The 2006 Earnings Public-Use Microdata File: An Introduction

by Michael Compson*

This article introduces the 2006 Earnings Public-Use File (EPUF) and provides important background information on the file's data fields. The EPUF contains selected demographic and earnings information for 4.3 million individuals drawn from a 1-percent sample of all Social Security numbers issued before January 2007. The data file provides aggregate earnings for 1937 to 1950 and annual earnings data for 1951 to 2006. The article focuses on four key items: (1) the Social Security Administration's experiences collecting earnings data over the years and their effect on the data fields included in EPUF; (2) the steps taken to "clean" the underlying administrative data and to minimize the risk of personal data disclosure; (3) the potential limitations of using EPUF data to estimate Social Security benefits for some individuals; and (4) frequency distributions and statistical tabulations of the data in the file, to provide a point of reference for EPUF users.

Introduction

This article introduces the 2006 Earnings Public-Use File (EPUF), a data file containing earnings records for individuals drawn from a 1-percent sample of all Social Security numbers (SSNs) issued before January 2007. EPUF is the latest public-use data file released by the Social Security Administration (SSA) to contain earnings data from its administrative files. EPUF comprises a much larger sample than previously released public-use files containing earnings histories, and significantly enhances the ability of researchers and policy analysts to analyze SSA programs.

EPUF consists of two linkable files. One contains selected demographic and aggregate earnings information for all 4,348,254 individuals in the file, and the second contains annual earnings records for the 3,131,424 individuals who had positive earnings in at least 1 year during 1951–2006. EPUF data reflect capped Social Security taxable earnings. As such, the earnings data contained in EPUF do not present complete measures of the number of workers or the amount of wage-and-salary and self-employment income in the US economy.

The data fields included in EPUF are nearly identical to those in SSA's most recent public-use file containing administrative earnings, the 2004 Benefits and Earnings Public-Use File (BEPUF). This was done (1) to address the critical need to meet data disclosure standards, (2) because of the complexity of the earnings data that SSA has collected over the life of the program, and (3) to maximize EPUF's timeliness. SSA plans to continue working on data disclosure standards for several key detailed earnings data fields from its administrative files. Combining this work with direct

Selected Abbreviations

BEPUF	Benefits and Earnings Public-Use File
EPUF	Earnings Public-Use File
IRS	Internal Revenue Service
MEF	Master Earnings File
QC	quarter of coverage
SSA	Social Security Administration
SSN	Social Security number
YOB	year of birth

* *Michael Compson is with the Division of Policy Evaluation, Office of Research, Evaluation, and Statistics, Office of Retirement and Disability Policy, Social Security Administration.*

feedback from EPUF users, SSA hopes to include new data fields in future releases.

This article informs potential users about the EPUF and provides background information about the data contained in the file. Specifically, the article discusses SSA's experiences collecting earnings data over the years and the effect of those experiences on the data fields included in EPUF; the steps taken to "clean" the data and to minimize the risk of personal data disclosure; and the potential limitations of using the data to estimate benefits for some individuals. Finally, the article presents frequency distributions and statistical tabulations of the data to provide points of reference for EPUF users.

Developing the Earnings Public-Use File

In 2006, SSA released BEPUF, a data file based on a systematic random 1-percent sample of all individuals who were receiving Social Security benefits in December 2004. The file contains benefit and earnings information for the 473,366 individuals in the sample. SSA and Internal Revenue Service (IRS) Data Review Boards reviewed the file to assess the risk of personal data disclosure before approving its release to the public.

The critical question in the initial EPUF development phase involved which data fields to include in the file. Users would undoubtedly like SSA to include all of the data fields from its administrative files. However, SSA has a legal obligation to protect the confidentiality of the individuals included in the file. This creates a tradeoff between the user's need for complete and accurate data and the need to ensure that the file's data fields do not disclose individual identities. Because BEPUF met the disclosure standards set by SSA and the IRS, its data fields served as a starting point for selecting fields for EPUF.

A second critical issue was the need to balance the desire to add data fields with the time needed to prepare the underlying data and conduct the required data-disclosure analysis. SSA originally hoped to include earnings data fields beyond those included in BEPUF. However, choosing fields to add to the file was complicated by more than data-disclosure limitations. Reconciling the types of earnings data in SSA's administrative files with the different data-collection timelines over the life of the program made seemingly simple choices fairly complicated.

To include new data fields would be much more complex because the additional fields would come from the detailed segment of the Master Earnings File (MEF).[1] For each individual, the detailed segment is likely to contain more than one earnings record in a given year. As a result, working with the detailed segment of the MEF is much more complicated and would take more time and effort than working with data fields from the summary segment of the MEF, as was done for BEPUF.

In addition, the only earnings data field that is available for all years from 1951 through 2006 is taxable earnings. Other fields of interest, such as noncovered earnings, covered earnings above the taxable maximum, and contributions to 401(k) retirement plans, are only available for selected years.[2] Consider self-employment income: From 1951 through 1977, self-employment income is included in the earnings data field only to the extent that it is covered under the Social Security program. If an individual had wage-and-salary earnings above the taxable maximum and also had self-employment income, none of the self-employment income would be included in the earnings record. This produces undercounts of both the number of individuals with self-employment income and the dollar amount of that income. From 1978 through 1993, the detailed segment of the MEF contains a separate value for covered self-employment income. However, the amount reported in this field is still limited to earnings covered under the program. The full amount of self-employment income does not appear in the MEF until 1994, when the cap for covered earnings subject to the Medicare Hospital Insurance payroll tax was eliminated. As a result, the administrative files do not contain a complete history of an individual's self-employment income.

After accounting for all of these considerations, SSA designed EPUF to contain nine data fields in two linkable data tables. The first linkable file contains a single record for each of the 4,348,254 individuals included in EPUF. Each record contains the following data fields:

- ID (a unique identification number)
- year of birth (YOB)
- sex
- aggregate capped Social Security taxable earnings from 1937 through 1950

- aggregate quarters of coverage (QCs) earned from 1937 through 1950
- aggregate QCs earned in 1951 and 1952

The second linkable file contains 60,326,474 earnings records with positive earnings values. There are 3,131,424 individuals in this file who had positive earnings for at least 1 year during 1951–2006. Each of the records in this file contains the following data fields:

- ID (a unique identification number)
- the year(s) when the individual had taxable Social Security earnings
- the amount of capped Social Security taxable earnings for each of those years
- the number of QCs earned for each year (except 1951 and 1952) based on the amount of capped Social Security taxable earnings

These data fields are identical to those included in the BEPUF with one minor exception. EPUF contains multiple data fields for the QCs: aggregate QCs earned 1937–1950 and aggregate QCs earned in 1951 and 1952 in the first linkable file; and annual QCs earned from 1953 through 2006 in the second linkable file. By contrast, the BEPUF contains a single aggregate value for QCs earned as of December 31, 2004. Because of this difference, an EPUF user can determine an individual's eligibility for retired-worker and disabled-worker benefits at any given time.

Overview of Earnings Records

SSA's primary objective in collecting earnings data is to meet the operational needs of the program.[3] As a result, the data contained in EPUF will be, in some aspects, somewhat limited from a researcher's perspective. However, the uniqueness of the data and the large sample size should outweigh these limitations in many cases.

To use EPUF appropriately, users must understand the nature of its earnings data. For example, analysts must be aware that the earnings data in EPUF do not reflect all workers in the US labor market, nor the aggregate earnings generated by those workers.[4] Putting the EPUF earnings data in their proper context requires an understanding of three measures of earnings distinct to the Social Security program: covered earnings, Social Security taxable earnings, and capped Social Security taxable earnings.

The first measure refers to earnings "covered" for purposes of determining eligibility for the Social Security program. The Social Security Act defines the types of employment covered under the program, and coverage has expanded significantly over the years.[5] Currently, nearly all types of employment are covered under Social Security. There are three primary exceptions: "state and local government employees whose employer has not elected to be covered under Social Security and who are participating in an employer-provided pension plan, current Federal civilian workers hired before 1984 who have not elected to be covered, and self-employed workers earning less than $400 in a calendar year" (Board of Trustees, 2010). "Covered earnings" has two components: wage-and-salary earnings from covered employment, and self-employment income covered under the program.

The second measure is called Social Security taxable earnings because it reflects all covered earnings that are subject to the payroll tax.[6] The annual earnings data in the MEF summary segment are a running total of an individual's taxable earnings up to the taxable maximum for each job in a given year, plus any taxable self-employment income. For the self-employed, "taxable earnings consists of net self-employment income which, when combined with any taxable wages for that individual, is at or below any applicable annual maximum taxable amount" (SSA 2009, G.17). If an individual has more than one employer, the amount of earnings in this data field may be greater than the taxable maximum in a given year.

EPUF uses the third measure, capped Social Security taxable earnings, defined as the total amount of a worker's taxable earnings (including any taxable self-employment income) up to the taxable maximum in a given year. It does not include any earnings beyond the taxable maximum, as the previous measure can when a worker has multiple employers. This measure allows an observer to determine total amounts contributed to the program by workers and self-employed individuals.[7] The primary reason EPUF uses this measure is that capped taxable earnings do not need to be top-coded for data disclosure purposes. Second, because the IRS and SSA approved BEPUF for release using capped taxable earnings, using the same measure in EPUF was deemed likely to expedite its approval.

Two adjustments were made in moving the taxable earnings data from the MEF summary segment to the capped taxable earnings information contained in

EPUF. First, all earnings values were top-coded at the taxable maximum in a given year. Second, any records with negative covered earnings were set to zero (this occurred very infrequently).

