# Codebook for the SIPP Synthetic Beta

This codebook documents version 5.0 of the SIPP Synthetic Beta (SSB). The SSB is a set of files containing individual-level data synthesized from linked survey and administrative data. The SSB is produced by the US Census Bureau as part of a joint project with the Social Security Administration (SSA), and the Internal Revenue Service (IRS). The goal of the project is to make some of the benefits of linked survey and administrative data available to researchers outside of restricted-access Census Bureau facilities in a manner that protects the confidentiality of the underlying data.

Creation of the SSB

The SSB is created from several data sources. The survey data are drawn from multiple panels of the Survey of Income and Program Participation (SIPP): the 1990, 1991, 1992, 1993, 1996, 2001, and 2004 Panels. The administrative data are drawn from the following SSA files: the Master Earnings File, the Master Beneficiary Records (MBR), the Supplemental Security Records (SSR), the 831 Disability File (F831), and the Payment History Update System (PHUS).

The creation of the SSB begins with the construction of the Gold Standard File (GSF). To construct the GSF, a set of variables from the 1990-2004 SIPP panels are standardized to produce consistent measures across panels. The SIPP respondent identifiers are mapped to Social Security Numbers (SSN) using the Census Bureau's Person Information Validation System (PVS). Using the list of SSN's for the sample of SIPP respondents, SSA creates Summary Earnings Records (SER) and Detailed Earnings Records (DER) extracts from the Master Earnings File. SSA also creates extracts from the four benefit files (MBR, SSR, F831, and PHUS) from the corresponding master files. Using the mapping between the SIPP identifiers and SSN's, Census then links these extracts to the SIPP data. The GSF consists of person-level research variables created from these linked data.

The next step in the creation of the SSB is to impute missing values in the GSF multiple times. This process results in four files (implicates) referred to as the Completed Data implicates. Each of these implicates contains original GSF values where non-missing and imputed values where the original value is missing. The imputations across Completed Data implicates are independent of each other.

The Completed Data implicates form the basis of the data synthesis that produces the SSB files. From each Completed Data file, four synthetic data sets are created by synthesizing variables conditional on the values in the Completed Data file. Thus, the SSB consists of sixteen files (implicates). All but the following data are synthesized in the SSB implicates: gender, OASDI benefit type, and spouse link (specific variables described in the data items section below). Detailed documentation of the process of data synthesis is available in the publication "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project" which can be downloaded from www.census.gov/sipp/synth_data.html.

The Completed Data and SSB implicates need not all have the same number of records. In order to be included in a Completed Data or SSB implicate, an individual's (possibly imputed or synthesized) age must be at least fifteen years as of January 1 in the first year of his or her SIPP panel. The interaction between this restriction and the variation in imputed and synthesized ages across implicates causes the exclusion of a slightly different set of individuals from each Completed Data and SSB implicate.

Using the SSB

The GSF and Completed Data implicates contain personally identifiable information protected by Titles 13, 26, and 42 and cannot be accessed without Census Bureau Special Sworn Status nor outside of Census Bureau facilities. The SSB files, however, have been cleared by the Census Bureau Disclosure Review Board, SSA, and IRS for use by individuals without Census Bureau Special Sworn Status and outside of Census Bureau facilities.

Researchers interested in using the SSB can submit an application to the Census Bureau. The application form and instructions can be downloaded from www.census.gov/sipp/synth_data.html. Applications will be judged solely on feasibility of the proposed project (i.e., that the necessary variables are available on the SSB). Once an application has been accepted, the new user will be given an account on a server where the data can be accessed and analyzed. While no SSB data downloads are permitted at this time, users do not have to operate behind the Census Bureau firewall to access this server.

The SSB is designed to be analytically valid in that sense that point estimates should be unbiased and estimated variances should lead to inferences similar to those that would be drawn from an identical analysis on the Completed Data implicates. Initial tests of analytic validity of the SSB have been promising. All SSB users are invited to help further test the analytic validity of the SSB by submitting programs used to analyze the SSB to be run on the Completed Data and/or Gold Standard files. Users need only inform Census Bureau staff of the location on the server of such programs and work with Census Bureau staff to ensure that the programs run without error. Census Bureau staff will run the programs on the confidential data and release to the user resulting output that are cleared for release by the Census Bureau Disclosure Review Board. In order to evaluate the effects of data synthesis separate from the effect of imputing missing data, comparisons should be made between results from the SSB and the Completed Data. To evaluate the effects of missing data imputation, comparisons should be made between results from the Completed Data and the Gold Standard.

When analyzing the SSB, users should account for the multiple imputation aspect of the SSB by averaging statistics of interests across all sixteen implicates. Variance measures should be created following the appropriate multiple imputation formulae as described in the document "Using the SIPP Synthetic Beta for Analysis" which can be downloaded www.census.gov/sipp/synth_data.html.

Description of Data Items

The remainder of this document lists and describes the data items contained in the (SAS format) SSB files. For each data item, the variable name, SAS label, variable type and length, source, range of values, and a description are provided. For data items for which the universe depends on another data item, the parent variable and the values that the parent variable must take in order for the data item to be in-scope are also listed.

Any questions related to the SSB can be directed to hhes.synthetic.data.use.list@census.gov.

## Identifiers

### PANEL

**Label:** SIPP Panel Year

Yes

**Values:** 1990-2004

**Description:** indicates panel of source record

### PERSONID

**Label:** Unique person identifier

Yes

**Values:**

**Description:** Across the Gold Standard and Completed Data files, personid uniquely identifies SIPP respondents. In the SSB, personid uniquely identifies records within a particular implicate. In order to strengthen confidentiality protection, personid in the SSB does not link records across implicates or to the Gold Standard and Completed Data files.

**SPOUSE_PERSONID**

**Label:** personid of spouse

Yes

**Values:**

**Description:** Personid of linked spouse. Across the Gold Standard and Completed Data files, spouse_personid uniquely identifies spouses of SIPP respondents. In the SSB, spouse_personid uniquely identifies records within a particular implicate. In order to strengthen confidentiality protection, spouse_personid in the SSB does not link records across implicates or to the Gold Standard and Completed Data files.

