EPOP
Understanding Entrepreneurship in the US

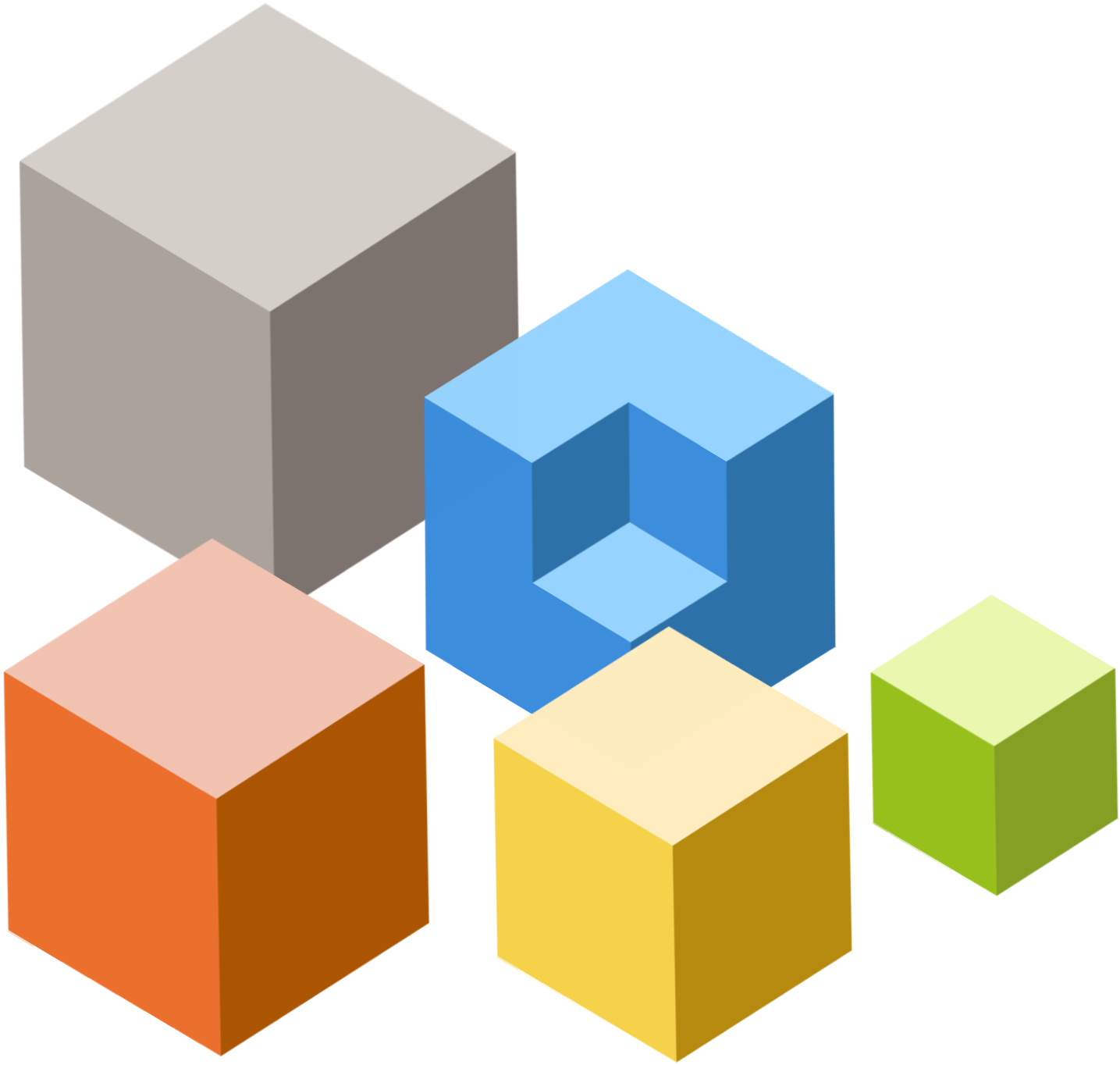
Entrepreneurship in the Population Survey

EPOP: 2022 Data User Guide

ISSUED: October 12, 2022
VERSION: 1

Created by
NORC at the University of Chicago
55 East Monroe Street, 30th Floor
Chicago, IL 60603
(312) 759-4000 Main
(312) 759-4004 Fax

Point of Contact
The NORC EPOP Research Team
EPOPresearch@norc.org



The Entrepreneurship in the Population Survey Project is being conducted by researchers at NORC at the University of Chicago with funding from a grant from the Ewing Marion Kauffman Foundation. Questions about this research project should be directed to EPOPresearch@norc.org.

The full title of the survey is "The Entrepreneurship in the Population Survey" and the abbreviation is EPOP Survey. In referencing the project or document, follow these standards:

Full Project Title:	The Entrepreneurship in the Population Survey Project: 2022
Project Abbreviation:	EPOP Survey
Survey Cycle Abbreviation:	EPOP:2022
Full Document Title:	Entrepreneurship in the Population Survey Data User Guide: 2022
Document Abbreviation:	EPOP:2022 Data User Guide
Citation:	Entrepreneurship in the Population Survey Data User Guide: 2022." NORC at the University of Chicago. October 12, 2022. https://EPOP.norc.org.

TABLE OF CONTENTS

1. Overview 1

About the EPOP Survey 1

Sponsor and Partner 1

About this Document 2

2. EPOP Survey Design and Response Rates 3

Target Population 3

Sample Design 3

Sample Sources 4

Data Collection and Response Rates 4

3. Survey Content 5

Screener 5

Employment Status Measurement 5

Job Type and Gig Work Measurement 5

Entrepreneurial Activities 6

Pathways and Prioritization 6

1 Current Business Owners 7

2 Current Freelancers 7

3 Nascent Entrepreneur 7

4 Former Business Owners 7

5 Former Freelancer 7

6 Withdrawn Entrepreneur 7

7 Non-Entrepreneur 8

Ancillary Questions by Entrepreneurship Categories 8

Pursuing Entrepreneurship Section 9

Business Ownership Operations Section 9

Non-Entrepreneur Section 9

Demographics 10

Primary 10
Secondary 10
Difference in Demographic Survey Administration by Sample Type 10
4. Working with EPOP Data Files 11
Administrative Variables 11
Location Variables 12
Weighting Variables 12
Data File Conventions 13
Variables Name 13
Reserve Codes 13
Data Protection 14
Weights 15
Development of Weights 15
How to Use Weights 17
Variance Estimation 18
Using the .csv data file 20
Using the STATA data file 20
Using the SAS data file 20
5. Reporting, Dissemination and Future Files 22
Abbreviations and Citations 22
EPOP Website: News and Publications 22
Where to find other Publications 23
Publish your analysis on the EPOP Website 23
Anticipated Data Release Schedule 23
References 24
Appendix 25
Appendix A: Entrepreneurship Pathways and Topical Areas A-1