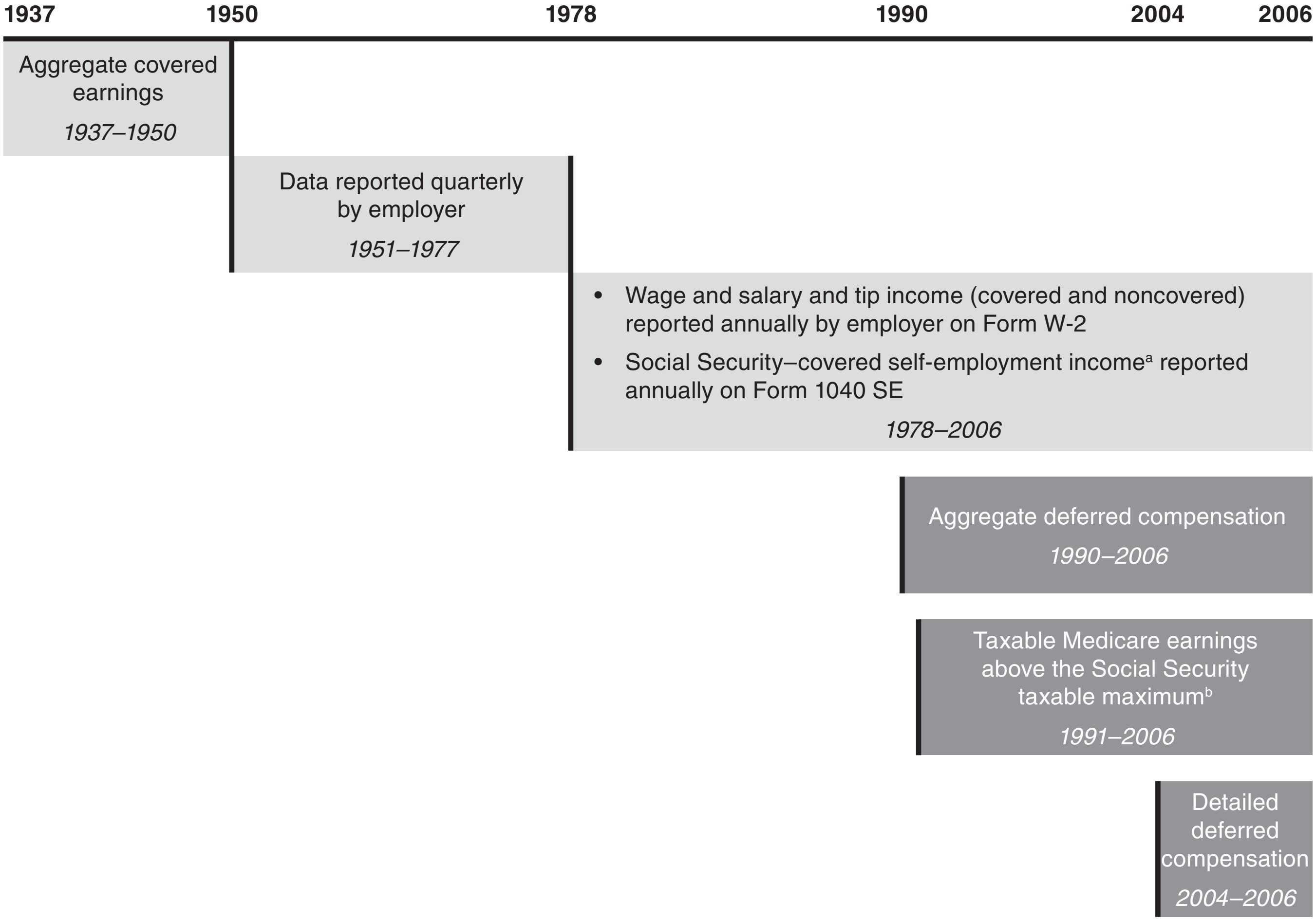
Through 2006, SSA used three distinct mechanisms to collect the earnings data required to administer its programs: (1) paper and microfilm records that yield an individual's total covered earnings from 1937 through 1950, (2) quarterly earnings data reported by the individual's employer from 1951 through 1977, and (3) annual earnings reported by the individual's employer on Form W-2 from 1978 through 2006 (Chart 1).

In the years since the adoption of Form W-2, three additional types of earnings data have been collected to reflect expanded data needs: (1) aggregate deferred compensation, used to calculate the national average wage index, beginning in 1990; (2) Medicare taxable wage-and-salary and self-employment income, beginning in 1991; and (3) detailed items for the deferred compensation field, beginning in 2004.[8] These changes are also reflected in Chart 1.

1937–1950 Earnings Data

Before the arrival of electronic data storage, SSA stored earnings data on either paper or microfilm.

Chart 1.
Types of earnings data available from Social Security administrative files, 1937–2006


1937
1950
1978
1990
2004
2006
Aggregate covered earnings
1937–1950
Data reported quarterly by employer
1951–1977
• Wage and salary and tip income (covered and noncovered) reported annually by employer on Form W-2
• Social Security–covered self-employment income[a] reported annually on Form 1040 SE
1978–2006
Aggregate deferred compensation
1990–2006
Taxable Medicare earnings above the Social Security taxable maximum[b]
1991–2006
Detailed deferred compensation
2004–2006

SOURCE: SSA.

a. From 1978 to 1990, data for only that portion of self-employment income that it is taxable for Social Security purposes are available. In general, during this period there is no way to distinguish between amounts of covered earnings from wages and salary, self-employment income, and earnings from agriculture. Beginning in 1991, the taxable maximum earnings amounts for Social Security and Medicare differed. Beginning in 1994, the cap on taxable Medicare covered earnings was eliminated, and data on total earnings amounts from each source became available.

b. Beginning in 1991 the Medicare taxable maximum earnings amount exceeded the Social Security taxable maximum, until the Medicare taxable maximum was eliminated altogether in 1994.

Given the limited storage capacity of early computers and the prohibitive costs associated with converting these data to electronic format, the earnings data for 1937–1950 on the MEF summary segment are available only as an aggregate number. As a result, the data extract from which EPUF is drawn contains two data fields for aggregate taxable earnings—one for 1937–2006, and the other for 1951–2006. The EPUF data field for aggregate Social Security taxable earnings from 1937–1950 was generated by subtracting the 1951–2006 aggregate earnings from the 1937–2006 aggregate earnings.

Another data field of interest is the QCs earned during this period. An individual can earn up to four QCs in a year depending on his or her taxable earnings amount. QCs determine an individual's eligibility for retirement and disability benefits and a family's eligibility for survivor benefits. The MEF summary segment contains no annual values for QCs for 1937–1953. Instead, the extract contains data fields from the MEF that contain the "known" aggregate number of QCs earned during the following periods: 1947–2006, 1951–2006, 1947–1952, and 1953–2006. For EPUF, these data fields are manipulated to generate the aggregate number of QCs earned for the periods 1947–1950 and 1951–1952.

Because the MEF has no known values for QCs from 1937 through 1946, SSA devised a three-step method to estimate the aggregate number of QCs earned by individuals with covered earnings during these years.[9] The first step assigns one QC for each $500 of aggregate taxable earnings from 1937 through 1950. The second step subtracts the known sum of QCs earned from 1947 through 1950. (The QCs from 1947 through 1950 are generated by subtracting the known number of QCs earned from 1951 through 2006 from the known number of QCs earned from 1947 through 2006.) If the resulting number is positive, this value is assigned to the number of QCs earned from 1937 to 1946. If this number is negative, a value of 0 is assigned for the number QCs earned from 1937 to 1946. The final step adds the estimated QCs from 1937 to 1946 to the known QCs from 1947 to 1950 for the estimated number of QCs earned from 1937 to 1950.[10]

1951–1977 Earnings Data

From 1951 through 1977, the earnings data used to administer Social Security came from two sources: the individual's employer and the IRS. SSA required employers to report covered wage-and-salary income quarterly. For the self-employed, the IRS processed the annual Social Security taxable self-employment income reported on the individual's Form 1040 on Schedule C and Schedule SE and transferred the data to SSA. Values in these data fields were added together to create a single entry for taxable Social Security earnings, which is stored on the Summary Earnings Record. As a result, it is not possible to determine whether covered earnings in a given year are from wages and salaries or from self-employment income. The MEF also contains separate indicators for the presence of self-employment income (Schedule C) or agriculture income (Schedule F) in a given year. However, if there are combinations among salary and wages, self-employment income, and income from agriculture, the amounts attributable to each source cannot be determined. As a result, these flags were not included in EPUF.[11]

As previously noted, the MEF has no annual values for the number of QCs earned in 1951 and 1952. This value is estimated by manipulating data used to calculate QCs from 1937 through 1950. Beginning in 1953, the MEF contains annual QC values based on quarterly earnings data.

1978–2006 Earnings Data

In 1978, SSA earnings data underwent major changes involving sources, processing, and types of data collected. Because requiring quarterly earnings reports had led to processing delays and administrative burdens, new legislation required employers to report their employee's earnings annually on Form W-2. The legislation also made SSA responsible for processing the W-2 earnings data. The source for self-employed taxable earnings, Form 1040 Schedule SE, remained unchanged.

The move to annual collection of earnings data resulted in three significant changes in the types of data collected:

- The W-2 included earnings from employment that was not covered under Social Security. Prior to 1978, SSA was only concerned with taxable earnings from covered employment.
- The ability to store data electronically and the need for more detailed earnings information to administer the program led SSA to establish separate data fields for taxable wage-and-salary income and taxable self-employment income. Prior to 1978,

administrative data contained a single entry for all taxable earnings.

- The W-2 allowed SSA to capture covered wage-and-salary income above the taxable maximum. Earnings reported to SSA for all previous years were capped at the taxable maximum.

It is important to note that the inclusion of taxable self-employment income on an individual's earnings record reflects the reporting criteria used during two distinct periods. For 1978 through 1993, self-employment income appears on an individual's earnings record only when Social Security or Medicare taxes were due on that income. It was not until 1994, when the cap for taxable earnings subject to the Medicare payroll tax was eliminated, that SSA's earnings data began to include uncapped values for covered self-employment income.

Several examples illustrate how the amount of taxable self-employment income differs from the amount of self-employment income reported for federal income tax purposes across these two periods. Suppose an individual earned $25,000 in covered wages and $25,000 in self-employment income, and assume a taxable maximum of $40,000. Prior to 1994, the individual's earnings record for that year would contain $25,000 for wage-and-salary income and only $15,000 for self-employment income. Now consider an individual with self-employment income of $55,000 and no covered wages. In this example, the individual's earnings record would have $40,000 for taxable self-employment income. From 1994 onward, there is no cap on the amount of covered earnings subject to the Medicare payroll tax. As a result, the full amount of both wage-and-salary and self-employment income in the examples above would be included in the individual's earnings record on the MEF, but is not in EPUF.

The Revenue Act of 1978 also affected the earnings data collected by SSA by allowing the elective deferral of wage earnings.[12] Elective deferrals enabled individuals to postpone the receipt and the taxation of certain types of earnings. This led to the creation of 401(k) retirement plans, 403(b) plans for employees of nonprofit organizations, and 457 plans for state and local government employees. From 1978 through 1983, these elective deferrals were not covered under Social Security. As a result, the taxable earnings data in EPUF for these years do not include contributions to these plans.

Starting in 1984, elective deferrals are covered under the program and are reflected in the taxable earnings in EPUF (up to the taxable maximum). In 1990, SSA was required to include elective deferrals in the calculations of the average wage index, and created a separate data field in the MEF detailed section to capture this information.

Data on annual QCs earned during 1978–2006 are based on taxable earnings in a given year. As noted earlier, the MEF contains annual QC values after 1952.

Sample Selection, Data Cleaning, and Disclosure Protection

EPUF consists of earnings records drawn from a 1-percent sample of the MEF (the "underlying EPUF sample"). A series of data cleaning and disclosure protection procedures produced the final EPUF. This section describes the process of selecting the underlying EPUF sample, the data cleaning steps, and the disclosure protections that were applied to the data to produce the EPUF.

Sample Selection

The sample universe for the EPUF is all SSNs issued before January 2007. Thus, any individual who does not have an SSN cannot be included in the EPUF. The EPUF sample was created using a systematic sampling process that closely approximates a random sample. For each area-group combination, an algorithm selects 100 out of the possible 10,000 SSNs.[13] SSA then determines if the SSNs have been issued. The sampling algorithm is systematic in that it avoids any overlap between the BEPUF, EPUF, and any potential future public-use samples generated using the algorithm.[14] SSA has determined that the design effect for the systematic random sample is effectively equal to one.[15]

The SSNs generated using this algorithm were checked for inclusion in the Numident file to confirm their presence in the Social Security administrative files.[16] A final check verified that none of the SSNs in the sample overlapped those in the BEPUF. The individuals in the resulting underlying EPUF sample numbered 4,413,024.[17] Note that the sample is not strictly representative of the US population because the sampling universe (all SSNs issued) includes individuals in Puerto Rico and the US territories.

Data Cleaning

A number of analyses were undertaken to determine if there were any problems with the data and, if so, what to do about them. Three key issues were identified: (1) a coding error incorrectly assigned a YOB value equal to 1900 to many individuals, (2) some YOB values were missing, and (3) some extreme age values occurred for individuals who had taxable earnings (values ranged from -47 years to 179 years).[18] Several other smaller issues were discovered in the process of generating the EPUF and a number of steps were taken to "clean" the data before releasing the file to the public.