Linked spouse is defined as the first person to whom the SIPP respondent was married during the time period covered by the SIPP panel. Individuals could enter the panel already married and then each would be linked to the other. Individuals could also get married during the course of the panel. If this was the first observed marriage for each member of the couple, they were linked together. Individuals could also get divorced during the course of the panel and then remarry. In many cases, this later marriage caused a new individual to join the panel. This new SIPP respondent would only be linked to his or her spouse if the spouse (and original SIPP sample member) had not already been observed married to someone else. If the original SIPP sample member had been previously linked by marriage to another SIPP sample member, this original link was maintained in spouse_personid. However the marital history reflects the ending of this marriage and the occurrence of the next marriage for the original SIPP sample member. Likewise, the new SIPP sample member who joins through marriage will have that marriage date recorded in his or her marital history but will have a blank spouse_personid.

In summary, this variable captures only one marriage partner and does not provide a history of marriage partners even if this history is (partially) observed in the SIPP.

The link between SIPP respondents and their spouses has not been perturbed in any way in the SSB. The same individuals will be linked as married partners in the Gold Standard, the Completed Data, and the SSB.

## Demographic Variables

The variables in this section are all drawn from the SIPP and represent demographic information gathered by the survey at a specific point in time. Entries for individual variables describe the exact SIPP source variable and the reference point in time.

**MALE**

**Label:** Male

Yes

**Values:** 0/1

**Description:** 0=Female
1=Male
In the 1990-2004 Census-internal SIPP panels, a value for sex is included on each wave file. Thus, there are actually as many sex variables as there are waves of the survey and some changes occur across waves as a result of data collection error. Sex is selected from the array of variables sex1-sex{max number of waves} in which the wave corresponds either to the month in which marital status is first observed (for those without spouses during the course of the SIPP) or to the month in which the respondent's spouse is assigned instead of from a fixed point in the survey. Thus when a spouse is never assigned, an individual's gender comes from the first wave where they report being not married. For individuals who are assigned a spouse, gender comes from the first wave where they reveal their spouse. Finally, an indicator variable for males was created from the categorical sex variable for analytic convenience.

This variable is unsynthesized on the SSB and is never missing so there are no imputed values in the Completed Data.

## RACE

**Label:** Race

Yes

**Values:** 1 to 3

**Description:** 1=White
2=Black
3=Other
In the 1990-2004 Census-internal SIPP panels, a value for race is included on each wave file. Thus, there are actually as many race variables as there are waves of the survey and some changes occur across waves as a result of data collection error. Race is chosen by creating an array of variables race1-race{max number of waves} and choosing the first non-missing value. Thus race comes from the first wave in which the individual was interviewed instead of from a fixed point in the survey.

## BLACK

**Label:** Black/African American Race

Yes

**Values:** 0/1

**Description:** 0=Non-black
1=Black

## HISPANIC

**Label:** Hispanic

Yes

**Values:** 0/1

**Description:** 0=Non-Hispanic
1=Hispanic
In the 1990-1993 SIPP panels, a value for ethnicity is included on each wave file. Thus, there are actually as many ethnicity variables as there are waves of the survey and some changes occur across waves as a result of data collection error. Ethnicity is chosen by creating an array of variables ethncty1-ethncty{max number of waves} and choosing the first non-missing value. Thus, ethnicity comes from the first wave in which the individual was interviewed instead of from a fixed point in the survey. Respondents are coded as Hispanic if they have an ethnicity code between 14 and 20.
In the 1996-2004 panels, the longitudinally-edited version contains only one value for ethnicity across all waves (eorigin) and this value is used. Respondents are coded as Hispanic if they have an ethnicity code between 20 and 28 in 1996 and 2001, or if they have an ethnicity code of 1 in 2004.

## FOREIGN_BORN

**Label:** Foreign Born

Yes

**Values:** 0/1

**Description:** Immigrant Status, born in country other than U.S.
Taken from wave 2 topical module (TM8730, TM8734, TM8709 1990-1993 panels; eprstate, ebrstate and rcitiznt 1996 panel; eprstate, ebrstate and tcitiznt 2001 panel; eprstate, ebrstate, citiz, and ebornus 2004 panel)
0=Born in U.S.
1=Born in country other than U.S.

### TIME_ARRIVE_USA

**Label:** Decade of Arrival to US (Foreign Born)

Yes

**Values:** 1 to 10

**Description:** Decade arrive in U.S. (answered when SIPP respondent was foreign_born)
The year of arrival to the U.S. is from the Census-internal SIPP files (TM8736 1990-1993 panels; rmoveus 1996 panel; tmoveus 2001-2004 panels)
.=Structurally missing, out of scope for question (foreign_born=0)
1=Before 1959
2=1960 - 1964
3=1965 - 1969
4=1970 - 1974
5=1975 - 1979
6=1980 - 1981
7=1982 - 1984
8=1985 - 1993
9=1994 - 1999
10=2000 - 2004

## Disability Variables

### SUM_DISAB_IN_SCOPE

**Label:** In-Scope for Disability (Sum of Core and TM)

Yes

**Values:** 0/1

**Description:** In scope to be asked questions about whether health limits the type or amount of work one can do
0=No
1=Yes
For the 1996-2004 panels, information on disability comes from core questions (edisabl and edisprev) during wave 2 for people ages 15-69, when respondents were asked both whether health limited and prevented the type or amount of work. These indicators are supplemented with details from the Functional Limitations and Disability topical module (wave 5 for 1996-2004 panels, variables ejobdif and ejobcant) that covers people ages 16-67. For the 1990-1993 panels, disability information comes from the core question (disab) during wave 2 for people ages 15-69, when respondents were asked whether health limited the type or amount of work. This information is supplemented with details from the Functional Limitations and Disability topical module (waves 3, 3, 6, 3 for 1990-1993 panels, variables TM8914, TM8918, TM8920, TM8922, and TM8924) that covers people ages 16-67. The core and topical modules were used in conjunction to construct summary measures of disability. In order to make the two disability questions consistent for all panels, both responses were taken from the core and topical modules, when available. If the respondent was interviewed during the appropriate wave and provided a response to either the core or topical module question, then the individual was in scope. If the respondent was not interviewed during the core or topical module, respondent age at the time of the core and topical module waves was calculated and the person was judged to be in scope or not based on their age at the time the question should have been asked of them. Thus, a positive flag for the variable sum_disab_in_scope indicates that the person was in scope to answer at least one of the core or topical module questionnaires.