1. Overview

About the EPOP Survey

The Entrepreneurship in the Population Survey, or EPOP Survey, was first conducted in 2022; four more annual collections are planned. The survey is designed to understand the scope of entrepreneurial activities from adults 18 years and older in United States and result in a variety of measures of entrepreneurial behavior including current and former business ownership, whether individuals are currently taking or have in the past taken steps towards starting a business, the extent to which individuals engage in freelance work, and engagement with the "gig economy." In addition to capturing the characteristic profile of the individuals involved in these various entrepreneurial activities across the U.S., the survey collects information on the behaviors, challenges, and resources available to individuals during the entrepreneurial process.

Information about the EPOP Survey methods, data availability, publications, and access to data user support may be found on the project's website: EPOP.norc.org.

EPOP Survey Manager and Conservator

NORC at the University of Chicago (NORC) is developing and conducting the EPOP Survey Project with grant funding from the Ewing Marion Kauffman Foundation. NORC is responsible for collecting, maintaining, disseminating, and safeguarding the resulting EPOP Survey data. For the project, NORC is both the manager of the enterprise and conservator of the resulting data.

NORC is an independent research institution that delivers reliable data and rigorous analysis to guide critical programmatic, business, and policy decisions. We conduct objective, non-partisan research to help inform people in government, nonprofits, and businesses making decisions on key issues of the day. Our research addresses important issues like employment, education, and health care. Since 1941, NORC has conducted groundbreaking studies, created and applied innovative methods and tools, and advanced principles of scientific integrity and collaboration. Today, government, corporate, and nonprofit clients around the world partner with NORC to transform increasingly complex information into useful knowledge. For more information, visit norc.org and connect with us via Twitter (twitter.com/norcnews) or Facebook (facebook.com/NORCatUofC.)

Sponsor and Partner

The Ewing Marion Kauffman Foundation is a private, nonpartisan foundation based in Kansas City, MO., that seeks to build inclusive prosperity through a prepared workforce and entrepreneur-focused economic development. The Foundation uses its $3 billion in assets to change conditions, address root causes, and break down systemic barriers so that all people –

regardless of race, gender, or geography – have the opportunity to achieve economic stability, mobility, and prosperity.

For more information, visit their website at Kauffman.org or connect with Kauffman via Twitter (twitter.com/kauffmanfdn) or Facebook (facebook.com/kauffmanfdn).

ABOUT THIS DOCUMENT

This document is designed to help data users analyze and better understand the EPOP:2022 Public Use File data. If there is methodology that is not covered in this document, a full methodology report for the 2022 survey may be found on EPOP.norc.org.

2. EPOP Survey Design and Response Rates

Target Population

The target population of the EPOP Survey includes noninstitutionalized adults 18 years or older in the United States.

The sample design supports the following estimation objectives:

- National estimates of entrepreneurial activity by demographics such as race/ethnicity, gender, age, and education, but not necessarily by the cross of these demographic variables,
- State-level estimates of entrepreneurial activity by race/ethnicity and gender, but not necessarily by the cross of these variables, and
- Metropolitan statistical area (MSA) level estimates of entrepreneurial activity for the top 50 MSAs by population[1] by race/ethnicity and gender, but not necessarily by the cross of these variables.

Sample Design

A stratified sampling design is used to achieve these objectives. Each state that does not contain a top 50 MSA constitutes a primary sampling stratum or a geography. For states that contain one or more of the top 50 MSAs, each MSA and the rest of state outside MSAs make a primary sampling stratum. For example, seven (7) strata or geographies are defined for the state of California, including the six (6) MSAs within the state plus the rest of the state. In addition, MSAs that are made up of counties from multiple states are divided into multiple primary sampling strata, one for each state. For example, Minneapolis-St. Paul-Bloomington, MN-WI, contains counties from both Minnesota and Wisconsin. One important objective of the study is to support estimation and analysis of entrepreneurship characteristics of underrepresented minorities, particularly Black and Hispanic individuals, within states and MSAs. Therefore, each primary stratum is further divided into three secondary sampling strata: Hispanic, non-Hispanic Black, and non-Hispanic Other.

[1] Top 50 MSAs are defined according to total population size in the 2020 decennial census. In the remainder of this document, “MSAs” refer to these top 50 MSAs.

Sample Sources

The study sample is selected from three frame sources:

1. NORC's AmeriSpeak® Panel,
2. An address-based sample (ABS) frame built from the United States Postal Service (USPS) Delivery Sequence File (DSF), and
3. Opt-in online survey panels.

Samples selected from the AmeriSpeak Panel and the ABS frame are probability samples with explicit stratification and known sample selection probabilities while the sample obtained from the opt-in online survey panels is a nonprobability sample with unknown frame coverage and unknown selection probabilities. Subsequent to data collection, the completed surveys from the three (3) samples are combined using NORC's TrueNorth® weighting method to generate a set of combined sample weights for estimation.

Data Collection and Response Rates

EPOP:2022 survey data collection began on February 15, 2022, for the AmeriSpeak sample, February 28, 2022 for the ABS sample, and May 13, 2022 for the opt-in online survey panel samples. Differential data collection protocols were followed for each of the sample types. After sending survey requests by USPS letter, email, and prompting calls, data collection ended on June 3 for the opt-in online survey panel, and on June 6, 2022 for both ABS and AmeriSpeak Panel samples. Data were primarily collected via an online survey; computer-assisted telephone interviewing was a secondary mode and available upon request. The survey was available in both English and Spanish. All participants were compensated for their participation.

The response rate varied by sample type. For the AmeriSpeak Panel sample, the response rate was 38.2%, and for the ABS sample, the response rate was 6.0%. For the opt-in survey panels, the response rate is not reported.