The first check involved graphing the distribution of individuals in the underlying sample by their YOB. This graph produced an abnormally large spike in the number of individuals with a YOB value equal to 1900. For these 24,843 individuals, a check against the Numident file confirmed a YOB value of 1900 on 21,269 records. There were 3,464 individuals whose YOB value was missing on the Numident file; these were removed from EPUF. This left 110 individuals with an alternative (non-1900) YOB value on the Numident file. The Numident's alternative value was assigned for those individuals.

The next data-cleaning issue involved the 13,405 individuals in the underlying sample whose MEF records had a missing value for YOB. The overwhelming majority (12,142) also had a missing value for YOB on the Numident file; these individuals were removed from EPUF. Of the remaining 1,263 individuals, 1,234 had a single YOB value on the Numident file; for them, the Numident YOB was used. This left 29 individuals who had multiple YOB values on the Numident file; for these, we assigned a "best" YOB value.

The analysis of the age at which an individual in the underlying sample recorded taxable earnings found 77,458 individuals who either had age values of less than 14 or greater than 79, or had earnings during 1937–1950 but a YOB value after 1950. Again, MEF records were validated against the Numident file. Records for 5,810 individuals were removed for one of the following reasons: there was no logical choice among multiple alternative YOB values on the Numident, age when recording taxable earnings was either negative or greater than 100, or the YOB value was after 1950 although earnings were recorded during 1937–1950.

The final adjustments included removing 5,935 individuals whose YOB value was before 1870, removing 1,096 individuals whose YOB value was equal to 2007, and removing 4 individuals who were assigned a missing YOB value. Individuals born before 1870 were removed because they were unlikely to have received Social Security benefits. The data for the underlying sample were extracted in 2007 and it is possible that a small number of individuals who were enumerated after December 31, 2006 were part of the sample.

Data "cleaning" procedures resulted in the removal of records for 28,451 individuals from the underlying sample. The effect of removing these individuals on the number of earnings records and on the amount of earnings by year is discussed later in conjunction with the effect of the data disclosure procedures.

Disclosure Protection

The most critical determinant of whether data fields can be included in the public-use file is disclosure risk. To protect confidentiality, SSA removes all identifying information, evaluates disclosure risk posed by administrative earnings data for individuals that overlap other public-use files,[19] and modifies any distinguishing characteristics that could identify individuals in the file. The data disclosure procedures applied to the EPUF fall into three broad categories: (1) removing any identifiable information from the file and evaluating the disclosure risk of public-use file overlap, (2) adjusting the earnings amounts to create a range of uncertainty between the amount of earnings reported to SSA and the amount released in EPUF, and (3) zeroing out earnings records because of age considerations. These categories are described in detail below.

Removing identifiable information and evaluating disclosure risk from public-use file overlap. To minimize disclosure risk, the following steps were taken:

- All SSNs were removed from the file.
- The records in the final EPUF were randomly sequenced.
- Where possible, EPUF sample records were checked for overlap with other public-use files.

As previously noted, there is no overlap between individuals in BEPUF and EPUF. There were 319 individuals in the underlying EPUF sample who were included in the New Beneficiary Data System (NBDS). These individuals were removed from the sample.[20]

PREVIOUS PUBLIC-USE DATA FILES WITH EARNINGS DATA

SSA has released a number of public-use microdata files that contain earnings data from its administrative files. The first six items listed below are products of two interagency studies undertaken in the 1970s and 1980s: the 1963 Pilot Link Study and the 1973 Exact Match Study, conducted by SSA, the Census Bureau, and the IRS. SSA produced items 7 and 8 independently.

1. The 1964 Current Population Survey—Administrative Record Pilot Link File
2. The 1973 Current Population Survey—Summary Earnings Record Exact Match File
3. The 1973 Current Population Survey—Administrative Record Exact Match File
4. The Social Security Longitudinal Earnings Exact Match Public Use File, 1937–1975
5. The 1972 Augmented Individual Income Tax Model Exact Match File
6. The Retirement History Longitudinal Survey, 1969–1973, and Summary of Social Security Earnings: Merged Data
7. The New Beneficiary Data System
8. The 2004 Beneficiary and Earnings Public-Use File

The 1963 Pilot Link Study matched data from Census Bureau's Current Population Survey with SSA and IRS administrative data files. The 1973 Exact Match Study refined the 1963 Pilot Link Study processes. The primary objective of both studies was to improve the quality of statistical output related to income distribution and redistribution.

The Retirement History Study matched survey data with Social Security administrative data to create public-use data files useful for researching retirement decisions and circumstances.

The New Beneficiary Data System consists of two separate surveys. The original survey was the New Beneficiary Survey, a nationally representative survey of beneficiaries who were in payment status during a 12-month period from mid-1980 to mid-1981. In 1992, SSA conducted the New Beneficiary Followup (NBF) survey and attached limited earnings data to all 18,599 individuals in the original survey.

The 2004 Beneficiary and Earnings Public-Use file, released in 2006, is a systematic random sample of individuals who were on the benefit rolls as of December 2004.

Although minimal overlap between individuals in EPUF and individuals in the Synthetic SIPP Beta files (SSB) is likely, the SSA and IRS have concluded that there is no disclosure risk because all of the earnings data in the SSB are synthetic.[21]

The number of individuals in EPUF who are potentially included in the public-use files created from the 1964 Pilot Link Study, the 1973 Exact Match Study, and the Retirement History Study is very small (see text box). SSA and the IRS have determined that disclosure resulting from overlap of these files is very unlikely.

Adjusting earnings to create a range of uncertainty and limit potential disclosure. With a few exceptions, the earnings amounts in EPUF were random-rounded to a base of $25, $100, or $1,000, depending on the amount of earnings reported to SSA.[22] Specifically,

- earnings greater than $100 and less than $1,000 were random-rounded to a base of $25;
- earnings greater than $1,000 and less than $50,000 were random-rounded to a base of $100; and
- earnings greater than $50,000 were random-rounded to a base of $1,000.

Using this process, earnings near the taxable cap could be rounded up to the taxable maximum, and very low earnings could be rounded down to zero. SSA was concerned that this could affect two key research issues: (1) analyses of the differences between workers and nonworkers (as defined in terms of covered employment) and (2) analyses comparing individuals with earnings above and below the taxable maximum in a given year. To maintain the integrity of the data in these two areas, and to eliminate the possibility of rounding down to zero or rounding up to the taxable maximum in a given year, the following steps were taken:

- All annual earnings values less than $100 were replaced with the average amount of all earnings less than $100 in a given year.

- All annual earnings within the random rounding base of the taxable maximum ($100 or $1,000, depending on the taxable maximum in a given year) were replaced by the average of all values within the rounding base for that year.
- Any values for the aggregate amount of earnings from 1937 to 1950 greater than $37,000 were replaced with $41,500 (the average value of all aggregate earnings amounts greater than $37,000).
- Any values for the aggregate amount of earnings from 1937 to 1950 that were less than $100 were replaced with $39 (the average dollar amount for all values of aggregate earnings less than $100).

These adjustments to the random-rounding process may reduce the amount of uncertainty between the earnings reported to SSA and those contained in EPUF for a select group of individuals. Consider an individual with $100 in earnings. We know that the actual value of earnings reported to SSA for this individual had to be between $100 and $124. This creates a range of uncertainty of only $25 instead of plus or minus $25. However, this limited range of uncertainty only occurs for the $100 value of earnings.

Second, consider an individual with earnings of $95,250 in a year when the taxable maximum was $96,000. This individual's earnings value was replaced with the average value for all individuals with earnings from $95,001 and $95,999. In this case, we know the actual value of earnings reported to SSA to within $1,000. This is a much smaller range of uncertainty than the difference of plus or minus $1,000 that applies to earnings greater than $50,000 and not within the random-rounding base of the taxable maximum.

Third, the random-rounding process may also affect the number of annual QCs included in EPUF for 1953–2006. On the MEF, QCs are calculated based on the quarterly earnings (1951 to 1977) and on annual earnings (1978 to 2006) recorded for a given year. However, the random-rounding process can change the value of earnings by plus or minus $25, $100, or $1,000, depending on the amount of taxable earnings in a given year. Thus, QCs based on randomly rounded earnings values may differ from those based on the MEF.

This potential discrepancy raises questions about the effectiveness of the random-rounding process. Consider a case in which the amount of earnings on the MEF is $735 and the rounded earnings value is $750 for a year in which $250 are needed to earn a QC. The QCs based on MEF earnings would be two, and the rounded-earnings QC value would be three. By using the MEF QC value in EPUF we would know that the actual earnings reported to SSA would be between $725 and $750. In addition to reducing the range of uncertainty for the individual's earnings, this could affect analyses of eligibility for benefits.

In this light, the question arises: What is the appropriate value for QCs to include in EPUF? A comparison of the QC measure on the MEF with that based on randomly rounded earnings found the following four items:

- Of 60,326,474 records with positive earnings, QC values differed on only 175,609 (0.29 percent).
- When records differed, the maximum difference was plus or minus one QC.
- The aggregate number of QCs based on randomly rounded earnings (213,915,632) was 39,389 fewer than the aggregate number of quarters on the MEF, a difference of only 0.018 percent.
- The net impact of random rounding on total QCs earned at the individual level was very small. Among those whose records were affected, nearly 97 percent had a net difference of plus or minus one quarter over their work histories.

Given the very small differences between the two QC measures, SSA included the MEF measure in EPUF because it reflects an individual's actual number of QCs earned.

Zeroing out earnings for certain ages. When the BEPUF was created, the IRS requested that SSA zero out all earnings for individuals born after 1937 who had earnings at ages 14 or younger to prevent disclosure of potentially identifiable data.

SSA applied these same data disclosure procedures to EPUF. In addition to zeroing out any earnings for individuals who were very young, SSA assigned a value of zero to any earnings records that had a positive value when the individual was aged 86 or older.

Table 1 shows the number of records that SSA either removed from the underlying EPUF sample because of data cleaning or assigned a value of $0 because of data disclosure procedures, along with the dollar value of earnings represented by these omitted records.[23] Table 2 shows the number of records and the value of earnings represented in the entire underlying EPUF sample, in the omitted records, and in the resulting final EPUF, revealing that the omitted records are a very small share of the original underlying sample.