**SUM_DISAB**

**Label:** Disability (Sum of Core and TM)

Yes

**Values:** 0/1

**Description:** Health limits kind or amount of work

For the 1996-2004 panels, information on work-limiting disability comes from core question (edisabl) during wave 2 for people ages 15-69, when respondents were asked whether health limited the type or amount of work. This indicator is supplemented with details from the Functional Limitations and Disability topical module (wave 5 for 1996-2004 panels, variable ejobdif) that covers people ages 16-67. For the 1990-1993 panels, disability information comes from the core question (disab) during wave 2 for people ages 15-69, when respondents were asked whether health limited the type or amount of work. This information is supplemented with details from the Functional Limitations and Disability topical module (waves 3, 3, 6, 3 for 1990-1993 panels, variables TM8914, TM8918, and TM8920) that covers people ages 16-67. In order to make the sum_disab consistent for all panels, both responses were taken from the core and topical modules, when available, with any positive indication of health limiting the kind or amount of work flagging the positive response.

.=Structurally missing, out of scope for question (sum_disab_in_scope=0)
0=No (sum_disab_in_scope=1)
1=Yes (sum_disab_in_scope=1)

**SUM_DISAB_NOWORK_IN_SCOPE**

**Label:** In-Scope for Disab Prev Work (Sum of Core and TM)

Yes

**Values:** 0/1

**Description:** In scope to be asked questions about whether health prevents work

0=No
1=Yes

**SUM_DISAB_NOWORK**

**Label:** Disability Prevents Work (Sum of Core and TM)

Yes

**Values:** 0/1

**Description:** Health prevents work

For the 1996-2004 panels, information on work-preventing disability comes from core question edisprev during wave 2 for people ages 15-69, when respondents were asked whether health prevented work. This indicator is supplemented with details from the Functional Limitations and Disability topical module (wave 5 for 1996-2004 panels, variable ejobcant) that covers people ages 16-67. For the 1990-1993 panels, the core questionnaire does not ask respondents whether health prevents work. This information is solely obtained from the Functional Limitations and Disability topical module (waves 3, 3, 6, 3 for 1990-1993 panels, variables TM8922 and TM8924) that covers people ages 16-67. When available, the core and topical modules were used in conjunction to construct summary measures of disability, with any positive indication of health preventing work flagging the positive response.

.=Structurally missing, out of scope for question (sum_disab_in_scope=0 or { sum_disab_in_scope=1, sum_disab=0})
0=No (sum_disab=1)
1=Yes (sum_disab=1)

## Education Variables

### YEAR_BEG_POSTHS

**Label:** Year Began Post-HS Education

Yes

**Values:**

**Description:** The wave 2 Education and Training History topical module provides knowledge of the year that post-high school education began (variable TM8420 for 1990-1993 panels; variable tcollstr for 1996-2004 panels).
.=Structurally missing (educ_5cat=1 or educ_5cat=2)
=Range

### CURRENT_ENROLL_COLL

**Label:** Flag currently enrolled collage

Yes

**Values:** 0/1

**Description:** indicates whether an individual is enrolled in college at time of SIPP education history topical module and has not finished his/her education.  This variable can be used to differentiate between individuals who completed some college and stopped school and those who have finished some college but not yet stopped attending school.

### YEAR_BACH

**Label:** Year of Bachelor Degree

Yes

**Description:** year bachelor's degree was finished

**FIELD_BACH**

**Label:** Field of Bachelor Degree

Yes

**Values:** 1-20

**Description:** Field in which bachelor's degree was obtained.  Taken from topical history module on education history.

Catgories for 1996-2004 panels

1. Agriculture/Forestry
2. Art/Architecture
3. Business/Management
4. Communications
5. Computer and Information Sciences
6. Education
7. Engineering
8. English/Literature
9. Foreign Language
10. Health Sciences
11. Liberal Arts/Humanities
12. Math/Statistics
13. Nature Sciences (Bilogical and Physical)
14. Philosophy/Religion/Theology
15. Pre-Professional
16. Psychology
17. Social Sciences/History
18. Other

Categories for 1990-1993 panels

1. Agriculture/forestry
2. Biology
3. Business/management
4. Economics
5. Education
6. Engineering (including computers and computing)
7. English/journalism
8. Home eonomics
9. Law
10. Liberal arts/humanities (including arts, architecture, music, languages, philosophy)
11. Mathematics or statistics
12. Medicine
13. Nursing, pharmacy, health technologies
14. Physical or earth sciences
15. Police science or law enforcement
16. Psychology
17. Religion/theology
18. Social Sciences (history, sociology, political science)
19. Vocational or technical studies
20. Other

**YEAR_END_POSTHS**

**Label:** Year Ended Post-HS Education

Yes

**Values:**

**Description:** The wave 2 Education and Training History topical module provides knowledge of the year that post-high school education ended (variables TM8426 and TM8440 for 1990-1993 panels; variables tlastcol, tvocyr, tassocyr, tbachyr, and tadvncyr for 1996-2004 panels).

.=Structurally missing (educ_5cat=1 or educ_5cat=2)

=Range

### EDUC_5CAT

**Label:** Education Category (5)

Yes

**Values:** 1 to 5

**Description:** Highest level of education attained at the time of the education history topical module.
1=No high school degree
2=High school degree
3=Some college
4=College degree
5=Graduate degree
This variable was created from information gathered in the topical module on education history and represents the highest level of education achieved up to the point of the administration of the topical module questions.

For individuals who did not answer the topical module education history questions, we created a highest level of education variable from the SIPP core and used this variable as a predictor to impute the education variable based on the topical module. See description of educ_cat for details on how the SIPP core education variable was created.