3. Survey Content

Screener

The screener section of the survey determines a respondent's working status (e.g., currently employed, retired, student, etc.) and, if working, their working arrangements and any potential entrepreneurial activities they might be engaged in. Through a multiple step process, the screener section identifies various possible entrepreneurial activities capturing current and former business ownership, current and former freelance/consultant/independent contracting work, and any current new business planning as well as situations where respondents were planning to start a business of some type at one point but withdrew from the planning process. Additionally, the screener was designed to capture flexible work arrangements provided by the gig economy.

Employment Status Measurement

The EPOP Survey first establishes the respondent's work status by asking, "*In the last week, did you work for pay at a job or business?*" Following the results of Abraham and Amaya (2019), the questionnaire also asks, "*In the last week, did you do ANY work for pay, even for as little as one hour?*" By asking this follow-up question of respondents who report they are not working, the survey ensures more informal work activities are captured and asked about which is important for determining an accurate measurement of gig work and the full suite of entrepreneurial activities.

Job Type and Gig Work Measurement

After establishing employment status, the EPOP survey collects key job information from those employed to construct a typology for entrepreneurs. This includes three main job types: (1) self-employed/business owner (respondents who select they either own their own business or are freelancers), (2) working for a for-profit or non-for-profit company, or (3) working for the government. This information is collected for both primary and secondary jobs.

Additionally, respondents are asked if the primary or secondary job is gig work. Given the potential for lack of clarity in what counts as gig work, the survey includes extensive examples of gig work activities and includes a definition of gig work in the main text of the question:

> "*Some people earn money through short, paid tasks or jobs online or in-person that are conducted through companies that coordinate payment for the service. This is sometimes referred to as 'gig work.'*"

A final question is asked to determine if respondents are engaged in gig work regardless of the prior responses about the primary and secondary jobs to ensure all gig work activities are reported even if that gig work is not the primary or secondary job.

Entrepreneurial Activities

Once job information has been captured, the EPOP Survey asks questions to gauge entrepreneurial activities directly. These survey items include a series of questions designed to determine if a respondent currently owns a business but does not work at it; has owned any sort of business enterprise in the past which is now closed; is planning a new business enterprise; or considered starting a business in the past, but ultimately withdrew from the enterprise.

1. **Former business ownership and freelancer**. Respondents are asked if they have ever owned a business or freelancer and if so when this activity stopped. Importantly, some respondents at this step report they currently own a business even though it is not reported as a current job. This likely reflects individuals who are currently passive business owners, and the business ownership is not considered a job. The results presented below combine these business owners with those who report business ownership as their primary or secondary job.

2. **Nascent entrepreneur**. To measure whether respondents are currently taking steps towards owning a business venture of any type, respondents are asked, "*Are you, alone or with others, currently trying to start a new business, including any form of self-employment, freelancing, consulting, or independent contracting, or selling any goods or services to others?*"

3. **Withdrawn entrepreneur.** Respondents are asked if they have ever considered starting a business, but withdrew from planning the enterprise: "*Have you, alone or with others, ever considered starting a new business, including any form of self-employment, freelancing, consulting, or independent contracting, or selling any goods or services to others but decided to wait or change your mind?*"

PATHWAYS AND PRIORITIZATION

The EPOP Survey screener is intentionally designed to capture the full range of entrepreneurial activities in which an individual might be engaged. As a result, some respondents qualify for multiple categories. To limit the burden on survey participants, each respondent is assigned to just one entrepreneurship category for follow-up survey questions using a priority order schema. That priority schema and description of each entrepreneurship category are presented here and shown in Table 1.

1 Current Business Owners

Respondents who report they currently own a business. Importantly, some respondents report that they currently own a business even though it was not reported as a current job. This likely reflects individuals who are currently passive business owners, but for whom the business ownership is not considered a job. The results presented below combine these business owners with those who report business ownership as one of their two primary jobs. Therefore, "current business owners" includes individuals who report that they still own a business even if it is not one of their two primary jobs.

2 Current Freelancers

Respondents who report they are currently freelancers, consultants, or independent contractors. Similar to the current business owner category, this category includes individuals who report that they are freelancers, consultants, or contractors even if they do not report their freelance work as one of their two primary jobs.

3 Nascent Entrepreneur

To measure whether respondents are currently taking steps towards owning a business, respondents are asked "*Are you, alone or with others, currently trying to start a new business, including any form of self-employment, freelancing, consulting, or independent contracting, or selling any goods or services to others?*" For the purposes of survey categorization, this classification does not condition on specific steps being taken towards entrepreneurship (such as Bennet and Chatterji (2019)), but this information is available in the EPOP Survey's follow questions. In this way, individual researchers can create measures suited to different definitions of nascent business development.

4 Former Business Owners

Respondents who answer they used to own a business but are no longer current business owners.

5 Former Freelancer

Respondents who report they were at one time a freelancer, consultant, or independent contractor but are no longer engaged in freelance work.

6 Withdrawn Entrepreneur

Respondents who answer yes to the following question regarding whether they have considered starting a business, "*Have you, alone or with others, ever considered starting a new business, including any form of self-employment, freelancing, consulting, or independent contracting, or selling any goods or services to others but decided to wait or change your mind?*"

7 Non-Entrepreneur

Respondents who are not engaged in any of the previous six entrepreneurial activities. These respondents receive "general population" questions.

Table 1. Assigned Entrepreneurship Categories by Reported Entrepreneurship Activities

Survey Pathway Priority	Assigned Entrepreneurship Category	Total Surveys	Reported Entrepreneurship Activity					
			Current Business Ownership	Current Freelancing	Entrepreneur-ship Planning	Former Business Ownership	Former Freelancing	Withdrawn Entrepreneur-ship Planning
1	Current Business Owner	4,907	4,907	3,189	2,892	0**	728	0
2	Current Freelancer	4,213	0	4,213	1,754	842	0**	0
3	Nascent Entrepreneur	1,467	0	0	1,467	257	274	0
4	Former Business Owner	3,030	0	0	136*	3,030	1,766	0
5	Former Freelancer	3,144	0	0	347*	0	3,144	0
6	Withdrawn Entrepreneur	2,649	0	0	0	0	0	2,649
7	Non-Entrepreneur	12,611	0	0	0	0	0	0
	Total	32,021	4,907	7,402	6,596	4,129	5,912	2,649

*These individuals worked as a freelancer, consultant, or independent contractor within the last 5 years of the survey administration date, although they report no longer working in that capacity of self-employment.