Table 1.
Earnings records removed from underlying EPUF sample or with earnings values set to zero for data cleaning or disclosure protection procedures, 1951–2006

	Records removed for data cleaning		Records with earnings values set to zero for individuals aged—				Total	
			14 or younger		86 or older			
Year	Records	Dollar amount	Records	Dollar amount	Records	Dollar amount	Records	Dollar amount
1951	2,759	5,665,897	1,646	254,528	0	0	4,405	5,920,425
1952	2,829	5,941,257	1,793	283,133	0	0	4,622	6,224,390
1953	2,805	6,027,761	1,778	316,999	0	0	4,583	6,344,760
1954	2,712	5,880,985	1,216	211,168	0	0	3,928	6,092,153
1955	3,024	6,959,324	1,496	269,778	0	0	4,520	7,229,102
1956	3,113	7,458,390	1,560	298,590	56	77,183	4,729	7,834,164
1957	3,085	7,594,978	1,494	304,594	88	140,290	4,667	8,039,862
1958	3,032	7,390,087	1,036	235,218	115	179,890	4,183	7,805,195
1959	3,037	8,135,307	1,048	247,442	135	204,584	4,220	8,587,334
1960	2,997	8,186,207	1,129	246,054	148	273,315	4,274	8,705,575
1961	2,945	8,086,654	1,080	238,310	170	315,373	4,195	8,640,337
1962	2,937	8,339,769	1,022	241,864	173	340,236	4,132	8,921,869
1963	2,928	8,465,681	1,158	260,460	182	358,582	4,268	9,084,723
1964	2,919	8,789,314	1,208	286,514	181	397,263	4,308	9,473,091
1965	2,987	9,166,718	1,454	366,245	189	425,929	4,630	9,958,893
1966	3,035	11,318,086	1,963	477,524	210	506,443	5,208	12,302,053
1967	3,027	11,629,233	2,128	544,917	193	511,459	5,348	12,685,609
1968	3,071	13,106,921	2,459	707,891	212	549,174	5,742	14,363,986
1969	3,084	13,678,081	2,887	903,985	217	567,872	6,188	15,149,938
1970	3,084	13,777,730	2,758	987,296	225	563,126	6,067	15,328,153
1971	3,060	14,117,871	2,758	966,145	203	579,624	6,021	15,663,640
1972	3,069	15,613,850	3,224	1,254,230	234	680,321	6,527	17,548,401
1973	3,069	17,630,854	4,007	1,565,846	246	870,503	7,322	20,067,203
1974	3,096	19,670,766	4,083	1,828,581	258	950,736	7,437	22,450,084
1975	2,948	20,124,329	3,587	1,817,022	247	1,082,987	6,782	23,024,338
1976	2,965	21,504,597	3,606	2,023,184	270	1,142,883	6,841	24,670,664
1977	2,972	22,805,353	4,035	2,484,999	275	1,228,865	7,282	26,519,217
1978	2,948	24,277,547	4,569	3,479,281	299	1,459,620	7,816	29,216,447
1979	2,927	27,336,728	4,339	3,915,380	302	1,615,747	7,568	32,867,855
1980	2,852	28,188,933	3,754	4,130,883	296	1,694,111	6,902	34,013,927
1981	2,736	28,247,820	3,433	4,092,412	278	1,680,975	6,447	34,021,207
1982	2,557	28,006,503	3,019	4,123,014	320	1,963,325	5,896	34,092,842
1983	2,498	28,325,155	2,886	4,092,376	339	2,175,128	5,723	34,592,659
1984	2,525	29,220,576	3,474	4,682,009	325	2,140,629	6,324	36,043,213
1985	2,482	30,028,067	3,893	5,404,605	344	2,151,727	6,719	37,584,399
1986	2,452	30,415,341	3,593	5,086,159	358	2,245,808	6,403	37,747,309
1987	2,403	30,272,513	3,896	5,377,760	345	2,220,378	6,644	37,870,651
1988	2,410	30,171,045	4,402	4,589,446	324	2,461,835	7,136	37,222,326
1989	2,336	30,739,323	4,693	4,514,906	354	3,015,574	7,383	38,269,803
1990	2,293	30,787,395	4,039	4,082,369	337	3,115,513	6,669	37,985,278
1991	2,198	29,839,368	3,427	3,380,565	354	3,031,322	5,979	36,251,255
1992	2,151	30,622,053	3,444	3,320,321	385	3,233,655	5,980	37,176,029
1993	2,311	31,248,868	3,453	3,833,342	497	3,287,167	6,261	38,369,376
1994	2,331	32,203,088	3,847	4,116,755	554	3,072,048	6,732	39,391,891
1995	2,334	33,036,888	3,725	4,345,292	548	3,489,519	6,607	40,871,699

(Continued)

Table 1.
Earnings records removed from underlying EPUF sample or with earnings values set to zero for data cleaning or disclosure protection procedures, 1951–2006—*Continued*

	Records removed for data cleaning		Records with earnings values set to zero for individuals aged—				Total	
			14 or younger		86 or older			
Year	Records	Dollar amount	Records	Dollar amount	Records	Dollar amount	Records	Dollar amount
1996	2,318	33,864,278	3,868	4,744,775	553	3,596,750	6,739	42,205,802
1997	2,305	35,451,643	3,928	5,780,153	614	4,349,378	6,847	45,581,174
1998	2,308	37,255,636	4,126	6,576,731	638	4,517,465	7,072	48,349,833
1999	2,284	38,915,191	4,010	7,408,910	678	5,136,825	6,972	51,460,925
2000	2,250	40,225,040	4,122	7,885,400	751	5,085,548	7,123	53,195,988
2001	2,184	40,499,362	3,712	7,971,572	764	5,649,651	6,660	54,120,585
2002	2,078	40,125,933	3,271	7,919,378	733	6,126,470	6,082	54,171,781
2003	1,986	39,695,001	2,869	7,885,607	848	7,454,773	5,703	55,035,381
2004	1,936	40,701,220	2,686	8,262,664	933	8,533,980	5,555	57,497,864
2005	1,845	40,491,527	2,582	8,311,535	915	9,109,953	5,342	57,913,014
2006	1,759	40,382,967	2,584	8,320,514	999	9,476,470	5,342	58,179,951
Total	148,586	1,267,641,009	163,257	177,256,628	19,212	125,037,983	331,055	1,569,935,621

SOURCE: Author's calculations based on underlying EPUF sample.

Table 2.
Earnings records contained in the underlying EPUF sample, affected by data cleaning or disclosure protection procedures, and included in final EPUF, 1951–2006

	Records from the underlying EPUF sample with positive earnings		Records affected by data cleaning or disclosure protection procedures[a]		Final EPUF		Final EPUF as a percentage of underlying EPUF sample	
Year	Records	Dollar amount	Records	Dollar amount	Records	Dollar amount	Records	Dollar amount
1951	579,071	1,182,038,005	4,405	5,920,425	574,666	1,176,121,621	99.24	99.50
1952	595,005	1,256,504,791	4,622	6,224,390	590,383	1,250,218,697	99.22	99.50
1953	605,891	1,321,673,609	4,583	6,344,760	601,308	1,315,308,988	99.24	99.52
1954	594,469	1,301,518,421	3,928	6,092,153	590,541	1,295,436,078	99.34	99.53
1955	650,393	1,540,292,673	4,520	7,229,102	645,873	1,533,057,873	99.31	99.53
1956	675,958	1,667,196,602	4,729	7,834,164	671,229	1,659,358,545	99.30	99.53
1957	706,274	1,775,031,770	4,667	8,039,862	701,607	1,766,986,216	99.34	99.55
1958	699,009	1,760,718,703	4,183	7,805,195	694,826	1,752,916,336	99.40	99.56
1959	714,773	1,973,721,356	4,220	8,587,334	710,553	1,965,128,948	99.41	99.56
1960	724,277	2,023,372,141	4,274	8,705,575	720,003	2,014,641,299	99.41	99.57
1961	727,019	2,046,121,645	4,195	8,640,337	722,824	2,037,456,281	99.42	99.58
1962	742,198	2,133,834,749	4,132	8,921,869	738,066	2,124,909,855	99.44	99.58
1963	754,582	2,194,781,542	4,268	9,084,723	750,314	2,185,708,897	99.43	99.59
1964	773,598	2,292,872,077	4,308	9,473,091	769,290	2,283,413,867	99.44	99.59
1965	804,466	2,418,879,156	4,630	9,958,893	799,836	2,408,907,420	99.42	99.59
1966	845,200	3,053,032,399	5,208	12,302,053	839,992	3,040,762,112	99.38	99.60
1967	864,648	3,201,085,410	5,348	12,685,609	859,300	3,188,408,570	99.38	99.60
1968	891,688	3,677,060,356	5,742	14,363,986	885,946	3,662,694,039	99.36	99.61
1969	920,804	3,924,915,106	6,188	15,149,938	914,616	3,909,791,660	99.33	99.61
1970	926,593	4,047,308,546	6,067	15,328,153	920,526	4,031,955,717	99.35	99.62

(Continued)

Table 2.
Earnings records contained in the underlying EPUF sample, affected by data cleaning or disclosure protection procedures, and included in final EPUF, 1951–2006—*Continued*

	Records from the underlying EPUF sample with positive earnings		Records affected by data cleaning or disclosure protection procedures[a]		Final EPUF		Final EPUF as a percentage of underlying EPUF sample	
Year	Records	Dollar amount	Records	Dollar amount	Records	Dollar amount	Records	Dollar amount
1971	928,927	4,154,580,909	6,021	15,663,640	922,906	4,138,931,362	99.35	99.62
1972	957,932	4,725,131,546	6,527	17,548,401	951,405	4,707,580,541	99.32	99.63
1973	995,014	5,499,708,261	7,322	20,067,203	987,692	5,479,673,083	99.26	99.64
1974	1,010,681	6,266,031,784	7,437	22,450,084	1,003,244	6,243,556,827	99.26	99.64
1975	1,000,671	6,560,822,942	6,782	23,024,338	993,889	6,537,771,640	99.32	99.65
1976	1,025,235	7,272,380,800	6,841	24,670,664	1,018,394	7,247,765,424	99.33	99.66
1977	1,057,528	8,034,161,719	7,282	26,519,217	1,050,246	8,007,612,706	99.31	99.67
1978	1,091,783	9,003,657,698	7,816	29,216,447	1,083,967	8,974,444,824	99.28	99.68
1979	1,117,921	10,568,459,651	7,568	32,867,855	1,110,353	10,535,550,984	99.32	99.69
1980	1,123,641	11,588,053,871	6,902	34,013,927	1,116,739	11,553,996,366	99.39	99.71
1981	1,124,468	12,808,231,847	6,447	34,021,207	1,118,021	12,774,215,295	99.43	99.73
1982	1,109,975	13,447,471,166	5,896	34,092,842	1,104,079	13,413,406,844	99.47	99.75
1983	1,120,926	14,320,140,280	5,723	34,592,659	1,115,203	14,285,581,480	99.49	99.76
1984	1,164,250	15,733,184,777	6,324	36,043,213	1,157,926	15,697,179,349	99.46	99.77
1985	1,199,486	16,954,192,478	6,719	37,584,399	1,192,767	16,916,577,414	99.44	99.78
1986	1,222,942	18,072,210,162	6,403	37,747,309	1,216,539	18,034,475,665	99.48	99.79
1987	1,253,504	19,277,082,505	6,644	37,870,651	1,246,860	19,239,275,056	99.47	99.80
1988	1,293,120	20,699,177,394	7,136	37,222,326	1,285,984	20,661,907,215	99.45	99.82
1989	1,317,740	22,114,192,632	7,383	38,269,803	1,310,357	22,075,919,050	99.44	99.83
1990	1,327,049	23,320,377,715	6,669	37,985,278	1,320,380	23,282,326,410	99.50	99.84
1991	1,321,141	23,947,887,306	5,979	36,251,255	1,315,162	23,911,705,384	99.55	99.85
1992	1,329,671	25,038,192,482	5,980	37,176,029	1,323,691	25,000,961,124	99.55	99.85
1993	1,350,606	26,020,626,627	6,261	38,369,376	1,344,345	25,982,355,871	99.54	99.85
1994	1,379,206	27,519,441,609	6,732	39,391,891	1,372,474	27,480,153,319	99.51	99.86
1995	1,401,604	28,817,889,800	6,607	40,871,699	1,394,997	28,777,048,663	99.53	99.86
1996	1,424,677	30,325,434,565	6,739	42,205,802	1,417,938	30,283,145,483	99.53	99.86
1997	1,451,322	32,381,811,355	6,847	45,581,174	1,444,475	32,336,383,309	99.53	99.86
1998	1,479,545	34,688,002,415	7,072	48,349,833	1,472,473	34,639,656,847	99.52	99.86
1999	1,503,546	36,837,645,411	6,972	51,460,925	1,496,574	36,786,136,938	99.54	99.86
2000	1,529,060	39,253,537,670	7,123	53,195,988	1,521,937	39,200,496,095	99.53	99.86
2001	1,531,311	40,822,309,702	6,660	54,120,585	1,524,651	40,767,753,758	99.57	99.87
2002	1,525,643	41,636,130,619	6,082	54,171,781	1,519,561	41,581,840,812	99.60	99.87
2003	1,526,341	42,646,073,822	5,703	55,035,381	1,520,638	42,590,915,589	99.63	99.87
2004	1,541,064	44,453,363,308	5,555	57,497,864	1,535,509	44,395,547,826	99.64	99.87
2005	1,555,944	46,181,182,273	5,342	57,913,014	1,550,602	46,123,343,357	99.66	99.87
2006	1,568,139	48,431,660,720	5,342	58,179,951	1,562,797	48,373,174,994	99.66	99.88
Total	60,657,529	864,212,398,878	331,055	1,569,935,621	60,326,474	862,641,549,923	99.45	99.82

SOURCE: Author's calculations based on underlying EPUF sample.

a. Includes records removed because of data cleaning and records with earnings values set zero for indivudals with earnings at age 14 or younger or at age 86 or older.