### CURRENT_ENROLL_HS

**Label:** Flag currently enrolled high school

Yes

**Values:** 0/1

**Description:** indicates whether individual is still enrolled in high school (or lower grade) and has not completed education at time of SIPP education history topical module

### YEAR_END_HS

**Label:** Year Ended HS (or less) Education

Yes

**Values:**

**Description:** The wave 2 Education and Training History topical module provides knowledge of the year that high school was last attended (variables TM8404 and TM8412 for 1990-1993 panels; variables tlstschl and thsyr for 1996-2004 panels).
.=Structurally missing (educ_5cat =1 and current_enroll_hs=1)
=Range

## Economic Variables

The variables in this section are all drawn from the SIPP and represent economic information gathered by the survey at a specific point in time. Entries for individual variables describe the exact SIPP source variable and the reference point in time.

### TOTNETWORTH

**Label:** Total Net Worth

Yes

**Values:** ge 0

**Description:** total net worth

### PENSION_IN_SCOPE_EMPL

**Label:** In-Scope for Pension (Level II)

Yes

**Values:** 0/1

**Description:** Individual must have been employed at time of pension topical module in order to answer the pension questions.

### DC_PENSION

**Label:** Defined Contribution Pension Plan
Yes
**Values:** 0/1
**Description:** 1=had defined contribution pension plan
0=no defined contribution pension plan

### DB_PENSION

**Label:** Defined Benefit Pension Plan
Yes
**Values:** 0/1
**Description:** 1=had defined benefit pension plan
0=no defined benefit pension plan

### OWN_HOME

**Label:** Own a Home
Yes
**Values:** 0/1
**Description:** 1=own a home
0=do not own a home

### HOMEEQUITY

**Label:** Home Equity
Yes
**Values:** pos/neg
**Description:** Self-reported home equity value

### NONHOUSWEALTH

**Label:** Non-Housing Financial Wealth
Yes
**Values:** pos/neg
**Description:** Non-housing wealth = total wealth minus home equity

### IND_EXIST

**Label:** Flag: Industry Assigned
Yes
**Values:** 0/1
**Description:** Does person have valid industry from a job held during survey?
0=No, last worked 1984 or earlier, or no valid industry reported
1=Yes

**IND_4CAT**

**Label:** Industry Category (4)

Yes

**Values:** 1 to 4

**Description:** 1=Manufacturing
2=Wholesale/retail trade
3=FIRE, services, public administration, military
4=Agriculture, mining, construction, transportation, communications, and public utilities
Industry is a characteristic of an individual's job and hence varies over time. There are industry values reported for (potentially) two jobs in each wave of the survey. Industry is chosen by summing earnings associated with the array of variables ws1ind1-ws1ind{max number of waves} and ws2ind1-ws2ind{max number of waves} in the 1990-1993 panels, and ejbind1_1-ejbind1_{max number of waves) and ejbind2_1-ejbind2_{max number of waves} in the 1996-2004 panels and choosing the industry associated with the greatest total earnings. Thus industry is the industry from which greatest earnings are derived in the survey.

**OCC_EXIST**

**Label:** Flag: Occupation Assigned

Yes

**Values:** 0/1

**Description:** Does person have valid occupation from a job held during survey?
0=No, last worked 1984 or earlier, or no valid industry reported
1=Yes

**OCC_3CAT**

**Label:** Occupation Category (3)

Yes

**Values:** 1 to 3

**Description:** 1=Managerial and professional specialty occupations
2=Technical, sales, and administrative support occupations
3=Other
Occupation is a characteristic of an individual's job and hence varies over time. There are occupation values reported for (potentially) two jobs in each wave of the survey. Occupation is chosen by summing earnings associated with the array of variables ws1occ1-ws1occ{max number of waves} and ws2occ1-ws2occ{max number of waves} in the 1990-1993 panels, and tjbocc1_1-tjbocc1_{max number of waves) and tjbocc2_1-tjbocc2_{max number of waves} in the 1996-2004 panels and choosing the occupation associated with the greatest total earnings. Thus occupation is the occupation from which greatest earnings are derived in the survey.

## Marital History Variables

Marital history is presented as two arrays of 8 elements describing up to 4 marriages. This history retains at most 4 dates of origin and the associated 4 dates of dissolution (whether due to divorce or death), if applicable, along with the corresponding type of event (marriage, divorce, or widowhood) for each SIPP respondent. The wave 2 Marital History topical module provides the majority of this information for up to 3 marriages. If an individual had more than 3 marriages, no dates for those marriages between the second and most recent are collected during the topical module interview.

For individuals who participate in the topical module, we supplement this information by searching for new marriages or the termination of an existing marriage utilizing our knowledge of monthly marital status covering the period of the panel beyond wave 2. We rely exclusively on the complete set of monthly marital status indicators for people who do not participate in the topical module.

The two marital history arrays, one comprised of dates and the other of event types, are edited to ensure internal consistency for linked spouses. Missing event dates or reasons are acquired from the spouse who has provided this information during a SIPP interview. This means that in the case of a deceased individual, the surviving spouse's report of widowhood is transferred to their former spouse. Likewise, for respondents who leave the household due to divorce (and are no longer interviewed by the SIPP), the spouse remaining in the household supplies the details regarding marital dissolution for both.

It is possible for beginning or ending date information to not identically match for the linked spouses. In this case, we evaluate whether topical module details were supplied. If both spouses participated in the topical module, the data are considered as likely to be reliable from either individual. We examine whether the first spouse's beginning date occurs before the previous marriage end date for either spouse while the second spouse's beginning date occurs after the previous marriage end date of both spouses. If so, then the second spouse's beginning date replaces the first spouse's beginning date. Alternatively, if the reverse is true, then the first spouse's beginning date replaces the second spouse's beginning date. When no obvious conflicts with the date of the start of the current marriage and the date of termination of the previous marriages of either spouse exist, then a random number is used to determine which spouse's information to retain. A similar algorithm is implemented to resolve issues relating to non-matching ending dates of the linked spouses (in this case, each spouse's subsequent marital start date is taken into consideration).