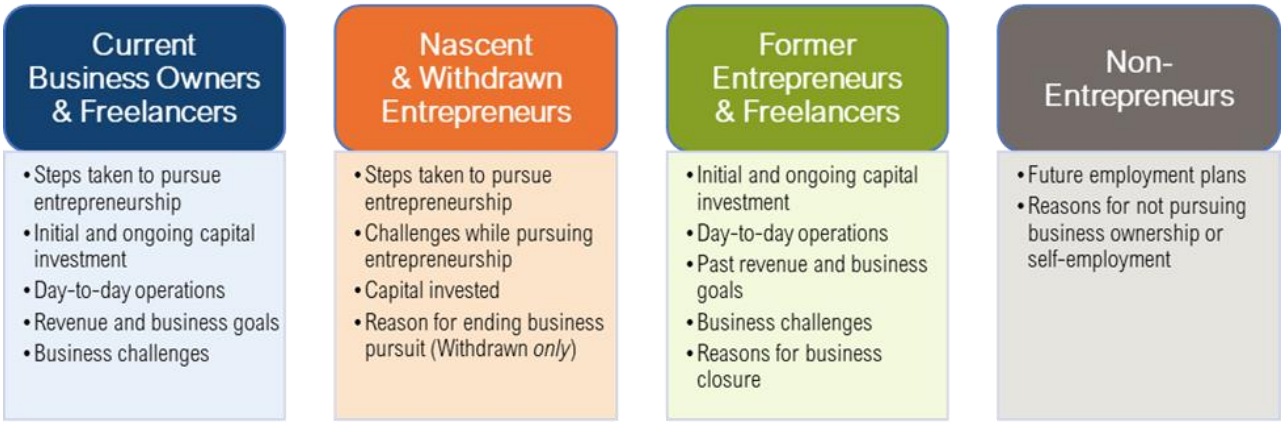
**There is no former business ownership activity reported for current business owners and no former freelancing or consulting activity for those categorized as current freelancers. This is a result of the survey construction as the Year 1 version of the EPOP Survey does not capture former entrepreneurship activity when an individual is currently engaged in the activity. The survey questionnaire will be revised for Year 2 to capture this type of former entrepreneurial activity for current business owners and freelancers.

Source: NORC, Entrepreneurship in the Population Survey: 2022.

Ancillary Questions by Entrepreneurship Categories

The focus of the EPOP Survey follow-up questions asked of each assigned entrepreneurship category is shown in Figure 1 and are briefly described below. Survey questions for current and former business owners and freelancers focus on concepts such as the operations and goals of the activities, whereas nascent and withdrawn entrepreneurs are asked more about concepts such as challenges starting a business. Non-entrepreneurs are asked more general questions about future employment plans as well as reasons for not pursing business ownership or self-employment.

Figure 1. EPOP Survey Content Focus by Assigned Entrepreneurship Categories



Pursuing Entrepreneurship Section

The questions in this section focus on the steps respondents took to pursue starting a business or working for themselves. The topics covered within this section are asked of current business owners and Freelancers as well as nascent and withdrawn entrepreneurs.

Business Ownership Operations Section

The questions in this section focus on the day-to-day operations of business owners and freelancers/consultants/independent contractors when their businesses were in full operation. For former business owners and freelancers, the questions pertained to the last year when their business was in operation. Specifically, the topics covered in this section include questions on when they started the business or self-employment and general descriptions of the type of business, legal status of the business, and how they came up with the idea for the business or self-employment.

For current and former business owners and freelancers, this section also asked for the types and amounts of additional financing they requested and/or received to continue the business or self-employment, the number and types of employees they used in their business, how much time they spent managing or working in their business, their revenue and profit/loss margin, their goals for the next five years, their biggest challenges facing their business (or former business), and their post-entrepreneurship plans and exit strategy.

Finally, all entrepreneurship categories were asked to indicate the industry that best classifies their current, former, or idea for a business.

Non-Entrepreneur Section

The questions in this section focus on respondents who did not have any prior experience with business ownership and were not currently taking steps to own a business or be self-employed. These respondents are asked to provide reasons why they have not considered entrepreneurship and to describe their current work arrangements including how long they've been in their current job, how many coworkers they have, how much they've worked in the past year, what type of benefits they receive with their work, and their likelihood of starting a new job or changing jobs in the future.

See Appendix A from a detailed graphic showing the question topics asked across all of the assigned entrepreneurship groups.

DEMOGRAPHICS

The EPOP Survey asks a comprehensive set of demographic questions. The demographic questions asked at the start of the survey and are critical for weighting are consider primary. The remaining demographic survey items are asked last and considered secondary.

Primary

The primary demographic variables that were asked of the ABS and nonprobability samples included age, ethnicity, and race. For the AmeriSpeak sample, panel member data on age, ethnicity, and race were already known and were not asked again as part of the EPOP Survey questionnaire.

Secondary

Secondary demographic variables included household income, student status, health insurance and benefits, debt amount, education, marital status, number of household members, number of children, military status, and citizenship status. The secondary demographic section also contained some additional attitude questions around the respondent's assessment of their community's economic outlook as well as lingering concerns related to the COVID-19 pandemic. These secondary demographic questions were asked of all respondents from all sample types.

Difference in Demographic Survey Administration by Sample Type

Data users should be aware that demographic variables were collected somewhat differently for the different sample types. The AmeriSpeak Panel members collected all primary variables and most secondary variables when the individual was originally recruited for the panel and the EPOP Survey did not re-ask those questions of those sample members. The ABS and opt-in online survey sample respondents were asked all demographic questions during EPOP survey administration. To mitigate inconsistency and potential bias across samples, the EPOP Survey aligned the demographic survey items to the AmeriSpeak Panel's version.

