User's Guide for the Paradata Data File and Codebook Documentation

Special file covering 2011–2015 of the National Survey of Family Growth

Introduction

The 2011-2015 NSFG paradata file includes data describing the NSFG data collection process, for fieldwork between September 2011 and September 2015. The sample includes housing units that were randomly sampled for the NSFG, including housing units that responded to a screener and/or main NSFG interview as well as housing units that ultimately did not respond to either interview. This file corresponds to the four years of data contained in the two 2-year public-use data files for interviews between 2011-2013 and interviews between 2013-2015. The file does not contain any data from the survey questionnaire, but includes a common identification variable for merging with the public use files. Types and sources of data included on this file are described in the paragraphs below. These data can be utilized to assess potential non-response bias as well as other methodological questions of interest. The Paradata File is intended as both a stand-alone data file and one that can also be linked with the NSFG public-use data files and restricted files.

The Paradata File includes several categories of data from several sources:

- Variables derived from the household roster (i.e., screening interview data)
- "Process paradata" (variables derived from interviewer-generated call records), for the screener and main interview attempts
- Process paradata indicating phase and outcome codes
- Interviewer observations of neighborhoods, housing units, contact with households, and sampled persons
- Interviewer characteristics
- Geographic identifiers from the U.S. Census Bureau and the National Institute of Standards and Technology (NIST)

The geographic identifiers from the U.S. Census Bureau and the NIST may be used to link the Paradata File to data from the U.S. Census Bureau.

There are 69,132 records in the Paradata File. This includes data from 59,179 screener interviews (including Eligible, Ineligible or Ineligible by proxy) and 20,621 main interviews (complete and partial interviews). The main interviews are a subset of the screener interviews - they only occur after a screener interview has taken place and an eligible person is identified. Of the 59,179 households screened, approximately half contained an eligible person. The 20,621 main interviews were completed with those eligible persons. The number of interviewers totaled 120 (114 of whom completed both screener and main interviews and 6 of whom only completed screener interviews).

The records remaining (69,132 minus 59,179 = 9,953) are households for which no screener interview (thus, no main interview) took place. This could occur for two reasons: 1) some households never responded to the request to complete a screening interview, and 2) some households who did not respond to the screener as of a certain point in the fieldwork were intentionally not selected for followup. This feature was part of "Phase 2" in a two-phase design. More details on this design are below (see: Screener and Main Interview: Phase and Outcome codes) and are available on the NSFG webpage, in the methodology documents describing the sample design and statistical features of each release [*link to "Design and Data*

Collection Methods" within the 2013-2015 page]. The inclusion of these cases with no interview data allows users to perform analyses of phase 1 outcomes.

Types and Sources of Data

Household Screening Interview Data

These are data collected at the time of household screening. The screening interview lists all household members, along with their age, sex, race, and ethnicity. These data are summarized into a number of variables about the household composition and characteristics of the sampled person. For instance, RSCRAGE is the age of the sampled person. While the same variables are available on the public use files for 2011-2013 and 2013-2015, they are also provided in the Paradata File so that they can be used in analyses of nonresponse patterns among screened households.

Two variables derived from the screener interview data are not available on the public use files -- ANYKIDS14 and NUMKIDS14 -- which are based on the age of each person in the screener household roster. Their relationship to the respondent is not considered.

Process Paradata

Screener and Main Interview: Attempts

These measures are summaries of data that are recorded at the call level and aggregated up to the case (address line) level. Attributes of individual call attempts are recorded by the interviewer at the time of attempt and stored in a sample management system called "SurveyTrak" at the call attempt level. The data are entered into SurveyTrak on laptops issued to each interviewer. These data are then replicated to a central database on a routine basis.

A key distinction is made between attempts that occur up to and including any household screening interview ("screener" attempts) and any attempts made after a screening interview has been completed and an eligible person has been sampled for an in-depth main interview ("main" attempts).

Sampled units for which a screening interview was not completed, or for which a screening interview resulted in a determination that there were no eligible persons within the household, will not have any main attempts, nor will other fields associated with the main attempts have values.

Many of the process paradata variables derived from call records describe whether a type of outcome was ever achieved for a case. For example, SCRN_BLCONTACTFLAG indicates whether the screening case was ever contacted. Other variables count the number of times a type of outcome was achieved. For example, SCRN_NUMNOCONT indicates the number of times that a "no contact" outcome was achieved for the screening phase of each case.

Screener and Main Interview: Phase and Outcome Codes

An important feature of the NSFG 2011-2019 data collection is the use of two-phase sampling. During each quarter, after 10 weeks, a subsample of the remaining active cases is selected. This subsample receives an altered data collection protocol (See NSFG methodology documents: *[link to 2013-2015*

page, "Design and Data Collection Methods"]. In order to allow analysts to examine the impact of this phased design, a variable denoting when the case was completed (SCRN_PHASE and MAIN_PHASE) was added. Further, the outcome codes at the end of the first phase (SCRN_OUTCOME_PH1 and MAIN_OUTCOME_PH1) have been included to facilitate comparisons across the phases.