When disagreements between the beginning or ending dates occur and only one of the two spouses participated in the topical module, the participating spouse's information is considered to be more reliable (provided that the adoption of the date presents no conflicts with previous or subsequent marital events for either spouse). In the absence of topical module participation for both spouses, a random number is used to determine which spouse's information to retain. Any persisting conflicts between dates of marital events are remedied by utilizing a random number to determine the most reasonable event date for the pair. As a final step, the cleaned file is carefully reviewed once more to ensure internal consistency.

**MH1**

**Label:** Flag: Marital History Event 1

Yes

**Values:** 0/1

**Description:** First marital history event flag
0=never married
1=first marriage occurred

**FLAG_MAR4T**

**Label:** Flag: 4 or More Marriages

Yes

**Values:** 0/1

**Description:** Flag for existence of a marriage for which date is unknown because it was not collected in the SIPP.
1=An additional marriage occurred but with unknown date
0=No additional marriage with unknown date
The marital history topical module asks about a person's first and second marriages and then his or her most recent marriage. If any other marriages occurred after the second but before the most recent, no information about this marriage is collected. However, individuals are categorized as having 1, 2, 3, or more than 3 marriages. We create flag_mar4t to identify individuals who reported more than 3 marriages.

**MH2**

**Label:** Flag: Marital History Event 2

Yes

**Values:** 0 to 2

**Description:** 0=first marriage did not end over course of survey
1=first marriage ended in widowhood
2=first marriage ended in divorce

**OBS_FIRST_SIPP_MAR_NUM**

**Label:** Ordinal Number of First Obs Marriage

Yes

**Values:** 1 to 4

**Description:**

**SIPP_PANEL_END_DATE**

**Label:** SIPP Panel End Date

Yes

**Values:**

**Description:**

**MH3**

**Label:** Third marital history event flag

Yes

**Values:** 0,1

**Description:** 0=no second marriage
1=second marriage occurred

**MH4**

**Label:** Fourth marital history event flag

Yes

**Values:** 0,1,2

**Description:** 0=second marriage did not end over course of survey
1=second marriage ended in widowhood
2=second marriage ended in divorce/separation

**MH5**

**Label:** Fifth marital event flag

Yes

**Values:** 0,1

**Description:** 0=no third marriage
1=third marriage occurred

**MH6**

**Label:** Sixth marital history event flag
Yes
**Values:** 0,1,2
**Description:** 0=third marriage did not end over course of survey
1=third marriage ended in widowhood
2=third marriage ended in divorce/separation

**MH7**

**Label:** Seventh marital event flag
Yes
**Values:** 0,1
**Description:** 0=no fourth marriage
1=fourth marriage occurred

**MH8**

**Label:** Eighth marital history event flag
Yes
**Values:** 0,1,2
**Description:** 0=fourth marriage did not end over course of survey
1=fourth marriage ended in widowhood
2=fourth marriage ended in divorce/separation

**MH_DATE**

**Label:** Date of Marital History Event <n>
Yes
**Values:**
**Description:** SAS date value

## Fertility Variables

Number of children and dates of birth.

**OWN_KIDS_EVER**

**Label:** Number of Children Ever Born
Yes
**Values:** ge 0
**Description:** Number of children ever born
This is taken from the wave 2 Fertility history topical module (TM8752 and TM8754 for 1990-1993 panels; tfrchl and tmomchl for 1996-2004 panels).
0=No children ever born
=1-6

**FIRST_BIRTH_YEAR**

**Label:** Year of Birth of First Child

Yes

**Values:**

**Description:** Year of birth of first child

This is taken from the wave 2 Fertility history topical module (TM8762 and TM8794 for 1990-1993 panels; tfbrthyr for 1996-2004 panels).

.=Structurally missing (own_kids_ever=0)

=Range

**LAST_BIRTH_YEAR**

**Label:** Year of Birth of Last Child

Yes

**Values:**

**Description:** Year of birth of last child

This is taken from the wave 2 Fertility history topical module (TM8768 and TM8782 for 1990-1993 panels; tlbirtyr for 1996-2004 panels).

.=Structurally missing (own_kids_ever=0)

=Range

**TOTFAM_KIDS**

**Label:** Total Number of Children in Family

Yes

**Values:** 0 to 30

**Description:** Number of children under the age of 18 that live in a family in the interview month in which marital status is first observed (for those without spouses during the course of the SIPP) or in which the respondent's spouse is assigned This number is the same for all family members and does not indicate that the children are related to a particular individual (fnkids for 1990-1993 panels, rfnkids for 1996-2004 panels).

0=No children under age 18 live with the family

=1-12

## Lifespan Variables

Birth and Death dates.

### BIRTHDATE

**Label:** Date of Birth

Yes

**Values:**

**Description:** This variable was taken from a hierarchy of SSA sources instead of the respondent-provided value in the SIPP. Date of birth was selected from the first non-missing value in the following files: (i) SSA's Master Benefits Record (MBR) file, (ii) the Census Bureau's Person Characteristic File (PCF) whose main input is the SSA Numident file, and (iii) SSA's Supplemental Security Record (SSR) file. Thus, this variable is administrative and sometimes differs from the birth date reported in the SIPP survey itself. When missing due to the lack of a validated SSN for the SIPP respondent, date of birth was imputed using date of birth from the Census-internal version of the SIPP.

We chose the administrative source for two reasons. First, the administrative birth date was more often consistent with the other administrative data (benefits and earnings). For example, when age was calculated using the administrative birth date, there were fewer individuals who appeared to retire before age 62. Second, the differences between the administrative birth date and the birth date reported in the survey helped to increase the difficulty of re-identifying a record in the original SIPP public use data from a record in the synthetic data, thus improving the confidentiality protections. This variable is coded as a SAS date variable. This format gives the number of days between the date of birth and January 1, 1960. An individual born on January 1, 1959 would have birthdate=-365 and an individual born on January 1, 1961 would have birthdate=365.