4. Working with EPOP Data Files

EPOP Survey data is available via both a restricted use data file (RUF) and a public use data file (PUF). Each data file includes variables for all survey questions in a format suitable for analysis and does not present a disclosure risk. Many variables in the RUF and PUF represent all response choices in the original survey questions. However, based on the need to protect respondent identity, question responses have been aggregated as necessary. In some cases, as with open-ended questions, variables were omitted entirely as the answers are unique to the respondent. The EPOP:2022 Questionnaire shows how the survey item was asked and the data file codebooks (described below) will show if any response options were aggregated for the analysis files. Together, these documents allow data users to clearly see where and how survey item aggregation was implemented.

The RUF provides more finely grained response options aligned more closely to the original response choices in the survey. To obtain the RUF, data file data users must undergo training on disclosure and publishing considerations and sign an agreement. (Please see the "Data Protection" section of this document for more information on the disclosure review process and disclosure considerations). The PUF presents some survey data items in broader categories. The PUF is available on the project website and does not require a formal data use agreement.

Administrative Variables

Both the RUF and PUF contain a set of administrative variables. Administrative variables include information relevant to how the survey was administered, how data was edited, or sample information. Table 2 shows the administrative variables included in the files. Note that only the RUF contains the EPOP Survey sample type variable (SAMP_TYPE). More information on these variables is provided in the data codebooks.

Table 2. EPOP Survey Data File Administrative Variables

Variable Name	Variable Label	RUF or PUF
R_SUID	RESPONDENT ID	Both
SAMP_TYPE	EPOP SAMPLE TYPE	Only RUF
PARTIAL_CASE	INDICTOR FOR PARTIAL COMPLETE CASES	Both
Q_LANGUAGE	SURVEY LANGUAGE	Both
DEM_AGE_IMPUTED	INDICATOR FOR IMPUTED AGE VALUES	Both

Location Variables

Both the RUF and the PUF include a nine-category census division variable and four-category census region, both derived from respondent's ZIP code. The RUF also includes a county and state variable. Note that many ZIP code cross multiple counties. For these ZIP codes, the case was coded to the county with the highest population based on 2021 Census estimates. The derived variable ZIP_TO_M_COUNTY is set to '1' for these cases. Table 3 shows the location variables.

Table 3. EPOP Survey Location Variables

Variable Name	Variable Label	RUF or PUF
CENSUS_DIV_DRV	CENSUS DIVISION DERIVED FROM ZIP CODE	Both
REGION_DRV	CENSUS REGION DERIVED FROM ZIP CODE	Both
COUNTY_DRV	COUNTY FIPS CODE DERIVED FROM ZIP CODE	Only RUF
ZIP_TO_M_COUNTY	INDICATOR REPORTED ZIP LINKS TO MULTIPLE COUNTIES	Only RUF
DEM_STATE	RESPONDENT STATE	Only RUF

Weighting Variables

In addition to the survey weight (WTSURVY), design variables (PSU, STRATA) are included to allow for calculating accurate standard errors (more information on the weighting process is in the "Weights" section of this document). The RUF also includes a weight constructed for only the probability sample (WTPROB). Table 4 shows the available survey weight variables.

Table 4. EPOP Survey Weight Variables

Variable Name	Variable Label	Included in:
WTSURVY	SURVEY WEIGHT: APPLIES TO ALL CASES	Both
WTPROB	PROBABILITY SAMPLE WEIGHT: FOR WORK WITH ONLY ABS AND AMERISPEAK SAMPLES	Only RUF
STRATA	SAMPLING STRATA	Both
PSU	PSU (PRIMARY SAMPLING UNIT)	Both

DATA FILE CONVENTIONS

Variables Name

In most instances, variable names within the RUF and PUF match each other and the variable names in the EPOP Survey Questionnaire. In instances when survey responses were aggregated to protect respondent confidentiality, variables names have been modified. For example, the original variable name in the questionnaire for the highest level of education is "DEM_EDU." Both the RUF and PUF require a different level of aggregation based on disclosure considerations. The variable recoded for the RUF is appended with "_RUF." The variable recoded for the PUF is appended with "_PUF." When a variable is recoded using the same level of aggregation for both the RUF and the PUF, the variable name is appended with "_DRV." For instance, the nine-category census division derived from ZIP code uses the same grouping for both the PUF and the RUF. This variable is named "CENSUS_DIV_DRV" in both files. Table 5 shows the variable name convention used to indicate which variables are modified for the RUF or PUF.

Table 5. EPOP Survey Data Variable Name Conventions indicating Aggregation

Variable Source	Name Convention	Example
Original Questionnaire Variable	No change	DEM_EDU
Aggregated for RUF	_RUF	DEM_EDU_RUF
Aggregated for PUF	_PUF	DEM_EDU_PUF
Aggregated for RUF and PUF	_DRV	CENSUS_DIV_DRV

Reserve Codes

When respondents skipped or refused questions or indicated they did not know the response to a question, the response is coded with a reserve code value. Similarly, data points that present a disclosure risk either in isolation or in combination with other data points are masked with a reserve code value. Table 6 shows the list of reserve codes used in the EPOP Survey data files.

Table 6. EPOP Survey Reserve Code Values

Reserve Code	Label
-3	Missing
-5	Don't' Know
-7	Suppressed

DATA PROTECTION

To protect the EPOP Survey data from allowing the potential re-identification of respondents, these four Statistical Disclosure Limitation (SDL) techniques have been used:

1. **Recoding.** Recoding can be used for both categorical and continuous variables. For categorical variables, it involves combining smaller categories into larger categories. It can also be used for continuous variables to code numbers into categories.
2. **Local Suppression.** Local suppression creates missing values to replace some the values.
3. **Rounding.** Rounding is applied to continuous variables like numbers to make the data harder to re-identify
4. **Mirco-aggregation.** Micro-aggregation clusters records into small groups and then the average is released as the value for some of the sensitive units

The variables that have undergone recoding, rounding, or micro-aggregation are indicated in the data files with variable names with the suffixes _PUF, _RUF, or _DRV (see Table 5). Variables that have undergone local suppression are in Table 7. All suppressed values in the PUF and RUF are indicated by -7 (see Table 6). For a full description of the various recoding schemas used in the PUF and RUF and the SDL methods used in the EPOP Survey data files, see the methodology report on EPOP.norc.org.

