Codebook for the SIPP Synthetic Beta File
Version 4.1

October 2007

General Description

The SIPP Synthetic Beta (SSB) file was created by integrating data from the Survey of Program Participation (SIPP), Social Security Administration (SSA), and Internal Revenue Service (IRS) and then synthesizing these data. This work was performed as part of a joint project between the three data contributing agencies. The goal was to create a product that could be used by researchers outside of the regular Census restricted-access facilities. These synthetic data should reproduce the characteristics of the underlying confidential micro-data and, at the same time, assure the confidentiality of the actual data on the sampled individuals. The Census Disclosure Review Board, SSA, and IRS have cleared this file for use by individuals without Census Special Sworn Status. Researchers interested in using the file should submit an application to the Census Bureau. This application can be found on the SIPP home page. Applications will be judged solely on feasibility (i.e. the necessary variables are on the SSB) and researchers will be given accounts on the server housing these data. While no data downloads are permitted at this time, individuals do not have to operate behind the Census firewall to access this server.

Version 4.1 of the SSB is a person-level file with SIPP, Summary Earnings Records (SER), Detailed Earnings Records (DER), and Master Beneficiary Records (MBR) information. This person-level file was created by extracting variables from the 1990-1996 SIPP panels, standardizing these variables, and then stacking individuals across panels. SER data from 1951-2003 and variables from the MBR were then merged onto the SIPP variables. Finally, DER data were summarized for each year from 1978-2003 in order to create person-level records and these were also merged. Merges between administrative data and survey data were performed using the Census Master Consolidated SIPP Crosswalk that maps SIPP person identifiers to SSNs.

An individual was eligible to be included in the SSB if he or she met one major criterion: the individual must have been at least 15 years old at the time of the second wave of the SIPP panel in which that person's household was interviewed. For respondents who were interviewed in the second wave of their respective survey, the variable *popstat* (1990-1993 panels) or *epopstat* (1996 panel) from the wave 2 core data was used to determine eligibility. For those who were not interviewed in wave 2, their age at the end of wave 2 was calculated using their SIPP reported birth date. If the individual would have been 15, he or she was included in the file.

In order to create synthetic data, missing values in the original data first had to be imputed. Thus we created a type of file that we call the Completed Data files. The Completed Data files, which are confidential, are identical to the original data files except that there are no missing values. In the Completed Data files, when the value of a data item is non-missing, the Completed Data files contain original value of that data item. When the value of the data item in the original data file was missing, the Completed Data files contain an imputation. There are four Completed Data files, or four

completed data implicates as they are called in the literature. These contain four separate sets of imputations.

After imputing missing data, we used each Completed Data file as the basis for data synthesis. For every Completed Data file, we created four synthetic data sets by synthesizing conditional on the values in the indicated Completed Data file. Thus, the SSB is actually comprised of 16 total synthetic data sets, or implicates. The SSB contains original values for five unsynthesized variables (described below) and synthetic values, which replace original values, for all other variables. Further documentation is available which describes the synthesizing process in detail.

The goal of the three agencies involved in creating the SSB is to provide an analytically valid file. We define analytic validity to mean unbiased point estimates and variances of these estimates such that inferences drawn about the estimates are similar to inferences drawn from the Completed Data file. In other words, results from an analysis using the synthetic data should not be materially different from results using the original data plus missing data imputations. Initial tests of analytic validity have shown promising results. It is our hope that researchers outside Census, IRS, and SSA will help to further test the analytic validity by running analyses on the SSB and then submitting their programs to the Census Bureau to run on the Completed Data files. A protocol has been developed to facilitate this interaction and allow researchers to receive quantitative information about the validity of their results.

The remainder of this document provides a codebook listing and describing the variables available in the SSB. Further questions should be directed to hhes.synthetic.data.use.list@census.gov.

Person-level File: Variable Descriptions

***Personidm_r*: Person identifier on the SIPP Synthetic Beta File**
m=1 to 4 and r=1 to 4
Unique number for each SIPP respondent in the data set. This is a random identifier and refers to a different source record for each implicate file. It does not link individuals across synthetic implicates. It only identifies an individual within an implicate.