After all of the data cleaning and data disclosure procedures were applied, several steps were taken to evaluate the validity of the data contained in EPUF. A forthcoming Research and Statistics Note compares the data in the underlying sample and the final EPUF with the earnings estimates published by SSA in the *Annual Statistical Supplement to the Social Security Bulletin.*

Caveats on Using EPUF Data

Any user should be fully aware of three caveats on using the EPUF: (1) earnings data in EPUF are capped taxable Social Security earnings, (2) EPUF does not contain all of the information needed to calculate benefits accurately for everyone in the file, and (3) there may be some errors in the administrative data underlying EPUF.

Capped Taxable Social Security Earnings

As previously noted, earnings data in EPUF are limited to capped taxable Social Security earnings. The file excludes data for workers whose only earnings are from noncovered employment. Additionally, the file does not contain covered earnings above the taxable maximum.

Table 3 compares the number of workers covered under the Social Security program with all US workers. Although the percentage working in covered employment has increased dramatically over time—from 55 percent in 1939 to nearly 94 percent in 2006—6 percent of the US workforce in 2006 still worked in noncovered employment.

Chart 2 shows that the amount of covered earnings expressed as a percentage of all earnings in the economy increased from approximately 70 percent in 1950 to nearly 85 percent in 2006. This represents a large increase in the share of earnings covered under the program, but it also reveals that approximately 15 percent of earnings in 2006 were not in covered employment.

However, noncovered earnings account for only part of the earnings “missing” from EPUF. Chart 2 also shows taxable Social Security earnings and the capped taxable Social Security earnings measure used in EPUF. As a percentage of total earnings in the economy, EPUF’s capped taxable earnings ranges from around 55 percent in the early 1950s to 78 percent in 1986, then declines gradually to 70 percent by 2006.

The relatively large differences between covered and taxable earnings from 1951 through the mid-1970s stem from the low taxable maximum earnings amounts during those years. The jagged pattern of the differences results from ad hoc changes to the taxable maximum. Prior to the 1972 Social Security Amendments, the taxable maximum was set by statute. From 1937 to 1950, the taxable maximum was $3,000. The first increase in the taxable maximum, to $3,600, occurred in 1951, and it increased four more times through 1971. The 1972

Table 3.
Civilian workers covered by the Social Security system, selected years 1939–2006

Year	Paid civilian workers[a] (millions)	Workers in covered employment or self-employment: Number (millions)	Workers in covered employment or self-employment: As a percentage of paid civilian workers
1939	43.6	24.0	55.0
1944	51.2	30.8	60.2
1949	56.7	34.3	60.5
1955	62.8	51.8	82.5
1960	64.6	55.7	86.2
1965	71.6	62.7	87.6
1970	77.8	69.9	89.8
1975	86.0	77.9	90.6
1980	99.4	89.3	89.8
1985	107.7	100.0	92.9
1990	117.8	111.7	94.8
1991	117.1	110.3	94.2
1992	118.7	111.9	94.3
1993	121.3	114.6	94.5
1994	124.6	117.9	94.6
1995	125.0	118.1	94.5
1996	127.7	120.7	94.5
1997	130.6	123.4	94.5
1998	132.6	125.1	94.4
1999	134.6	127.0	94.4
2000	137.7	130.0	94.4
2001	136.1	128.2	94.1
2002	136.5	128.2	93.9
2003	138.4	129.9	93.9
2004	140.2	131.5	93.8
2005	142.8	133.8	93.7
2006	146.0	136.7	93.6

SOURCE: Unpublished data from SSA's Office of the Chief Actuary.

NOTE: Data for 1939, 1944, and 1949 are monthly averages; data for all other years are as of December.

[a] Includes wage-and-salary earners and the self-employed.

amendments provided an automatic annual increase in the taxable maximum proportional to the increase in the national average wage. The key point for EPUF users is that using different methodologies for increasing the taxable maximum has affected the number (and proportion) of workers with earnings at or above the taxable maximum. For example, in 1951, nearly 25 percent of workers with covered earnings had earnings equal to or greater than the taxable maximum. In 1960 and 1970, the percentages of workers with earnings at or above the taxable maximum were 28 percent and 26 percent, respectively. In 1980, the percentage dropped to 9 percent and by 2006, it had dropped even further, to 6 percent (SSA 2009, Table 4.B4).[24]

Chart 2 reveals that the earnings in EPUF do not account for a significant portion of the total earnings in the economy from 1951 through 2006. Thus, using EPUF to analyze work patterns for individuals with a mix of covered and noncovered earnings may produce inaccurate results. Suppose an individual started working in a noncovered job in 1945 that was redefined as covered employment in 1955. This individual's work history in the EPUF would begin in 1955, with no indication that he or she really started working in 1945. Another example is an individual who worked in covered employment during high school and college and subsequently worked in a job that was not covered. This would result in a covered work history that starts in the individual's early work years and stops shortly thereafter.

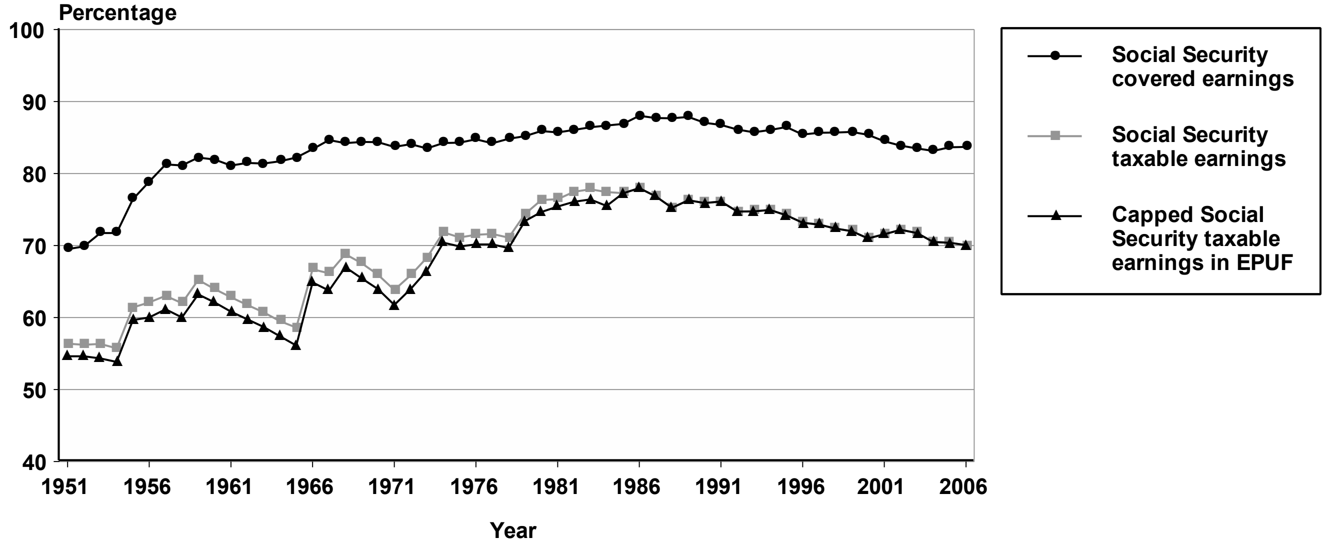
Limitations on Estimating Benefits

One expected use of EPUF is to evaluate how programmatic changes affect benefit amounts. However, such analysis is limited to estimating an individual's primary benefits; that is, benefits based on one's own earnings record. For example, auxiliary benefits—those to which individuals would be entitled based on their spouses' or parents' earnings record—cannot be estimated because there is no way to identify a spousal or parental link among individuals in EPUF.[25] This is problematic because many female beneficiaries receive part or all of their benefits based on a current or former spouse's higher earnings. Nevertheless, analysts can make reasoned assumptions about family size and estimate hypothetical family benefits based on an individual's own earnings records.

Analysts cannot use EPUF to estimate disability benefits because the file does not contain information about an individual's period(s) of disability. In addition, any calculation of retirement benefits for a disabled beneficiary would be inaccurate because it would exclude periods of disability. However, one can use EPUF to determine an individual's insured status in a given year and to estimate hypothetical disability benefits that could be awarded if an individual became disabled.

The EPUF does not contain a date of death for deceased individuals. As a result, one cannot determine if a string of years with zero earnings reflects that the individual has retired, become disabled, or died.

Chart 2.
Social Security earnings (weighted) as a percentage of all earnings



SOURCES: Bureau of Economic Analysis National Income and Product Account; SSA (2009a); 2006 EPUF.

The accuracy of estimates for primary benefits may be affected by the lack of detailed information for some individuals in the file. When calculating an individual's benefit amount, SSA uses the certified earnings record, which includes any ancillary earnings information such as military credits, railroad employment income, or having multiple SSNs.[26] Because EPUF omits this information, estimates of benefits for individuals who had these sources of income or had multiple SSNs are suspect. Although the number of individuals having multiple SSNs or railroad income is relatively small, accurate assessments of the effects of programmatic changes on these individuals would require such information. The number of individuals with military credits is likely to be much larger, but the impact on benefits is likely to be relatively small for those with limited military service.

Incomplete information in the EPUF also hinders accurate estimates of benefits for individuals with earnings during 1937–1950. Recall that SSA had to estimate the number of QCs associated with earnings from this period. Consider an individual who applies for benefits but is a couple of quarters short of being eligible. In such a case, SSA reviews the microfilm record to determine the individual's actual amount of covered earnings during the period. SSA posts this amount to the detailed segment of the MEF then determines the QCs earned using the usual procedures. However, EPUF does not include the information from the microfilm. Therefore, analysts should exercise caution when using EPUF data on QCs for this period, and should note this fact in any analysis using that data field.