Table 7. EPOP Survey Restricted and Public Data File Suppression Count by Variable

Variable Name	Suppression Case Counts
DEM_AGE	62
RACE	95
DEM_GENDER	5
DEM_EDU	884
DEM_MARITAL	999
DEM_HHINC	359
DEM_MILITARY_1	199
DEM_STUDENT	76
BO_INDUSTRY1	519
BO_STARTBIZ_1	4
BO_NUMEMPLOY_1A-I	522
BO_WORKHOME_1	1
BO_PLMARGIN_1	476

Variable Name	Suppression Case Counts
BO_REVENUE_1	181
BO_REVENUE_2	152

WEIGHTS

Development of Weights

The EPOP Survey contains two sets of weights: probability sample weights for probability samples (AmeriSpeak and ABS samples), and combined sample weights for the combined probability and nonprobability samples. The combined sample weights are available in both the RUF and PUF, whereas the probability sample weights are available only in the RUF.

Creation of the probability sample weights follows these steps:

1. **AmeriSpeak sample base weights**. computed as the AmeriSpeak Panel weight divided by the probability of selection from the AmeriSpeak Panel to the study sample.

2. **ABS sample base weights.** computed as the inverse of the selection probabilities that account for both the first and second phase of the ABS sample selection.

3. **Adjustment for unknown eligibility**. this adjustment is applied to the ABS sample because the eligibility status of some sample addresses is not determined at the end of the survey. Through this adjustment, the weights assigned to know eligible cases are inflated to account for the eligible cases among the unknown eligibility cases. All AmeriSpeak samples are assumed to be eligible.

4. **Adjustment for interview nonresponse**. The weighting class method is used to adjust the weights for interview nonresponse. For the AmeriSpeak sample, adjustment cells are constructed by cross-classifying:

    a. geography (or primary sampling strata)
    b. race/ethnicity (Hispanic and Non-Hispanic Black, and Non-Hispanic Other)
    c. age (18-34, 35-64, 65 and older)
    d. education (Some college or less and bachelor's degree or above); and
    e. gender (Male and Female).

    With more limited data for the ABS sample, adjustment cells are defined by cross-classifying:

a. geography (or primary sampling strata)
b. race/ethnicity (Hispanic and Non-Hispanic Black, and Non-Hispanic Other)
c. gender (Male and Female)

5. **Combined interview nonresponse adjusted weights for probability sample.** The interview nonresponse adjusted weights computed for AmeriSpeak and ABS sample completes are then combined by geography where the composition factor is proportional to the number of completed interviews from each sample source. This combination is carried out such that the combined sample represents the target population for each geography.

6. **Raking to derive probability sample final weights**. Raking benchmarks are developed using the 2019 American Community Survey (ACS) 1-year estimates. Raking adjustments are conducted along these dimensions:

    a. Geography by Race and Ethnicity (Non-Hispanic White, Non-Hispanic Black, Hispanic, Non-Hispanic Other)
    b. Geography by Gender (Male, Female)
    c. Geography by Age (18-24, 25-29, 30-39, 40-49, 50-59, 60-64, 65+)
    d. Geography by Education (Less than High School, High School/GED, Some College, and BA and Above)
    e. Geography by Household income (< $25,000, $25,000-$49,999, $50,000-$74,999, $75,000-$99,999, $100,000+)

7. **Trimming.** Following these adjustments, extreme weights are trimmed within a given geography so that no weights are lower than 1 or greater than the median plus three times the interquartile range of the weights. This trimming reduces weight variability and increases the effective sample size.

8. **Re-raking.** The weights after trimming are re-raked to the population benchmarks by geography and race/ethnicity to ensure that:

    a. Weight variations remain low per geography, and
    b. The sum of weights by geography and race/ethnicity does not deviate from their respective population benchmark by more than 5% for each geography and 15% for each racial/ethnic-specific population benchmark (Non-Hispanic Black, Hispanic and Non-Hispanic All Other) within each geography.

Such deviations are allowed because raking does not necessarily converge due to the large number of raking dimensions. This trimming and raking process is repeated until the weight variation and alignment with benchmarks are considered satisfactory.

Weights for the combined probability and nonprobability samples are then developed using two further steps:

9. **Statistical matching.** Matching each nonprobability sample unit to a probability sample unit, which divides the probability sample into two sets: the set of units matched to the nonprobability sample and the set not matched.

10. **Matched propensity weighting for the nonprobability sample.** using the matched probability sample as the reference sample to estimate the inclusion probability of the nonprobability sample units in the combined sample and develop the pseudo weights for the nonprobability sample based on the estimated probabilities.

How to Use Weights

The final EPOP:2022 analysis data contains 32,021 respondents, including 11,174 respondents from the probability sample and 20,847 respondents from the nonprobability sample. Two provided weights can be used to generate unbiased estimates of the population, the combined weight (WTSURVY) and the probability sample weight (WTPROB). WTSURVY is used for generating estimates using the full sample which combines the probability and nonprobability samples. WTPROB is used for generating estimates based on only the probability sample.

Probability sample weights **(**WTPROB**)** are developed for AmeriSpeak and ABS samples to correct for potential bias due to unequal sample selection probabilities, nonresponse, and coverage errors. These weights can be used to produce unbiased national estimates, state and MSA level estimates, and estimates for other domains defined by the user. Any software package that can handle sample weights should produce correct weighted points estimates. The probability sample weights are available only in the restricted use file.

The combined sample weights (WTSURVY) are available in both the restricted use file and public use file and make analysis and reporting for many smaller domains possible due to its larger sample size. The combined sample weight can be used to produce approximately unbiased national estimates, state and MSA level estimates, and estimates for other domains defined by the user.

Variance Estimation

The EPOP Survey uses a complex sample design that needs to be accounted for in variance estimation. Otherwise, statistical software will underestimate standard errors of estimates. To facilitate variance estimation, we provide two design variables: PSU and STRATA.

For samples selected from the AmeriSpeak Panel, these variables are pseudo-PSUs and pseudo strata that are defined to represent the first stage PSUs and strata associated with the NORC National Frame that was used as the sampling frame for AmeriSpeak Panel recruitment sampling. The ABS and opt-in samples are not clustered, so each PSU is a single sample unit, and each STRATA is an MSA or the rest of a state outside MSAs. Using PSU and STRATA with either probability sample-only weights or combined weights will provide approximately unbiased variance estimates.