***Spouse_personidm_r*: Spouse identifier on the SIPP Synthetic Beta File**
m=1 to 4 and r=1 to 4
Unique number for each SIPP respondent in the data set. This identifier is not the same as the Personid on the Gold Standard or completed Gold Standard files. It is a random identifier and refers to a different source record for each implicate file.

***M_implicate:* missing data implicate number**
= 1, 2, 3, or 4. Indicates which missing data implicate was the source for creating the synthetic implicate.

***R_implicate:* synthetic data implicate number (within *m_implicate*)**
= 1, 2, 3, or 4.

SIPP Variables that are not synthesized

Male

1=male
0=female

In the 1990-1993 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 chosen by creating an array of variables *sex1-sex*{max number of waves} and choosing the first non-missing value. Thus sex comes from the first wave in which the individual was interviewed instead of from a fixed point in the survey. In the 1996 panel, the longitudinally edited version contains only one value for sex across all waves (*ESEX*), and this value is used. Finally, an indicator variable for males was created from the categorical sex variable for analytic convenience.

Maritalstat

Point in time marital status
1=married, spouse present or absent
2=widowed
3=divorced or separated
4=never married

The original intent of the marital status variable was to identify those respondents who were married at the time of the wave 2 marital history topical module and to identify their spouses using the *spouseid* variable. However, in version 4.1 of the SSB, we have chosen to keep all respondents who meet our age requirement and to take marital status from another point in the survey if it was not collected at the time of the wave 2 topical module. In order to maintain consistency we also take the spouse identification from the same point in time as the marital status information.

Our preferred source of marital status is the variable contained in the wave 2 topical module, which corresponds approximately to month 8 of the survey.[1] If this variable was missing, we cycled through all the months of the survey looking for the first available non-missing marital status information. We began with the months closest to month 8 and gradually moved out. Hence, our second potential source was month 7, followed by months 9, 6, 10, 5, 11, 4, 12, 3, 13, 2, 14, 1, and then months 15 through the end of the survey. The spouse identifier was taken from the same point in the survey as the status information.

This method gave rise to some unforeseen problems. Individuals who did not have wave 2 marital status were often people who joined the survey in a later wave. One of the reasons for entering the panel was marriage to an already existing SIPP respondent. However, this led to a discrepancy in the marital status and *spouseid* reports for the partners of this marriage. The original sample member had marital status and *spouseid* taken from wave 2 while the new sample member had these values taken from whatever wave he or she entered. Thus, the new sample member reported being married to the

[1] In the 1996 survey, the marital status reported in the wave 2 topical module corresponds exactly to the marital status reported in month 8 of the core survey. However in the 1990-1993 panels, the topical module reference month was "month of survey" meaning the questions were asked about the status on the day of the interview. This leads to some discrepancies between the month 8 core marital status and the wave 2 topical module marital status. The major difference is that sometimes the core variable is missing while the topical module variable is not. We used the topical module marital status in all cases.

original sample member but the reverse was not true. If the original sample member had reported being unmarried in wave 2, we changed his or her marital status to be married and assigned the *spouseid* of the new sample member. We discovered several other variations of this miss-matching problem. These are described in detail in technical report describing the creation of the SSB. In general we tried to preserve marital relations whenever possible but we allowed each respondent to report only one spouse and we forced these reports to be consistent within the married couple.

SIPP Variables that are synthesized

Demographic

Black

1=black
0=non-black

In the 1990-1993 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. In the 1996 panel, the longitudinally edited version contains only one value for race across all waves (*ERACE*) and this value is used.

Educ_5cat

Highest level of education attained over course of the SIPP.
1=no high school
2= high school degree
3=some college
4=college degree
5=graduate degree

This variable is created by creating education variables for each wave and then scanning across waves to find the highest value. In the 1990-1993 panels, the variables *higrade* and *grdcmpl* are used to categorize education levels in the following way:
Higrade=1 through 11 is no high school degree
Higrade=12 and *grdcmpl*=2 is no high school degree
Higrade=12 and *grdcmpl*=1 is high school degree
Higrade=21, 22, 23 is some college
Higrade=24 and *grdcmpl*=2 is come college
Higrade=24 and *grdcmpl*=1 is college degree
Higrade=25, 26 is graduate degree
In the 1996 panel, the variable *eeducate* is used to classify education levels in the following way:
Eeducate=31 through 38 is no high school degree
Eeducate=39 is high school degree
Eeducate=40, 41, 42, 43 is some college
Eeducate=44 is college degree
Eeducate=45, 46, 47 is graduate degree

Hispanic

1=Hispanic

0=non-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 panel, 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.