The user should also note that precise computation of monthly benefits paid is not possible with the EPUF because age at entitlement, on which monthly benefit amounts are based, cannot be observed in the file. With EPUF, it is also not possible to adjust benefits for workers subject to the Windfall Elimination Provision, which reduces benefits of "individuals who have only minimal Social Security coverage and will receive a pension based on years of work in noncovered employment" (SSA 2009).

Errors in Underlying Earnings Data

SSA has been collecting data on individual workers covered under the program since its inception. The agency uses administrative files to determine eligibility for benefits, to determine benefit amounts, to estimate future benefit payments, and for a variety of other purposes.

Each year, capturing the earnings data reported on Form W-2 and used for program purposes is a massive undertaking. For earnings reported in tax year 2006, SSA processed W-2s for nearly 155 million workers and generated approximately 250 million wage items. SSA processed nearly 80 percent of the wage items reported on the W-2s electronically, and the remaining 20 percent were scanned using character recognition software or keyed in manually. In addition, SSA received information on self-employment income from the IRS based on data reported on Schedule SE. This information accounted for approximately 20 million items posted to the MEF. In total, SSA posted nearly 270 million earnings-related items for tax year 2006 to its MEF.

With so many items posted every year, the MEF is clearly susceptible to missing or erroneous earnings data. Each step of the process introduces potential errors. The employer may enter an incorrect amount for a given individual, or may put the correct information in the wrong box on the W-2. In addition, the SSN may not be valid or the name on the W-2 may not match the one to which the SSN was enumerated.[27] Errors can also arise as SSA posts the data in the MEF.

SSA has an elaborate set of checks to identify and correct improperly reported earnings information.[28] The agency verifies that the information on all the W-2s submitted by an employer corresponds to the amounts reported by the employer on Form W-3. SSA continuously updates the MEF as corrected W-2s (W-2c's) and delinquent W-2s stream in throughout the year. Workers may also file amended tax returns to correct errors reported in previous filings.

If SSA detects errors in a worker's earning record, it sends a letter to the employer seeking clarification. In response, the employer may file a W-2c. In some instances, an employer files a W-2c and the employee supplies information to correct the same error; the resulting double-correction also produces errors on the MEF.

Another opportunity to catch earnings-record errors arises when SSA mails out its annual Social Security statement to workers aged 25 or older. Errors detected by the worker can be resolved at any SSA field office.[29] Finally, workers can catch errors in their earnings data when they apply for benefits. Applicants see their complete earnings histories and can direct SSA to correct any verifiable errors they spot. Nevertheless, despite extensive efforts to ensure accurate earnings records, the EPUF may contain erroneous information.

Highlights from the EPUF

This section presents statistical highlights of the earnings data for the 4,384,254 individuals whose records are included in EPUF. Figures cited are unweighted.

Individuals by YOB

There are five distinct trends in the distribution of individuals by birth year in EPUF (Chart 3). The first is a steep increase in the number of individuals in the file, starting with 1,813 born in 1870 and peaking at 31,877 born in 1921. The second is a steady decline from 31,104 born in 1922 to 26,568 in 1933. The third trend is a dramatic increase to nearly 53,000 who were born in 1962, nearly doubling the number of individuals born in 1933. The fourth is a steep decline from 52,138 individuals born in 1963 to 41,792 born in 1975. The final trend reflects relatively flat numbers of individuals born from 1976 through 2006, from 41,822 to 41,241, respectively.

Chart 4 presents the distribution of individuals by YOB and sex.[30] For birth years from 1870 to about 1925, men outnumber women in EPUF. With a few exceptions, the numbers of women and men in the file are nearly the same for birth years from 1926 to 1947. The number of men born from 1948 to 2006 is consistently higher than the number of women, although not by very much.

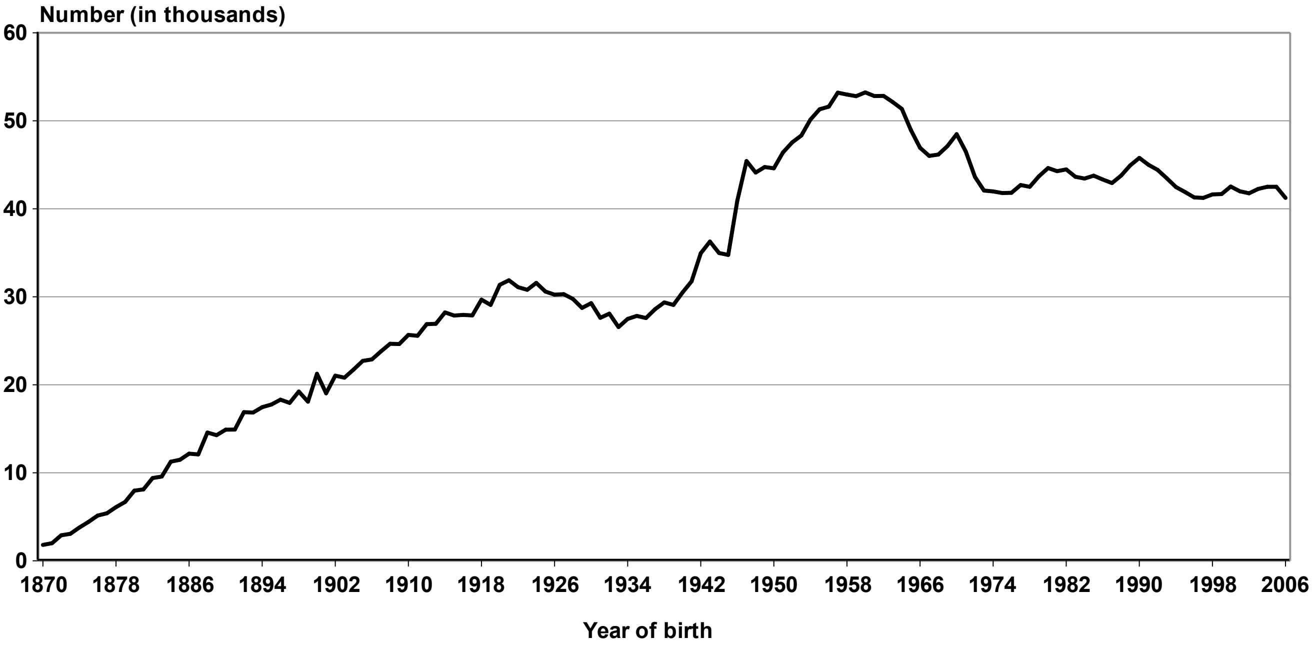
Workers and Nonworkers

There are four distinct categories of individuals in EPUF depending on whether they had any Social Security taxable earnings and, if so, the period in which they were earned. The four categories are nonworkers (individuals with no taxable earnings), workers with taxable earnings during 1937–1950 only, workers with taxable earnings during 1951–2006 only, and workers with taxable earnings in both periods. More than one-half of the individuals in EPUF had earnings during 1951–2006 only, about 4 percent had earnings only during 1937–1950, and 16 percent had earnings in both periods (Chart 5).

Initially, the 24.7-percent figure for individuals in EPUF who did not have any earnings seems very large. However, Chart 6 reveals that the bulk of these individuals (68 percent) were born after 1987. Thus, the main reason so many individuals in EPUF have no earnings is that most of them are not old enough to participate in the labor market.[31]

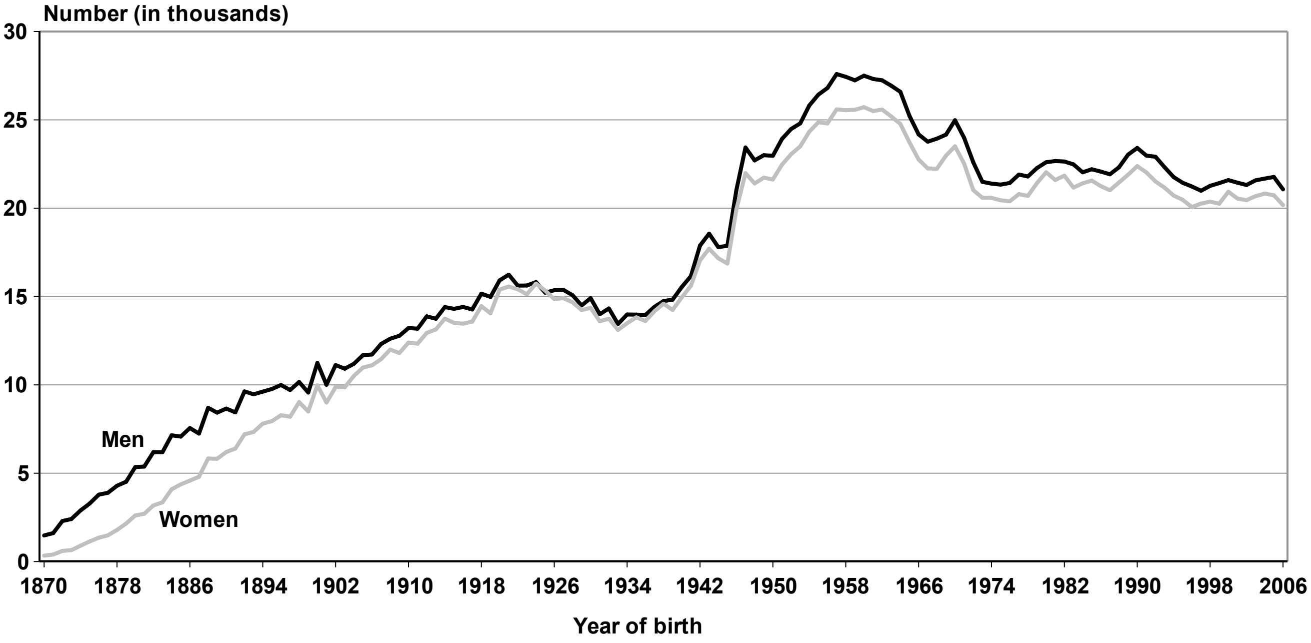
Chart 7 presents the distribution by sex of individuals in EPUF in each earner status. Women outnumber men among those who do not have any earnings

Chart 3.
Number of individuals in EPUF, by year of birth



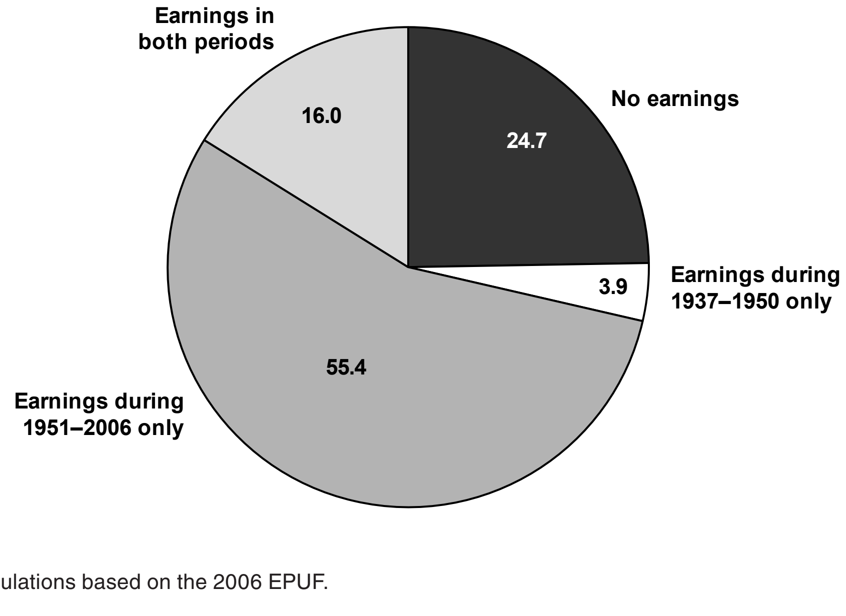
SOURCE: Author's calculations based on the 2006 EPUF.