In addition to help researchers we have created the variables:
brthdy_final
brthmn_final
brthyr_final
These variables are numeric, length 3 and contain the day, month, and year respectively of the birthdate.

### FLAG_DEATHDATE_EXIST

**Label:** Flag: Existence of Date of Death

Yes

**Values:** 0/1

**Description:** Flag to indicate that this respondent died after being interviewed and before 2006.
0=Death date does not exist, respondent did not die during this interval
1=Death date exists, respondent died during this interval

### DEATHDATE

**Label:** Date of Death

Yes

**Values:**

**Description:** Date of death from administrative data. This variable also is obtained using a hierarchy of administrative sources: (i) SSA's MBR file, (ii) the Census PCF with death information coming from the SSA Numident and Master Death Files, and (iii) SSA's SSR file. This variable is coded as a SAS date variable.

In addition to help researchers we include:
deathdy_final
deathmn_final
deathyr_final
which are numeric 3 and contain the day, month, and year, respectively, of the date of death.

## Benefits Variables

Social Security benefits - Old Age, Survivor, and Disability Insurance (OASDI) and Supplemental Security Income (SSI)

### MBR/PHUS Variables

**The Master Benefits Records (MBR) is SSA's main file to track who is receiving Old Age Survivor and Disability (OASDI) benefits, the reason for receipt, and the monthly benefit amounts payable to the individual.**

**The Payment History Update System (PHUS) contains actual payments delivered to OASDI beneficiaries. The data from the PHUS may differ from what are contained on the MBR due to discrepancies between the timing of SSA awarded amounts and the actual payments made to participants. This situation would be expected to affect disability cases more than aged cases because it takes more time to establish eligibility to receive disability.**

**Individuals are eligible to receive benefits due to their own earnings history and age, as well as due to a spouse's earnings history and age. In this section retirement and disability are "own" benefits while aged spouse, widowed spouse, and other are "spouse" benefits. The age requirements for receiving each type of benefit are as follows:**
**Retire - minimum age 62 (reduced benefit), full retirement age (full benefit)**
**Disability - under age 65 or full retirement age, whichever is greater; at full retirement age, these benefits convert to retirement.**
**Aged Spouse - minimum age 62(reduced benefit), full retirement age (full benefit), spouse must be retired or disabled**
**Widowed Spouse - minimum age 60(reduced benefit), full retirement age (full benefit), spouse must be deceased**
**Other - no age requirements**

**Until the year 2000, the full retirement age was 65. From 2000 to 2022, the full retirement age is increasing by 2 months each year so that by 2022 the full retirement age will be 67.**

**The benefits reported in this section are total benefits received at a point in time. The MBR research extract provided by SSA to create the Gold Standard contains information about different reasons for receiving benefits but does not always allow the amount due to each reason to be accurately separated from the total. Hence we have elected to report total benefits at a point in time and researchers should be careful to note that when an individual is receiving both own retirement and aged spouse benefits, the amounts listed for each benefit type will be redundant, i.e. there is really only one total amount and two reasons for receiving it.**

**SSA calculates benefits based on an individual's lifetime earnings history following rules which they publish in "Annual Statistical Supplement to the Social Security Bulletin," available for each tax year on the Social Security website, www.ssa.gov.**

#### MBR_AGEDSP_BENEFIT

**Label:** MBR: receive agedsp benefit
Yes

**Values:** 0/1

**Description:** Indicates that individual received aged spouse benefits at some point during the time period covered by the MBR extract.

This variable is not synthesized on the SSB. However it is missing due when the SIPP record cannot be linked to the MBR due to lack of an SSN. Hence the Completed Data contain imputed values for this variable.

#### MBR_AGEDSP_BENEFIT_STDATE

**Label:** MBR: benefit startdate
Yes

**Values:** SAS date value

**Description:** Date when the person first began receiving aged spouse benefits, conditional on having ever received this type of benefit.

**MBR_AGEDSP_BENEFIT_TOTAMT**

**Label:** MBR: total monthly benefit

Yes

**Values:** ge 0

**Description:** Total monthly amount of benefits received at beginning of aged spouse benefit entitlement. In most cases this amount is from the same month as in MBR_agedsp_benefit_stdate. However, if data for that month were missing in the MBR extract, we searched through the monthly benefit array to find the first positive value. This amount can be a combination of payments due to multiple entitlement reasons (i.e. dual entitlement).

**PHUS_AGEDSP_BENEFIT_STDATE**

**Label:** PHUS: startdate of benefits

Yes

**Values:** SAS date value

**Description:** Date aged spouse benefits began being paid, as recorded in the PHUS. This date must be greater than or equal to the MBR aged spouse benefit start date. It also must be 1984 or later because PHUS data began in 1984.

**PHUS_AGEDSP_BENEFIT_TOTAMT**

**Label:** PHUS: total monthly benefit

Yes

**Values:** ge 0

**Description:** total amount of benefits as recorded in the PHUS in the first month of receiving aged spouse benefits. This amount can be a sum of benefits received for different reasons (i.e. dual entitlement).

**MBR_DISAB_BENEFIT**

**Label:** MBR: receive disab benefit

Yes

**Values:** 0/1

**Description:** Indicates that individual received disability benefits at some point during the time period covered by the MBR extract.

This variable is not synthesized on the SSB. However it is missing due when the SIPP record cannot be linked to the MBR due to lack of an SSN. Hence the Completed Data contain imputed values for this variable.

**MBR_DISAB_BENEFIT_STDATE**

**Label:** MBR: startdate of benefits

Yes

**Values:** SAS date value

**Description:** Date at which individual began receiving own disability benefits. This date must be before individual reaches the full retirement age (FRA). FRA depends on the year the person reaches age 62. Any individual who turned 62 before 2000 had FRA=65 years old. Beginning in 2000, any individual turning 62 had a full retirement age of 65 years + 2*(year_age_62 - 1999) months.

### MBR_DISAB_BENEFIT_TOTAMT

**Label:** MBR: total monthly benefit

Yes

**Values:** ge 0

**Description:** Total monthly amount of benefits received at beginning of disability benefit entitlement. In most cases this amount is from the same month as in MBR_disab_benefit_stdate. However, if data for that month were missing in the MBR extract, we searched through the monthly benefit array to find the first positive value. This amount can be a combination of payments due to multiple entitlement reasons (i.e. dual entitlement).