Birthdate
This variable was taken from the Census Bureau's Person Characteristic File (PCF) whose main input is the SSA Numident file. Thus, this variable is administrative and sometimes differs from the birth date reported in the SIPP survey itself. 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.

Flag_deathdate_exist
Flag to indicate that this respondent died after being interviewed and before 2003.
1=death date exists, respondent died during this interval
0=death date does not exist, respondent did not die during this interval.

Deathdate
The source of this variable is also the Census PCF, with death information coming from the SSA Numident and Master Death Files. This variable is coded as a SAS date variable.

Disab_in_scope
In scope to be asked questions about whether health limits kind or amount of work can do;
1=Yes
0=No
For the 1996 panel, 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. For the 1990-1993 panels, disability information comes from the Functional Limitations and Disability topical module (waves 3,3,6,3 for 1990-1993 panels, variables *TM8914, TM8922, TM8924*) and covers people ages 16-67. The topical modules were used for the 1990-1993 panels because the core contained only a question on whether health limited the type or amount of work, while the question on whether disability prevented work was found only in the topical module. In order to make both the disability questions consistent with each other, both responses were taken from the topical module for these early SIPP panels. If respondent was interviewed during the appropriate wave, then the SIPP universe flags were used to determine whether an individual was in scope. If the respondent was not interviewed during the appropriate wave, respondent age at the time

of the wave was calculated and person was judged to be in scope or not based on their age at the time the question *should have* been asked of them.

***Disab*: Health limits kind or amount of work**
.=structurally missing, out of scope for question (*disab_in_scope*=0)
1=Yes (*disab_in_scope*=1)
0=No (*disab_in_scope*=1)

***Disab_nowork*: Health prevents work, taken from first topical module where question is asked (depends on panel)**
.=structurally missing, out of scope for question (*disab_in_scope*=0 or {*disab_in_scop*e=1, disab=0})
1=Yes (*disab*=1)
0=No (*disab*=1)

Fertility History
***Totfam_kids*: number of children under the age of 18 that live in a family in the interview month that was the source for marital status.**
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 panel).
range 1-12

Marital History
Taken from Wave 2 topical module
***Mh_category*: marital history category indicating marital history path of respondent**
0=never married
1=married, never divorced/separated or widowed
2=married once, widowed
3=married once, divorced/separated
4=married once, widowed, married second time
5=married once, divorced/separated, married second time
6=married once, widowed, married second time, widowed
7=married once, divorced/separated, married second time, widowed
8=married once, widowed, married second time, divorced/separated
9=married once, divorced/separated, married second time, divorced/separated
10= married once, widowed, married second time, widowed, married third time
11= married once, divorced/separated, married second time, widowed, married third time
12= married once, widowed, married second time, divorced/separated, married third time
13= married once, divorced/separated, married second time, divorced/separated, married third time
14= married once, widowed, married second time, widowed, married third time, widowed
15= married once, divorced/separated, married second time, widowed, married third time, widowed
16= married once, widowed, married second time, divorced/separated, married third time, widowed
17= married once, divorced/separated, married second time, divorced/separated, married third time, widowed
18= married once, widowed, married second time, widowed, married third time, divorced/separated
19= married once, divorced/separated, married second time, widowed, married third time, divorced/separated

20= married once, widowed, married second time, divorced/separated, married third time, divorced/separated
21= married once, divorced/separated, married second time, divorced/separated, married third time, divorced/separated
22= married once, widowed, married second time, widowed, married third time, widowed, married fourth time
23= married once, divorced/separated, married second time, widowed, married third time, widowed, married fourth time
24= married once, widowed, married second time, divorced/separated, married third time, widowed, married fourth time
25= married once, divorced/separated, married second time, divorced/separated, married third time, widowed, married fourth time
26= married once, widowed, married second time, widowed, married third time, divorced/separated, married fourth time
27= married once, divorced/separated, married second time, widowed, married third time, divorced/separated, married fourth time
28= married once, widowed, married second time, divorced/separated, married third time, divorced/separated, married fourth time
29= married once, divorced/separated, married second time, divorced/separated, married third time, divorced/separated, married fourth time