The outcome codes include RC (result code), OUTCOME and OUTCOME_CATEGORY for each stage of interviewing (main and screening) and for each phase (phase 1 and the final version of each). RC is the 4-digit code stored in SurveyTrak as the result of each call attempt. OUTCOME is the description of the RC and also distinguishes whether the case was ever contacted. OUTCOME_CATEGORY is used to categorize the outcomes into the following broader categories: 'Eligible', 'Not eligible', 'Interview', 'Non-interview', 'Non-Sample', and 'Refusal'. Cases may not have been finalized by the end of phase 1. Therefore, an additional category, 'Interim,' is used for OUTCOME_CATEGORY_PH1.

The two-phase feature of the design has implications for the creation of weights and use of appropriate weights. The phase 2 weight was created to adjust for the probability of selection in phase 2 sampling. For final analyses (i.e., analyses of final outcomes for phase 1 and phase 2 data), analysts should exclude phase 2 sub-selected non-sample and apply the phase 2 weight to the analyses. For analyses of phase 1 outcomes, phase 2 sub- selected non-sample should also be included, since they would have been active during phase 1, and the phase 2 weight should not be applied.

Interviewer Observations

Neighborhood/Segment Observations

Segment observations are data on the sampled study areas that are recorded during the listing of housing units within the area segments. Every address within a segment will have the same value for a given segment observation. These are recorded prior to any call attempt being made in a given segment. For example, BLACCESS_GATED indicates whether the area segment has locked buildings or gated communities in it.

Housing Unit Observations

Housing unit observations are data on individual addresses that are recorded prior to a call attempt being made at the housing unit. These observations are attached to individual addresses. For example, STRUCTURE_TYPE identifies the housing unit as a single-family home, housing unit in a structure with 2- 9 units, housing unit in a structure with 10 or more units, a mobile home, or some other type of structure.

Contact Observations

Contact observations are recorded at each call attempt where contact occurs. These observations note anything that the household member might have said. There are two broad categories of utterances – questions and statements. Each type of utterance is counted. For example, SCRN_NUMANYQUEST is a count of the number of times a question was asked on any screening attempt involving contact with the household. Further, specific types of comments are also counted. Some statements are considered "positive." For example, "I'd be happy to help by completing your survey," is a positive statement. The number of positive statements is also counted.

Sampled Person Observations

Two observations are gathered after an eligible person is sampled to complete the main interview and before the main interview begins: SEXACTIVE and PROB_MAINIW. For SEXACTIVE, interviewers record whether they believe the sampled person is in an active sexual relationship with a person of the opposite sex (yes or no). PROB_MAINIW is the interviewer assessment of the chances of completing a main interview with the sampled person and it is recorded as 'high', 'medium' or 'low'.

Interviewer ID and Characteristics

Alpha-numeric identifiers are used for assigning cases to interviewers. Each unique ID represents an interviewer. (These interviewer IDs are numeric on the data file). An interviewer ID is given for each stage of interviewing: screening and main. SCRN_IWRID is the interviewer who completed the screening interview. MAIN_IWRID is the interviewer who completed the main interview. For the cases that were finalized as refusal for the screening interview, SCRN_IWRID is the interviewer who made the first refusal attempt during the screening interview stage. For the cases that were finalized as refusal for main interview, MAIN_IWRID is the interviewer who made the first refusal attempt during the main interview stage. For the rest of the cases, SCRN_IWRID is the interviewer who made the last attempt during the screening interview stage and MAIN_IWRID is the interviewer who made the last attempt during the main interview stage.

Variables on the data file include several characteristics of the interviewer including age, year of hire, education, and number of survey projects on which the interviewer worked. These characteristics are available for cases for which the interviewer completed a screening interview and the case was eligible for the main interview. The response categories for these questions were ranges of values rather than individual values.

Weights and Related Variables

Weights and Sample Design variables

In the 2011-2015 Paradata File, each case represents a different number of housing units in the U.S. due to unequal probabilities of selection for different race/ethnicity domains and for phase 2 sampling. The housing unit selection weight (HU_SELECTION_WGT_4YR) should be applied to all cases to yield estimates representative of all housing units in the U.S. One person was sampled from each eligible household and the person selection weight (PERSON_SELECTION_WGT_4YR) should be applied to all sampled persons to yield estimates representative of men and women 15-44 years of age in the U.S. The person selection weight is the product of the housing unit selection weight and the within household selection weight. Both of these weight variables (HU_SELECTION_WGT_4YR and PERSON_SELECTION_WGT_4YR) include a factor that accounts for the second phase sample selection procedure. The phase 2 sample selection weight (PHASE2_WGT) is included in the paradata to allow analysts to examine the impact of the two-phase design. If analysts want to examine phase 1 outcomes, then each of the selection weights should be modified in the following manner in order to remove the factor due to the second phase sampling:

$$HU\_SELECTION\_WGT\_4YR\_PH1 = \frac{HU\_SELECTION\_WEIGHT\_4YR}{PHASE2\_WGT}$$

$$PERSON\_SELECTION\_WGT\_4YR\_PH1 = \frac{PERSON\_SELECTION\_WEIGHT\_4YR}{PHASE2\_WGT}$$

Also included are detailed sample design variables not included on the Public Use files These variables are the Primary Sampling Unit (PSU) and Secondary Sampling Unit (Segment).

Census Geographic Identifiers

In the Paradata File we include the geographic variables necessary to link the paradata file to data from the U.S. Census Bureau. These include the State and County FIPS variables (source: NIST) and the Tract number and Block Group number (source: Census Bureau).

Other

A variable, QUARTER, is included for quarter that the sample case was released, with values ranging from 1 to 16.