### PHUS_DISAB_BENEFIT_STDATE

**Label:** PHUS: startdate of benefits

Yes

**Values:** SAS date value

**Description:** Date disability benefits began being paid, as recorded in the PHUS. This date must be greater than or equal to the MBR disability benefit start date. It also must be 1984 or later because PHUS data began in 1984.

### PHUS_DISAB_BENEFIT_TOTAMT

**Label:** PHUS: total monthly benefit

Yes

**Values:** ge 0

**Description:** total amount of benefits as recorded in the PHUS in the first month of receiving own disability benefits. This amount can be a sum of benefits received for different reasons (i.e. dual entitlement).

### MBR_OTHER_BENEFIT

**Label:** MBR: receive other benefit

Yes

**Values:** 0/1

**Description:** Indicates that individual received other benefits at some point during the time period covered by the MBR extract. There were four types of other benefits: Spouse caring for minor children (TOB=4), Widow(er) caring for minor children (TOB=6), Disabled Widow(er) (TOB=7), Adult disabled in childhood (TOB=8). These types were combined because they are relatively rare and would be a confidentiality risk if reported on an individual basis.
It is important to note that all benefits included in our 5 benefit types are benefits that are received by adults. We do not include information about payments received as a child.

This variable is not synthesized on the SSB. However it is missing due when the SIPP record cannot be linked to the MBR due to lack of an SSN. Hence the Completed Data contain imputed values for this variable.

### MBR_OTHER_BENEFIT_STDATE

**Label:** MBR: benefit startdate

Yes

**Values:** SAS date value

**Description:** Date when the person first began receiving other benefits, conditional on having ever received this type of benefit.

### MBR_OTHER_BENEFIT_TOTAMT

**Label:** MBR: total monthly benefit

Yes

**Values:** ge 0

**Description:** Total monthly amount of benefits received at beginning of other benefit entitlement. In most cases this amount is from the same month as in MBR_other_benefit_stdate. However, if data for that month were missing in the MBR extract, we searched through the monthly benefit array to find the first positive value. This amount can be a combination of payments due to multiple entitlement reasons (i.e. dual entitlement).

### PHUS_OTHER_BENEFIT_STDATE

**Label:** PHUS: startdate of benefits

Yes

**Values:** SAS date value

**Description:** Date other benefits began being paid, as recorded in the PHUS. This date must be greater than or equal to the MBR other benefit start date. It also must be 1984 or later because PHUS data began in 1984.

### PHUS_OTHER_BENEFIT_TOTAMT

**Label:** PHUS: total monthly benefit

Yes

**Values:** ge 0

**Description:** total amount of benefits as recorded in the PHUS in the first month of receiving other benefits. This amount can be a sum of benefits received for different reasons (i.e. dual entitlement).

### MBR_RETIRE_BENEFIT

**Label:** MBR: receive retire benefit

Yes

**Values:** 0/1

**Description:** This variable indicates that a person received retirement benefits at some point during the time period covered by the MBR extract (for dates see …). These benefits were the result of the individual's own earnings history.

This variable is not synthesized on the SSB. However it is missing due when the SIPP record cannot be linked to the MBR due to lack of an SSN. Hence the Completed Data contain imputed values for this variable.

### MBR_RETIRE_BENEFIT_STDATE

**Label:** MBR: benefit startdate

Yes

**Values:** SAS date value

**Description:** Date when the person first began receiving own retirement benefits, conditional on having ever received this type of benefit.

### FLAG_RETIRE_BENEFIT_AMT

**Label:** MBR: in daterange for ben.amt

Yes

**Values:** 0/1

**Description:** This flag=1 if year of MBR_retire_benefit_stdate is greater than or equal to 1962 and less than or equal to last year of the MBR extract. The purpose of this variable is to indicate whether we should expect to know the dollar amount of retirement benefits. Pre-1962, amounts were not saved.

**MBR_RETIRE_BENEFIT_TOTAMT**

**Label:** MBR: total monthly benefit

Yes

**Values:** ge 0

**Description:** Total monthly amount of benefits received at beginning of own retirement benefit entitlement. In most cases this amount is from the same month as in MBR_retire_benefit_stdate. However, if data for that month were missing in the MBR extract, we searched through the monthly benefit array to find the first positive value. This amount can be a combination of payments due to multiple entitlement reasons (i.e. dual entitlement).

**PHUS_RETIRE_BENEFIT_STDATE**

**Label:** PHUS: startdate of benefits

Yes

**Values:** SAS date value

**Description:** Date retirement benefits began being paid, as recorded in the PHUS. This date must be greater than or equal to the MBR retirement benefit start date. It also must be 1984 or later because PHUS data began in 1984.

**PHUS_RETIRE_BENEFIT_TOTAMT**

**Label:** PHUS: total monthly benefit

Yes

**Values:** ge 0

**Description:** total amount of benefits as recorded in the PHUS in the first month of receiving own retirement benefits. This amount can be a sum of benefits received for different reasons (i.e. dual entitlement).

**MBR_WIDOWSP_BENEFIT**

**Label:** MBR: receive widowsp benefit

Yes

**Values:** 0/1

**Description:** Indicates that individual received widowed spouse benefits at some point during the time period covered by the MBR extract.

This variable is not synthesized on the SSB. However it is missing due when the SIPP record cannot be linked to the MBR due to lack of an SSN. Hence the Completed Data contain imputed values for this variable.

**MBR_WIDOWSP_BENEFIT_STDATE**

**Label:** MBR: benefit startdate

Yes

**Values:** SAS date value

**Description:** Date when the person first began receiving widowed spouse benefits, conditional on having ever received this type of benefit.

**FLAG_WIDOWSP_BENEFIT_AMT**

**Label:** MBR: in daterange for ben.amt

Yes

**Values:** 0/1

**Description:** This flag=1 if year of MBR_widowsp_benefit_stdate is greater than or equal to 1962 and less than or equal to last year of the MBR extract. The purpose of this variable is to indicate whether we should expect to know the dollar amount of widowed spouse benefits. Pre-1962, amounts were not saved.