***Mh1*: first marital history event flag**
=1 first marriage occurred
=0 never married
***Mh2*: second marital history event flag**
=1 first marriage ended in widowhood
=2 first marriage ended in divorce/separation
=0 first marriage did not end over course of survey
***Mh3*: third marital history event flag**
=1 second marriage occurred
=0 no second marriage
***Mh4*: fourth marital history event flag**
=1 second marriage ended in widowhood
=2 second marriage ended in divorce/separation
=0 second marriage did not end over course of survey
***Mh5*: fifth marital history event flag**
=1 third marriage occurred
=0 no third marriage
***Mh6*: sixth marital history event flag**
=1 third marriage ended in widowhood
=2 third marriage ended in divorce/separation
=0 third marriage did not end over course of survey
***Mh7*: seventh marital history event flag**
=1 fourth marriage occurred
=0 no fourth marriage

***Age_mh1*: age at date of first marital history event**
***Age_mh2*: age at date of second marital history event**
***Age_mh3*: age at date of third marital history event**
***Age_mh4*: age at date of fourth marital history event**
***Age_mh5*: age at date of fifth marital history event**
***Age_mh6*: age at date of sixth marital history event**
***Age_mh7*: age at date of seventh marital history event**

***Flag_mar4t*: 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.

***Foreign_born*: Immigrant Status, born in country other than U.S.**
Taken from wave 2 topical module (*TM8730, TM8734, TM8709* 1990-1993 panels, *Ebrstate, Rcitiznt* 1996 panel)
1=born in country other than U.S.
0=born in U.S.

***Time_arrive_usa*: Decade arrive in U.S. (answered when SIPP respondent was foreign_born)**
(*TM8736* 1990-1993 panels, *Rmoveus* 1996 panel)
.=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 – 1990

Economic

Variables that reference a specific year are made by summing month-level variables from the core. Data for missing months was completed (i.e. imputed) and then the annual variable was synthesized.

Detailed Labor Force Participation Characteristics
***Wkswp1990-Wkswp1999*: total number of weeks worked with pay in a year (sum of all months)**
Weeks worked with pay = weeks worked – weeks worked without pay;

***Wkspt1990-Wkspt1999*: total number of weeks worked part-time in a year (sum of all months)**

Weeks worked part-time is, by definition, less than or equal to weeks worked with pay. Part-time is defined as less than 35 hours in a week.

***Tothoursannual1990-Tothoursannual1999*: total number of hours worked at all jobs in a year**

***Ind_exist*: does person have valid industry from a job held during survey**
1=yes
0=no, last worked 1984 or earlier, or no valid industry reported

***Ind_4cat*: industry from first observed job in the SIPP**
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 creating an array of variables *ws1ind1-ws1ind*{max number of waves} in the 1990-1993 panels and *ejbind11-ejbind1*{max number of waves) in the 1996 panel and choosing the first non-missing value. Thus industry comes from the first reported job in the survey.

***Occ_exist*: does person have valid occupation from a job held during survey**
1=yes
0=no, last worked 1984 or earlier, or no valid industry reported

***Occ_4cat*: occupation from first observed job in the SIPP**
1=Managerial and professional specialty occupations
2=Technical, sales, and administrative support occupations
3=other
Created in same manner as industry using variables *ws1occ* from the 1990-1993 panels and *tjbocc1* from the 1996 panel.