Standard variance estimation method can be used to approximate the variance of estimates based on the combined probability and nonprobability sample. Users who are interested in closer approximations, especially for small domains, may use the variance estimation method described in Yang, et al. (2022). NORC can provide additional information to support such implementation.

The sample code provided in Figure 2 shows examples of how variable AVAR can be analyzed using corrections for weighting and sample design in R and Stata. This example uses the combined probability and nonprobability samples with the corresponding weight variable, WTSURVY. SAS users can use PROC SURVEYFREQ and PROC SURVEYMEANS to calculate the design-corrected standard errors.

Figure 2. EPOP Survey Sample Stata and R Code

STATA

Load data	`use EPOP_YR1_PUF.dta,clear`
Set survey design	`svyset [pweight=WTSURVY], ///` `  strata(STRATA) psu(PSU) singleunit(scaled)`
Weighted mean	`svy: mean AVAR`
Weighted percentage	`svy: proportion AVAR`
Weighted total	`svy: total AVAR`
Weighted one-way table	`svy: tabulate AVAR`
Subset mean	`svy, subpop (if SUBGROUP==1): mean AVAR`
Specifying subgroups	`svy: mean AVAR, over(GROUPVAR)`

R

Install & load required packages	`install.packages(c("tidyverse", "survey"))`
	`library(haven)`
	`library(survey)`

Load data	`mydata <- read_dta("EPOP_YR1_PUF.dta")`
Set survey design	`mydesign <- svydesign(id = ~PSU, weights = ~WTSURVY, strata = ~STRATA, data = mydata, nest = TRUE)`
Singleton PSU correction	`options(survey.lonely.psu = "adjust")`
Weighted mean	`svymean(~AVAR, mydesign, na.rm = TRUE)`
Weighted total	`svytotal(~AVAR, mydesign, na.rm = TRUE)`
Weighted one-way table	`svytable(~AVAR, mydesign)`
Subset mean	`svymean(~AVAR, subset(mydesign, SUBGROUP == 1), na.rm = TRUE)`
Specifying subgroups	`svyby(~AVAR, by = ~GROUPVAR, mydesign, svymean)`

Statistical software may return errors when conducting variance estimation on subsamples and/or variables with a large number of observations with missing values. STRATA and PSU were created so that there was a minimum number of respondents within a STRAT/PSU cell. However, if all respondents within a cell are missing on a variable, it will be impossible to calculate the standard error. This is sometimes referred to as a "lonely PSU" or "singleton PSU." If the dataset is subset (to current entrepreneurs, for example), this error becomes more likely to happen. In these situations, you may receive an error such as this:

```
STATA error handling: "missing standard error because of stratum with
single sampling unit"
```

The best workaround to avoid this type of error is to manually combine the single-PSU stratum with a similar stratum. Alternatively, the sample code provided addresses the lonely PSU issue using automatic adjustments. In Stata, the correction is made with the svyset option "singleunit(scaled)" and in R, with the command "options (survey.lonely.psu = "adjust")." These methods of adjustment involve taking variance averages from stratum with multiple sampling units. Users should refer to their software documentation for more information on automatic adjustment methods before implementing them in their own research.

DATA FILE FORMATS

Both the RUF and PUF are available in three file formats: .csv, SAS, and STATA. Each file provides a different set of meta-data requiring different accompanying programs. As an example, the PUF package will include the data files and formatting programs shown in Table 8.

Table 8. EPOP Survey PUF Package File and Formatting Program Contents

File	Name
CSV data file	EPOP_YR1_PUF.csv
STATA data file with formatting applied	EPOP_YR1_PUF.dta
SAS data file	EPOP_YR1_PUF.sas7bdat
SAS program containing labels	EPOP_YR1_PUF_LABELS.sas
SAS program containing formats	EPOP_YR1_PUF_FORMATS.sas
SAS program to apply labels and formats	EPOP_YR1_PUF_APPLY_FORMATS_LABELS.sas

Using the .csv data file

Users of the .csv can refer to the variable names in the header column and review the data codebooks to retrieve variable format information.

Using the STATA data file

The STATA file provided is a formatted file. The file contains all labels and format information. STATA users can simply import that file and run frequencies to review variable formats and variable labels.

Using the SAS data file

Users of the SAS file may apply label and variable format information by applying the provided label and format definitions. To do this, open the provided SAS program 'EPOP_YR1_PUF_APPLY_FORMATS_LABELS.sas' and update the folder reference to the location where the data user has saved the EPOP data files and programs. Then run the program to apply the formatting information.

CODEBOOKS

A separate codebook is provided for the RUF and PUF. Each codebook includes an index of all variables included in the file. For each variable, a table is presented containing the variable name, variable label, original question text, and any survey skip logic. For most variables, a frequency table was the appropriate format to report answer choices. Each frequency table includes unweighted and weighted counts and percentages for each response choice and reserve code (i.e., -3-missing, -5-don't know, -7-suppressed). Continuous variables were reported as a table containing descriptive statistics: valid n, mean, median, min, max. Variables that were not continuous but contained many categories (e.g., the case identifier [R_SUID]) were reported in a frequency table where rows were grouped by valid and reserve code categories.

DATA USER SUPPORT

If you are having issues accessing the link for the PUF, specific files that were sent, or you have other questions about the EPOP Survey data or methods, please contact the EPOP research team at EPOPresearch@norc.org.

5. Reporting, Dissemination and Future Files

Abbreviations and Citations

The full title of the survey is "The Entrepreneurship in the Population Survey" and the abbreviation is EPOP Survey. In referencing a specific year, follow these standards:

Full Project Title: The Entrepreneurship in the Population Survey Project: 2022

Project Abbreviation: EPOP Survey

User Guide Citation: "Entrepreneurship in the Population Survey User Guide: 2022." NORC at the University of Chicago. October 12, 2022. EPOP.norc.org.

Data File Citations: "Entrepreneurship in the Population (EPOP) Survey Project Restricted Use Data File: <year>." NORC at the University of Chicago. <month> <day>, <year>. EPOP.norc.org.

"Entrepreneurship in the Population (EPOP) Survey Project Public Use Data File: <year>." NORC at the University of Chicago. <month> <day>, <year>. EPOP.norc.org.