### MBR_WIDOWSP_BENEFIT_TOTAMT

**Label:** MBR: total monthly benefit

Yes

**Values:** ge 0

**Description:** Total monthly amount of benefits received at beginning of widowed spouse benefit entitlement. In most cases this amount is from the same month as in MBR_agedsp_benefit_stdate. However, if data for that month were missing in the MBR extract, we searched through the monthly benefit array to find the first positive value. This amount can be a combination of payments due to multiple entitlement reasons (i.e. dual entitlement).

### PHUS_WIDOWSP_BENEFIT_STDATE

**Label:** PHUS: startdate of benefits

Yes

**Values:** SAS date value

**Description:** Date widowed spouse benefits began being paid, as recorded in the PHUS. This date must be greater than or equal to the MBR widowed spouse benefit start date. It also must be 1984 or later because PHUS data began in 1984.

### PHUS_WIDOWSP_BENEFIT_TOTAMT

**Label:** PHUS: total monthly benefit

Yes

**Values:** ge 0

**Description:** total amount of benefits as recorded in the PHUS in the first month of receiving widowed spouse benefits. This amount can be a sum of benefits received for different reasons (i.e. dual entitlement).

## Supplemental Security Record Variables

**The Supplemental Security Records (SSR) is SSA's main file to track who is receiving Supplemental Security Income (SSI) benefits and the monthly benefit amounts payable. SSI benefits are paid to elderly, blind, or disabled individuals who fall below certain income threshholds. Eligibility and federal payment standards are uniform across all states but states have the option to supplement federal payments. The payment included here is the total of both federal and state SSI payments.**

### FLAG_IN_SSR

**Label:** Flag: In SSR

Yes

**Values:** 0/1

**Description:** This flag indicates that a person's SSN was found in the SSA Supplemental Securtiy Records (SSR). This database tracks people who receive SSI.

### SSR_SSI_DATE_INITIAL_ENTITLE

**Label:** SSR: SSI Date of Initial Entitlement

Yes

**Values:** SAS date values

**Description:** date of initial entitlement to SSI benefits

### SSR_SSI_AMT_INITIAL

**Label:** SSR: SSI Amount - Initial ($2000)

Yes

**Values:** ge 0

**Description:** amount of monthly SSI payment at time of initial receipt

## IRS/SSA Variables

The Census Bureau sent a list of validated SSNs from the seven included SIPP panels to SSA and extracts from the Master Earnings File (Summary and Detailed Earnings Records), Master Beneficiary Record, Supplemental Security Record, 831 Disability File, and Payment History Update System were created. The variables from these files that are included in the SSB are described below.

Not all SIPP respondents have linkages to SSA/IRS administrative data, including: those who refused to provide their SSN; those whose SSNs were not validated; and those with valid SSNs who never worked, and never applied for benefits or received benefits. In the Gold Standard, individuals without a validated SSN or without SSA/IRS administrative records had missing data for all SSA/IRS-derived variables described below. Among these people, those respondents without a validated SSN had all administrative data imputed as part of the data completion process. Hence in the completed Gold Standard and the synthetic data, only individuals with no work or benefit history have zero earnings and missing benefits.

### Detailed Earnings Record Variables

**The Detailed Earnings Records (DER) contains historical earnings reports for each person and job held from 1978 onwards. These reports include self-employment income. Earnings are not capped at the taxable maximum. For each tax year, we summed DER information for each person across all jobs and self-employment to create a total earnings amount.**

#### NONDEFER_DER_FICA

**Label:** DER: Non-Deferred FICA

Yes

**Values:** ge 0

**Description:** Non-deferred earnings (i.e. paid to individual) from jobs covered by FICA tax; summed across all employers in the DER to give a person-level total for each year.

#### DEFER_DER_FICA

**Label:** DER: Deferred FICA

Yes

**Values:** ge 0

**Description:** deferred earnings from jobs covered by FICA tax; summed across all employers in the DER to give a person-level total for each year.
While the variable exists on the Gold Standard for the years 1978-1986, it is always missing in this time period. The year 1987 is the first year with positive deferred wages. On the synthetic and completed gold standard files, we only keep 1990-2006 because so few people had deferred wages between 1987 and 1989 that we could not reliably synthesize these variables.

#### NONDEFER_DER_NONFICA

**Label:** DER: Non-Deferred Non-FICA

Yes

**Values:** ge 0

**Description:** Non-deferred earnings (i.e. paid to individual) from jobs NOT covered by FICA tax; summed across all employers in the DER to give a person-level total for each year.

#### DEFER_DER_NONFICA

**Label:** DER: Deferred Non-FICA

Yes

**Values:** ge 0

**Description:** deferred earnings from jobs NOT covered by FICA tax; summed across all employers in the DER to give a person-level total for each year.
While the variable exists on the Gold Standard for the years 1978-1986, it is always missing in this time period. The year 1987 is the first year with positive deferred wages. On the synthetic and completed gold standard files, we only keep 1990-2006 because so few people had deferred wages between 1987 and 1989 that we could not reliably synthesize these variables.

## Summary Earnings Record Variables

**The SSA/IRS Summary Earnings Records (SER) contain historical person-level earnings data. In addition to an array of annual FICA-taxed earnings (1951-2006) that are capped at the FICA taxable maximum, the SER provides information regarding quarters of covered work. Quarters of covered work are utilized by SSA to determine eligibility for participation in its old age, survivors, and disability insurance (OASDI) programs.**

### TOTEARN_SER

**Label:** SER: Total Earnings

Yes

**Values:** ge 0

**Description:** Annual earnings taxed by FICA; these variables include earnings only up to the FICA taxable maximum, i.e., these earnings measures are capped.

### WQC_YRTOT

**Label:** SER: Annual Total Covered Quarters of Work

Yes

**Values:** 0 to 4

**Description:** Indicates the total number of quarters of FICA-covered work.