Income Variables
***Fpov1990-Fpov1999*: Annual Family Poverty Threshold (created by summing monthly value and dividing by 12);**

***Ftotinc1990-Ftotinc1999*: Annual Total Family Income (sum of monthly values)**

***Totinc1990-Totinc1999*: Annual Total Personal Income (sum of monthly values)**

***Totearn1990-Totearn1999*: Annual Personal Earnings (sum of monthly values)**

Welfare Income variables
***Famwelpart1990-Famwelpart1999*: Annual Welfare Program Participation**
Created using *R27, R20, R21, R24* indicator flags in 1990-1993 panels, *ER27, ER20, ER21, ER24* in the 1996 panel.
1=family of respondent received welfare payments at least one month during the year

0=family of respondent never received welfare payments during the year

***Famwelamt1990-Famwelamt1999*: Annual family total combined benefit dollars from AFDC/TANF, food stamps, general assistance, other welfare sources**
Created using *S27amt, S20amt, S21amt, S24amt* in 1990-1993 panels, *T27amt, T20amt, T21amt, T24amt* in the 1996 panel.
0=structurally missing, out of scope for question, will remain missing (*famwelpart*{year}=0).
=Non-missing values greater than 0 (*famwelpart*{year}=1).

Health limitation income variables
***Helpart1990-Helpart1999*: Annual Program participation**
Created using *R10, R12, R13, R08* indicator flags in 1990-1993 panels, *ER10, ER12, ER13, ER08* in the 1996 panel.
1=respondent received payments for disabilities at some point during the year
0=respondent never received payments for disabilities during the year

***Helamt1990-Helamt1999*: Annual total combined benefit dollars from workers compensation, own sickness (disability) benefits, veterans disability benefits**
Created using *S10amt, S12amt, S13amt, S08amt* in 1990-1993 panels, *T10amt, T12amt, T13amt, T08amt* in the 1996 panel.
0=structurally missing, out of scope for question, will remain missing (*helpart*{year}=0).
=Non-missing values greater than 0 (*helpart*{year}=1).

Wealth Variables from topical module (Waves 4,7,4,7,3 for 1990-1996 panels)
***Totnetworth*: Total Net Worth (rounded to thousands)**
From *hh_tnw* in 1990-1993 panels, *thhtnw* in 1996 panel

***Own_home*: Does respondent own a home?**
From *TM8528, TM8530, TM8608* in 1990-1993 panels, *Ehreunv, Etenure* in 1996 panel
1=yes
0=no

***Homeequity*: Home Equity (rounded to thousands)**
From *hh_theq* in 1990-1993 panels, *thhtheq* in 1996 panel
.=structurally missing, out of scope for question (*own_home*=0)
Non-missing values greater than, less than, or equal to 0.

Nonhouswealth: Non-Housing Financial Wealth (rounded to thousands)
From *hh_twlth* in 1990-1993 panels, *thhtwlth* in 1996 panel
=Total wealth minus home equity

Pension variables from topical module (Waves 4,7,4,9,7 from 1990-1996 panels)
Created from *TM6000, TM6026, TM6028, TM6066, TM6068, TM6114, TM6116* 1993 panel
TM8324, TM8346, TM8348, TM8392, TM8394, TM8436, TM8427 1992 and 1991 panels
TM8324, TM8346, TM8348, TM8392, TM8394, TM8442, TM8444 1990 panel
Earpunv, Rmjb, Rmnjbbs, Eincpens, Emultpen, E1pentyp, E2pentyp 1996 panel

***Pension_in_scope_age*: first level of determination for in scope to be asked pension questions**
If respondent was interviewed during appropriate wave, then the SIPP universe flags were used to determine whether an individual was in scope. Respondents had to be at least 25 years old. If the respondent was not interviewed during the appropriate wave, respondent age at the time of the wave was calculated and *pension_in_scope_age* was set to 1 or 0 based on respondent age at the time the question *should have* been asked of them.
1=yes, old enough to be in scope for pension questions
0=no, not old enough to be in scope for pension questions

***Pension_in_scope_empl*: second level of determination for in scope to be asked pension questions**
If respondent was interviewed during appropriate wave, then the SIPP universe flags were used to determine whether an individual was in scope. Respondents had to work for an employer and could not be unemployed or self-employed. If the respondent was missing employer information or was not interviewed in the appropriate wave, this variable was completed (i.e. imputed).
1=yes, had employer and hence was in scope for pension questions
0=no, didn't have employer and was not in scope for pension questions

***Dc_pension*: Does individual have a defined contribution pension plan?**
.=structurally missing, out of scope for question (*pension_in_scope*=0)
1=yes
0=no