Chart 4.
Number of individuals in EPUF, by year of birth and sex



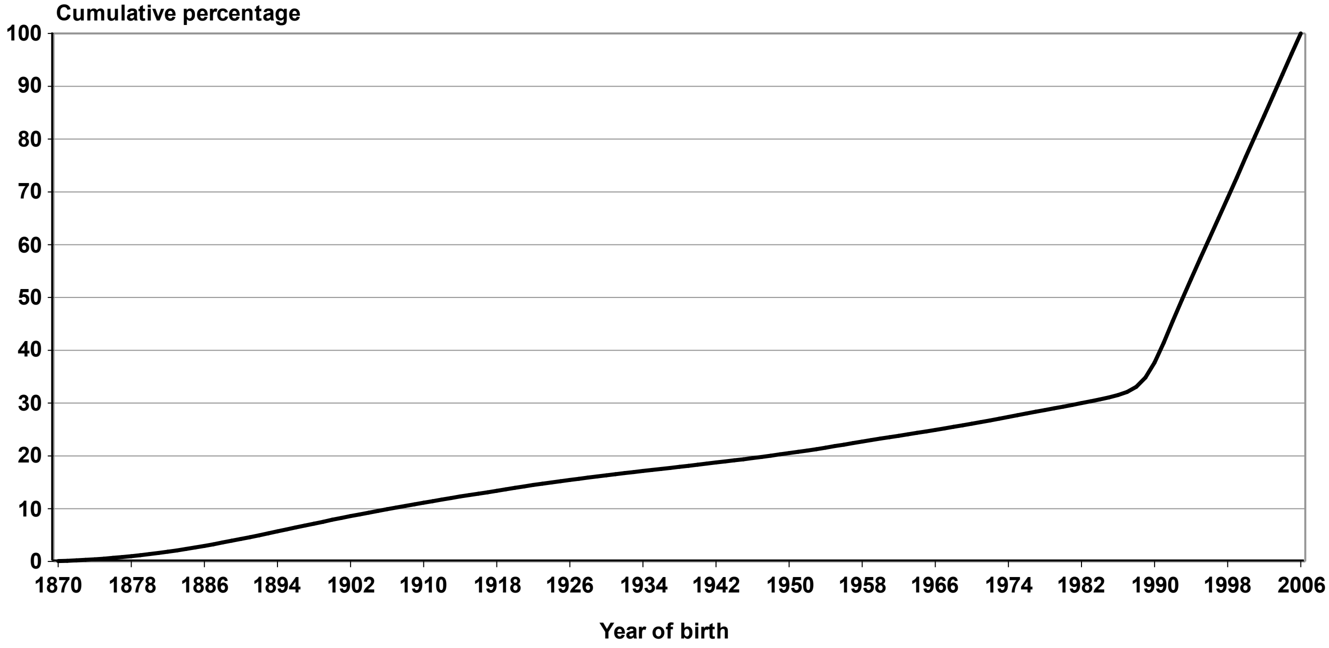
SOURCE: Author's calculations based on the 2006 EPUF.

Chart 5.
Percentage distribution of individuals in EPUF, by capped Social Security taxable earnings status



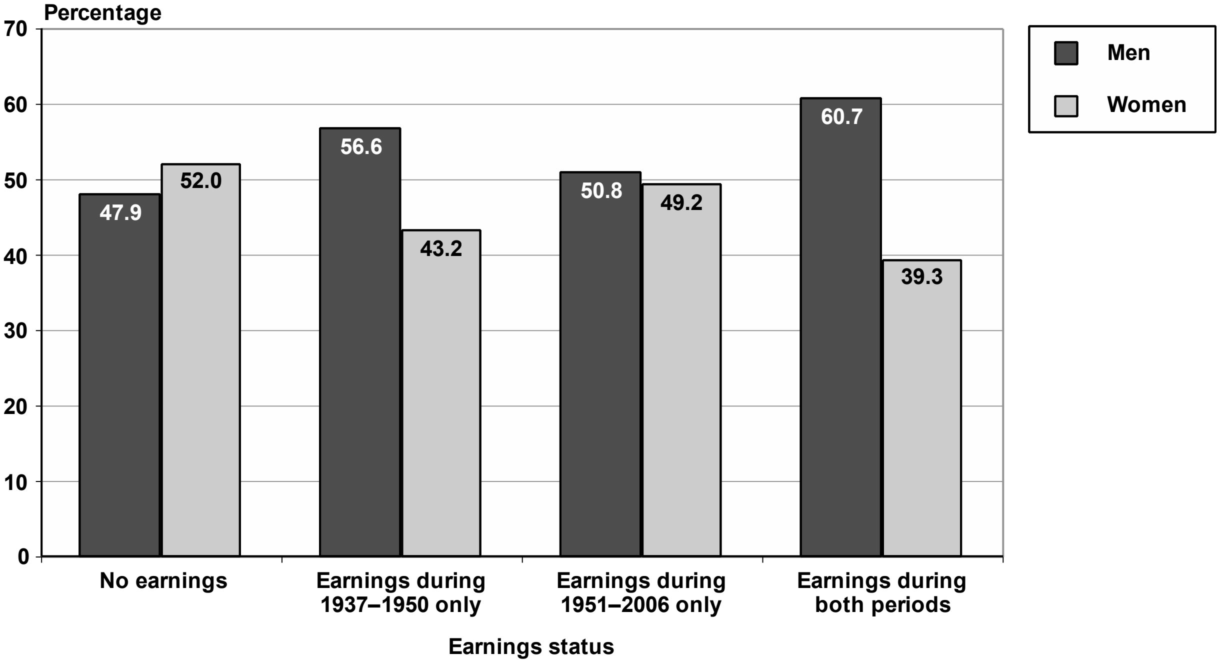
SOURCE: Author's calculations based on the 2006 EPUF.

Chart 6.
Cumulative distribution of individuals in EPUF with no capped Social Security taxable earnings, by year of birth



SOURCE: Author's calculations based on the 2006 EPUF.

Chart 7.
Percentage distribution of individuals in EPUF in each capped Social Security taxable earnings status, by sex



SOURCE: Author's calculations based on the 2006 EPUF.

NOTE: Rounded components of percentage distributions do not necessarily sum to 100.

(52 percent versus 48 percent). Among individuals with earnings during 1937–1950 only, a large majority are men (57 percent versus 43 percent). This result was expected because women were much less active in the labor market during that period. Individuals in EPUF with earnings during both periods skew even more towards men, 61 percent versus 39 percent. Individuals with earnings during 1951–2006 only are more evenly distributed between men (51 percent) and women (49 percent), reflecting women's substantial increases in labor force participation during the period.

Individuals in EPUF with any earnings during 1937–1950 number 874,287. Approximately 60 percent are men (523,465) and 40 percent are women (350,229). There are also records for 593 individuals whose sex is unknown and who had earnings during this period. Appendix Chart A1 presents the distribution of individuals with earnings during this period by YOB and sex. The average and median values for all earnings during this period are $9,106 and $4,600, respectively (not shown). The average earnings for men ($11,990) is much higher than that for women ($7,521). The median earnings for men and women diverge even more, at $7,900 and $1,800, respectively.

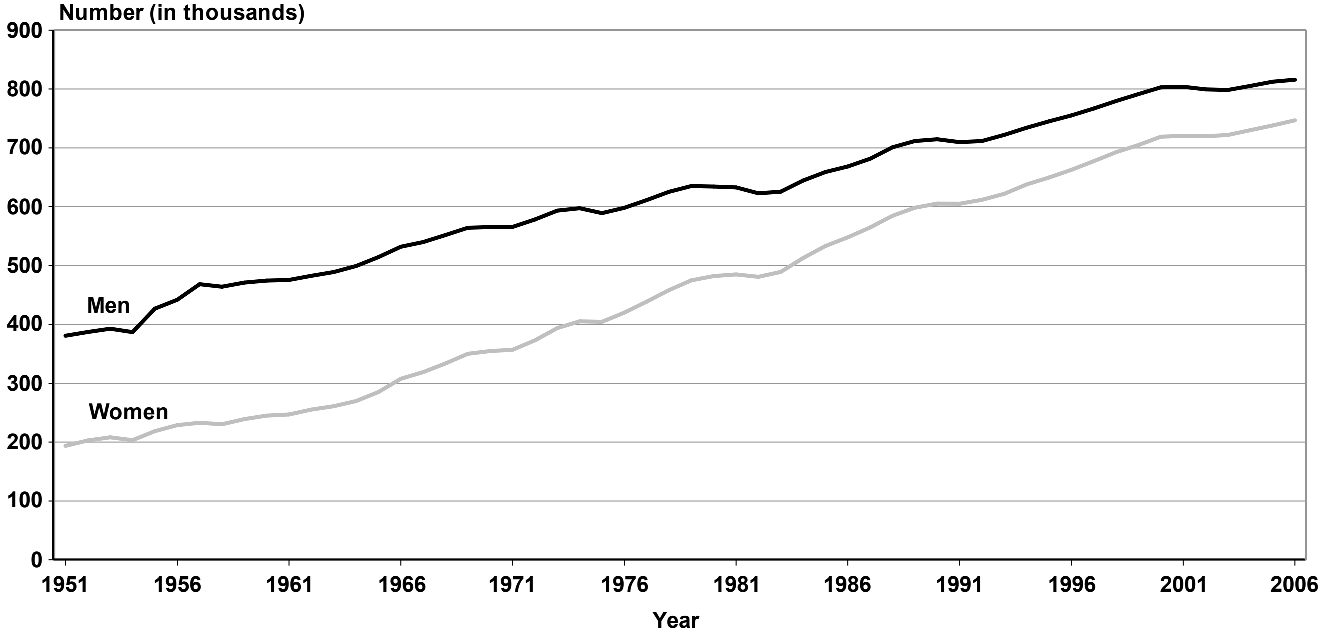
Earnings in EPUF

Chart 8 shows that the gap between the number of men and women with earnings in a given year has decreased significantly between 1951 and 2006. Chart 9 shows a slow but steady climb in aggregate earnings for men and women over the same period.[32] The difference between the total amount of earnings for men and women has been increasing over time. However, women's taxable earnings as a percentage of all taxable earnings has increased from 22.1 percent in 1951 to 39.7 percent in 2006 (see Table A2). Table 4 presents the average and median earnings of men, women, and individuals with unknown sex in the EPUF.

