Nonsampling Error Warning

Since the full extent of the nonsampling error is unknown, one should be particularly careful when interpreting results based on small differences between estimates. The Census Bureau recommends that data users incorporate information about nonsampling errors into their analyses, as nonsampling error could impact the conclusions drawn from the results. Caution should also be used when interpreting results based on a relatively small number of cases.

Standard Errors and Their Use

A sample estimate and its standard error enable one to construct a confidence interval. A confidence interval is a range about a given estimate that has a specified probability of containing the average result of all possible samples. For example, if all possible samples were surveyed under essentially the same general conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then approximately 90 percent of the intervals from 1.645 standard errors below the estimate to 1.645 standard errors above the estimate would include the average result of all possible samples.

A particular confidence interval may or may not contain the average estimate derived from all possible samples, but one can say with the specified confidence that the interval includes the average estimate calculated from all possible samples.

Standard errors may also be used to perform hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The most common type of hypothesis is that the population parameters are different. An example of this would be comparing the percentage of adults in households where someone had a loss in employment income since March 13, 2020 from week 1 to week 2.

Tests may be performed at various levels of significance. A significance level is the probability of concluding that the characteristics are different when, in fact, they are the same. For example, to conclude that two characteristics are different at the 0.10 level of significance, the absolute value of the estimated difference between characteristics must be greater than or equal to 1.645 times the standard error of the difference.

The Census Bureau uses 90-percent confidence intervals and 0.10 levels of significance to determine statistical validity. Consult standard statistical textbooks for alternative criteria.

Estimating Standard Errors

The Census Bureau uses successive difference replication to estimate the standard errors of HPS estimates. These methods primarily measure the magnitude of sampling error. However, they do measure some effects of nonsampling error as well. They do not measure systematic biases in the data associated with nonsampling error. Bias is the average over all possible samples of the differences between the sample estimates and the true value.

Eighty replicate weights were created for the HPS. Using these replicate weights, the variance of an estimate (the standard error is the square root of the variance) can be calculated as follows:

$$Var(\hat{\theta}) = \frac{4}{80}\sum_{i=1}^{80}(\theta_i - \hat{\theta})^2 \quad (1)$$

where $\hat{\theta}$ is the estimate of the statistic of interest, such as a point estimate, ratio of domain means, regression coefficient, or log-odds ratio, using the weight for the full sample and θ_i are the replicate estimates of the same statistic using the replicate weights. See reference Judkins (1990).

Creating Replicate Estimates

Replicate estimates are created using each of the 80 weights independently to create 80 replicate estimates. For point estimates, multiply the replicate weights by the item of interest to create the 80 replicate estimates. You will use these replicate estimates in the formula (1) to calculate the total variance for the item of interest. For example, say that the item you are interested in is the difference in the number of people with a loss in employment income in week 1 compared to the number of people with a loss in employment income in week 2. You would create the difference of the two estimates using the sample weight, $\hat{x}_0$, and the 80 replicate differences, x_i, using the 80 replicate weights. You would then use these estimates in the formula to calculate the total variance for the difference in the number of people with a loss in employment income from week 1 to week 2.

$$Var(\hat{x}_0) = \frac{4}{80}\sum_{i=1}^{80}(x_i - \hat{x}_0)^2$$

Where x_i is the ith replicate estimate of the difference and $\hat{x}_0$ is the full estimate of the difference using the sample weight.

Users may want to pool estimates over multiple weeks by creating averages for estimates with small sample sizes. For pooled estimates, where two or more weeks of data are

combined to make one estimate for a longer time period, one would divide the unit-level weights that formed $\hat{x}_0$ and x_i (for each of the 80 replicate weights) for each week by the number of weeks that are combined. Then, form 80 replicate pooled estimates, $\hat{x}_{i,pooled}$ and the estimate, $\hat{x}_{0,pooled}$. Then use the pooled estimates in formula (1) to calculate the pooled variance for the item of interest.

Example for Variance of Regression Coefficients

Variances for regression coefficients β_0 can be calculated using formula (1) as well. By calculating the 80 replicate regression coefficients $\beta_i's$ for each replicate and plugging in the replicate β_i estimates and the β_0 estimate into the above formula,

$$Var(\hat{\beta}_0) = \frac{4}{80}\sum_{i=1}^{80}(\beta_i - \hat{\beta}_0)^2$$

gives the variance estimate for the regression coefficient β_0.

TECHNICAL ASSISTANCE

If you require assistance or additional information, please contact the Demographic Statistical Methods Division via e-mail at dsmd.source.and.accuracy@census.gov.

REFERENCES

Judkins, D. (1990) "Fay's Method for Variance Estimation," Journal of Official Statistics, Vol. 6, No. 3, 1990, pp.223-239.