***Db_pension*: Does individual have a defined benefit pension plan?**
.=structurally missing, out of scope for question (*pension_in_scope*=0)
1=yes
0=no

Health Insurance Variables
First a monthly health insurance coverage indicator was created from *Hiind* in the 1990-1993 panels and *Ehimth* in the 1996 panel. Then an employer health insurance coverage indicator was created for each trimester (4 month period) of the year from *Hiscr* in the 1990-1993 panels and from *Ehemply* in the 1996 panel. An individual could only have employer coverage if they reported coverage in general, *i.e.*, employer-provided coverage is treated as a particular type of health insurance. Finally the annual variables were created by checking for any health insurance coverage throughout the year and classifying whether that insurance was employer-provided or not.

***Hicovannual1990*-*Hicovannual1999*: Health Insurance Coverage Variable**
1=respondent had health insurance coverage at some point during the year
0=respondent did not have health insurance coverage at any point during the year

***Hiempannual1990*-*Hiempannual1999*: Employer-provided health insurance coverage**
1=respondent had employer-provided health insurance at some point during the year
0=respondent never had employer-provided health insurance at any point during the year

Weights

Decen_SIPP_wgt_04_01_2000

Weight created by matching SIPP respondents to Decennial census. Reference date for the weight is April 1, 2000. See the technical report describing the creation of the SSB for more details on the creation of this weight variable.

SSA/IRS Administrative Data

The Census Bureau sent a list of validated SSNs from the five included SIPP panels to SSA and extracts from the Master Earnings File (Summary and Detailed Earnings Records) and Master Beneficiary Record were created. The variables from these files that are included in the SSB are described below. For those respondents without a validated SSN, all administrative data were imputed as part of the synthesizing process.

Summary Earnings Records

Earn1937_to_1951

Total earnings between 1937 and 1951 taxed by FICA

Totearn_ser_1951-Totearn_ser_2003

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

Ser_posearn_1951-Ser_posearn_2003

Indicator variable for the presence of positive fica-taxed earnings in a given year.

Wqcp1951-Wqcp1977

Indicates the pattern of FICA covered quarters. Character string that indicates which quarters of the year had FICA-covered earnings. Due to administrative data changes, this series only provides useful information until 1977.

SSA Benefits Data

Tob_initial

Initial type of benefit
1=Retired Worker
2=Disabled Worker
3=Aged Spouse
5=Aged Widow(er)
100=other type of benefit
Taken from *DOEI_TOB* in the Master Beneficiary Record. This variable was not synthesized and contains original data values.

Date_initial_entitle

Date of initial entitlement
Date first entitled to receive benefits from SSA
Taken from *DOEIYY* (year), and first day of first month had recorded MBA value

Mba_initial
Initial Monthly Benefit Amount
Dollar amount of benefit to which individual was entitled in first month of entitlement.
Taken from *mba* monthly array between 1962 and 2004, first month for which there was a positive value.

Tob_2000
Type of benefit in 2000
1=Retired Worker
2=Disabled Worker
3=Aged Spouse
5=Aged Widow(er)
100=other
Taken from either *DOEI_TOB* or *DOEC_TOB*, depending on which one was in effect in 2000. If date of current entitlement was post-2000, then *tob_2000* is the same as *tob_initial*. If date of current entitlement was prior to 2000, then *tob_2000* came from *DOEC_TOB* and was potentially different from *tob_initial*. If date of initial entitlement was post-2000, then this variable is SAS missing (structural zero). This variable was not synthesized and contains original data values.

Mba_2000
Monthly Benefit Amount in 2000
Dollar amount of benefit to which individual was entitled in April 2000. Taken from same *mba* array as *mba_initial*.