EPOP Website: News and Publications

The EPOP Survey project website at EPOP.norc.org posts up-to-date news and information on research using the EPOP Survey data. As researchers use the data to write journal articles, research briefs, book chapters, presentations, and other products, the EPOP research team will post links to their publications on the website at their request.

Other individuals and organizations—in government, non-profit, and for-profit sectors—will also find EPOP data compelling. Ideally, they will use the data for a variety of purposes including policy action, advocacy, media releases, and proposals. Should changes to policies or programs be made based on the EPOP Survey data, the EPOP research team would appreciate being notified and will create a post about it, if permissible.

Where to find other Publications

As the EPOP Survey data gets analyzed by NORC and other researchers and mentioned by news media, a repository of EPOP related research, publications and media mentions will be available on the EPOP website: EPOP.norc.org.

Publish your analysis on the EPOP Website

The EPOP research team welcomes information on research using EPOP Survey data. Please contact EPOPresearch@norc.org if you have analyzed EPOP data and would like your research displayed on the website.

ANTICIPATED DATA RELEASE SCHEDULE

The overall timeline for implementing this project is July 1, 2021, to June 30, 2026. Table 9 shows the anticipated release of future year EPOP Survey data releases Public Use Files.

Table 9. EPOP Survey Future Data Release Schedule

Data Release	Anticipated Release Period
2022	October 2022
2023	August – September 2023
2024	August – September 2024
2025	August – September 2025
2026	August – September 2026

Any changes to these release periods will be posted on the EPOP website: EPOP.norc.org.

REFERENCES

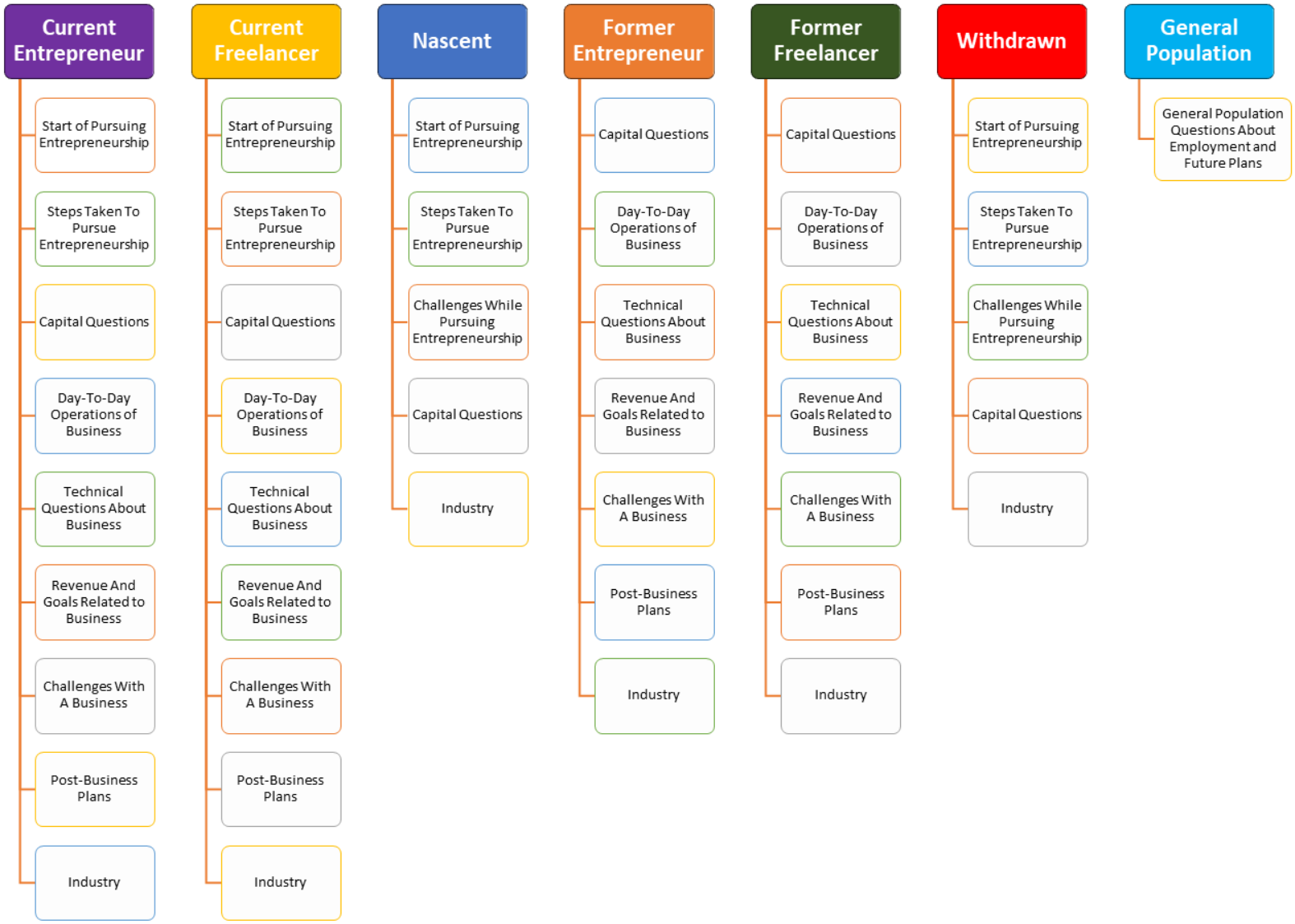
Abraham, K.G. and Amaya, A., (2019). Probing for Informal Work Activity. *Journal of Official Statistics*, 35(3), pp.487-508. https://doi.org/10.2478/jos-2019-0021.

Bennett, V. and Chatterji, R., (2019). The Entrepreneurial Process: Evidence from a Nationally Representative Survey. *Strategic Management Journal*, 23, pp.87-109. https://doi.org/10.1002/smj.307.

"Entrepreneurship in the Population (EPOP) Survey Project Questionnaire: 2022." NORC at the University of Chicago. October 12, 2022. EPOP.norc.org.

APPENDIX

APPENDIX A: ENTREPRENEURSHIP PATHWAYS AND TOPICAL AREAS



NORC at the University of Chicago

© Copyright 2022. NORC at the University of Chicago