AIME_AMW, flag_AIME_AMW_method, PIA, log_AIME_AMW
Average Indexed Monthly Earnings or Average Monthly Wage
Flag_AIME_AMW_method=1 if variable contains AIME
Flag_AIME_AMW_method=2 if variable contains AMW
Primary Insurance Amount

Once the synthetic data files had been created, we created two additional variables that were direct derivatives of SER earnings: Average Indexed Monthly Earnings (AIME) or Average Monthly Wage (AMW) and Primary Insurance Amount (PIA). The AIME/AMW calculation is the method used to summarize a person's lifetime earnings in order to make OASDI benefit calculations. The AIME/AMW is used to calculate the PIA, which in theory tells what benefit a person receives. However, additional rules about spouses, children, family maximums, etc., mean that the actual monthly benefit amount often differs from the PIA. The precise calculations for the AIME/AMW and the PIA depend on a person's gender, date of birth, type of benefit sought, and year of application. The rules governing these calculations are quite complicated (partly because they change a great deal over time) and depend on many things not necessarily observable in our data set. The PIA is an actual variable on the SSA Master Beneficiary File (MBR), but the decision was made by SSA and the Census Bureau not to synthesize this variable or include it on the file, primarily because of concerns that it would be inconsistent with the synthetic SER earnings array. Instead, it was decided that the AIME/AMW and the PIA would be calculated directly from the synthetic earnings using a simplified set of rules.

For individuals who reached age 62 before 1979, we calculated the AMW and for those who reached age 62 after 1979, we calculated the AIME. The variable flag_AIME_AMW_method indicates whether AIME or AMW was calculated for each individual. To compute the AMW, we first calculated the number of years between age 21 (or 1951 if later) and age 62, subtracted five years, and multiplied by 12 to get the number of months at risk. We then summed earnings between age 21 and age 62, dropping the five lowest years. Total summed earnings were then divided by the number of months at risk to give the Average Monthly Wage. There was one exception. For men (but not women) born before 1911, the calculation was performed using the years between age 21 and age 65 because the retirement age for men was three years older prior to 1973. The AIME calculation was essentially the same as the AMW but earnings were indexed to the year in which the individual turned 60.

Once the AIME/AMW had been calculated, the PIA was determined by applying the cut-off points and percentages applicable for the year of initial entitlement to benefits. In a given year, a% of the first X dollars of the AIME formed the initial portion of the PIA. The b% of the next Y dollars formed the next portion and c% of the next Z dollars formed the final portion. The sum of these three portions was the PIA. Prior to 1979, the cut-off points stayed constant across years and the percentages changed. Post 1979, the cut-offs changed every year while the percentages stayed constant. We used tables 2.A8, 2.A10, 2.A11, and 2.A16 from the SSA Statistical Supplement 2005 to make these calculations and consulted with Barbara Lingg at SSA to clarify details.

It is important to note that we calculated the AIME/AMW and the PIA for individuals based on the assumption that they were applying for retired worker benefits. We did not make separate calculations for individuals who received disability, spouse, or death benefits. Thus the AIME/AMW and PIA on the file will not correspond to the MBA for types of benefits other than retired worker. However, since the AIME/AMW and PIA do not contain any additional information and are direct calculations based on other variables in the file, any researcher interested in performing a different calculation may do so. We include these two variables solely for the convenience of retirement researchers. If AIME/AMW is positive, we take the log and include log_AIME_AMW also for convenience.

Detailed Earnings Records (W-2)

Pos_nd_der_fica_1978-pos_nd_der_fica_2003
Indicates presence of positive non-deferred, FICA covered earnings in a given year

Pos_d_der_fica_1978-pos_d_der_fica_2003
Indicates presence of positive deferred, FICA-covered earnings in a given year

Pos_nd_der_nonfica_1978-pos_nd_der_nonfica_2003
Indicates presence of positive non-deferred, non-FICA-covered earnings in a given year

Pos_d_der_nonfica_1978-pos_d_der_nonfica_2003
Indicates presence of positive deferred, non-FICA-covered earnings in a given year

Nondefer_der_fica_1978-nondefer_der_fica_2003
Earnings at all FICA-covered jobs in a year that were not deferred; uncapped

Defer_der_fica_1978-defer_der_fica_2003
Earnings at all FICA-covered jobs in a year that were deferred

Nondefer_der_nonfica_1978-nondefer_der_nonfica_2003
Earnings at all non-FICA-covered jobs in a year that were not deferred; uncapped

Defer_der_nonfica_1978-defer_der_nonfica_2003
Earnings at all non-FICA-covered jobs in a year that were deferred