Summary

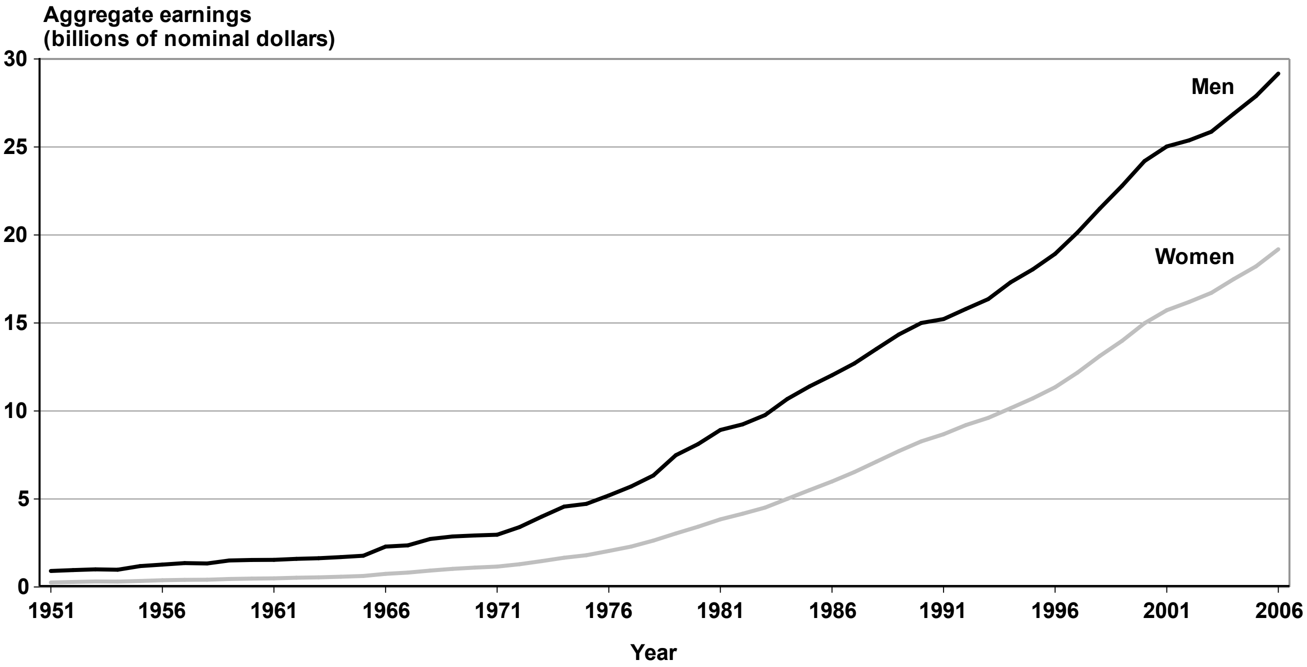
The 2006 EPUF contains earnings data for individuals drawn from a 1-percent sample of all SSNs issued before January 2007. The file contains limited demographic information and earnings data related to the Social Security program for 4,348,254 individuals. Although the file contains limited data fields, it is much larger than other public-use files with earnings histories. EPUF will provide policymakers and researchers with a unique tool to evaluate the Social Security programs and potential reforms.

Chart 8.
Number of individuals with capped Social Security taxable earnings in EPUF, by sex, 1951–2006



SOURCE: Author's calculations based on the 2006 EPUF.

Chart 9.
Aggregate amount of capped Social Security taxable earnings in EPUF, by sex of earner, 1951–2006



SOURCE: Author's calculations based on the 2006 EPUF.

Table 4.
Average and median Social Security taxable earnings in EPUF, by sex, 1951–2006 (in dollars)

	All workers		Men		Women		Sex unknown	
Year	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1951	2,047	2,100	2,404	2,900	1,344	1,200	1,978	1,950
1952	2,118	2,300	2,482	3,100	1,423	1,300	1,904	1,950
1953	2,187	2,400	2,553	3,300	1,499	1,300	2,043	2,000
1954	2,194	2,400	2,544	3,300	1,527	1,400	1,886	1,700
1955	2,374	2,400	2,779	3,300	1,583	1,300	1,932	1,600
1956	2,472	2,600	2,884	3,500	1,678	1,500	1,897	1,400
1957	2,518	2,700	2,900	3,600	1,752	1,500	1,874	1,450
1958	2,523	2,700	2,881	3,500	1,801	1,600	1,914	1,500
1959	2,766	2,800	3,204	3,800	1,903	1,600	2,040	1,600
1960	2,798	2,900	3,239	3,900	1,945	1,700	2,161	1,800
1961	2,819	2,900	3,248	3,900	1,994	1,700	2,190	1,800
1962	2,879	3,100	3,313	4,100	2,059	1,800	2,439	2,100
1963	2,913	3,100	3,345	4,300	2,104	1,800	2,534	2,200
1964	2,968	3,300	3,402	4,500	2,166	1,900	2,717	2,700
1965	3,012	3,400	3,459	4,754	2,206	2,000	2,871	2,900
1966	3,620	3,600	4,312	5,000	2,424	2,000	3,548	3,300
1967	3,710	3,700	4,380	5,200	2,576	2,200	3,580	3,500
1968	4,134	4,000	4,944	5,600	2,796	2,400	4,099	3,800
1969	4,275	4,200	5,093	6,000	2,956	2,600	4,090	3,900
1970	4,380	4,400	5,175	6,200	3,113	2,700	4,214	4,200

(Continued)

Table 4.
Average and median Social Security taxable earnings in EPUF, by sex, 1951–2006 (in dollars)—*Continued*

	All workers		Men		Women		Sex unknown	
Year	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1971	4,485	4,600	5,259	6,500	3,257	2,900	4,409	4,650
1972	4,948	4,900	5,893	7,000	3,482	3,000	5,067	5,100
1973	5,548	5,200	6,744	7,500	3,745	3,100	5,776	5,450
1974	6,223	5,500	7,653	8,000	4,115	3,400	6,572	5,950
1975	6,578	5,800	8,023	8,300	4,473	3,700	6,854	6,400
1976	7,117	6,300	8,693	8,900	4,870	4,100	7,806	6,950
1977	7,625	6,700	9,343	9,600	5,229	4,300	8,384	7,500
1978	8,279	7,300	10,132	10,400	5,750	4,900	8,816	8,200
1979	9,488	7,900	11,790	11,300	6,410	5,400	10,383	9,400
1980	10,346	8,600	12,804	12,000	7,112	6,000	11,052	9,600
1981	11,426	9,400	14,106	13,000	7,927	6,700	12,280	10,300
1982	12,149	9,900	14,844	13,300	8,659	7,200	13,362	10,950
1983	12,810	10,300	15,613	13,700	9,224	7,600	13,917	11,400
1984	13,556	10,900	16,572	14,500	9,766	7,900	14,365	11,300
1985	14,183	11,400	17,303	15,100	10,325	8,300	15,037	12,000
1986	14,824	11,900	18,002	15,600	10,948	8,800	16,204	12,950
1987	15,430	12,300	18,636	16,100	11,560	9,300	16,257	12,700
1988	16,067	12,900	19,304	16,600	12,186	9,800	16,943	13,250
1989	16,847	13,400	20,163	17,200	12,902	10,300	18,780	15,100
1990	17,633	14,000	20,990	17,800	13,669	10,900	19,937	16,500
1991	18,182	14,400	21,455	17,900	14,342	11,400	21,272	18,250
1992	18,887	14,900	22,199	18,400	15,034	11,900	22,037	18,800
1993	19,327	15,100	22,662	18,700	15,455	12,100	22,841	20,150
1994	20,022	15,600	23,576	19,400	15,930	12,400	23,137	18,700
1995	20,629	16,200	24,224	20,000	16,504	12,900	23,835	19,100
1996	21,357	16,800	25,066	20,800	17,129	13,400	24,950	20,200
1997	22,386	17,600	26,274	21,900	17,985	14,100	24,905	20,500
1998	23,525	18,600	27,582	23,100	18,958	14,900	26,290	21,800
1999	24,580	19,400	28,802	24,100	19,840	15,600	28,665	23,700
2000	25,757	20,300	30,149	25,200	20,851	16,400	30,797	25,450
2001	26,739	21,000	31,141	25,700	21,826	17,100	32,920	26,500
2002	27,364	21,300	31,743	25,900	22,499	17,500	33,686	26,800
2003	28,009	21,700	32,396	26,300	23,154	18,000	34,133	29,900
2004	28,913	22,500	33,396	27,200	23,965	18,500	34,542	29,100
2005	29,745	23,100	34,341	28,000	24,685	19,000	36,241	30,600
2006	30,953	24,000	35,764	29,100	25,696	19,700	36,799	32,500

SOURCE: Author's calculations based on the 2006 EPUF.

Appendix

Table A1.
Number and percentage distribution of individuals with Social Security taxable earnings records in EPUF, by sex, 1951–2006

Year	All workers	Men		Women		Sex unknown	
		Number	Percentage of workers	Number	Percentage of workers	Number	Percentage +I41of workers
1951	574,666	380,673	66.2	193,655	33.7	338	0.1
1952	590,383	387,176	65.6	202,841	34.4	366	0.1
1953	601,308	392,710	65.3	208,254	34.6	344	0.1
1954	590,541	386,904	65.5	203,317	34.4	320	0.1
1955	645,873	426,862	66.1	218,624	33.8	387	0.1
1956	671,229	441,870	65.8	228,933	34.1	426	0.1
1957	701,607	468,328	66.8	232,861	33.2	418	0.1
1958	694,826	464,175	66.8	230,290	33.1	361	0.1
1959	710,553	471,169	66.3	239,044	33.6	340	a
1960	720,003	474,604	65.9	245,085	34.0	314	a
1961	722,824	475,513	65.8	247,008	34.2	303	a
1962	738,066	482,590	65.4	255,187	34.6	289	a
1963	750,314	488,952	65.2	261,077	34.8	285	a
1964	769,290	499,171	64.9	269,834	35.1	285	a
1965	799,836	514,368	64.3	285,184	35.7	284	a
1966	839,992	531,966	63.3	307,743	36.6	283	a
1967	859,300	540,003	62.8	319,006	37.1	291	a
1968	885,946	551,920	62.3	333,731	37.7	295	a
1969	914,616	564,231	61.7	350,067	38.3	318	a
1970	920,526	565,453	61.4	354,749	38.5	324	a
1971	922,906	565,675	61.3	356,911	38.7	320	a
1972	951,405	578,237	60.8	372,840	39.2	328	a
1973	987,692	593,494	60.1	393,844	39.9	354	a
1974	1,003,244	597,517	59.6	405,375	40.4	352	a
1975	993,889	589,138	59.3	404,403	40.7	348	a
1976	1,018,394	598,171	58.7	419,885	41.2	338	a
1977	1,050,246	611,288	58.2	438,619	41.8	339	a
1978	1,083,967	625,380	57.7	458,246	42.3	341	a
1979	1,110,353	635,128	57.2	474,898	42.8	327	a
1980	1,116,739	634,313	56.8	482,099	43.2	327	a
1981	1,118,021	632,816	56.6	484,894	43.4	311	a
1982	1,104,079	622,799	56.4	480,974	43.6	306	a
1983	1,115,203	625,683	56.1	489,213	43.9	307	a
1984	1,157,926	644,631	55.7	512,978	44.3	317	a
1985	1,192,767	659,120	55.3	533,338	44.7	309	a
1986	1,216,539	668,310	54.9	547,925	45.0	304	a
1987	1,246,860	681,710	54.7	564,843	45.3	307	a
1988	1,285,984	700,961	54.5	584,711	45.5	312	a
1989	1,310,357	711,727	54.3	598,334	45.7	296	a
1990	1,320,380	714,671	54.1	605,422	45.9	287	a
1991	1,315,162	709,678	54.0	605,204	46.0	280	a
1992	1,323,691	711,615	53.8	611,804	46.2	272	a
1993	1,344,345	722,012	53.7	622,065	46.3	268	a
1994	1,372,474	734,324	53.5	637,884	46.5	266	a
1995	1,394,997	745,091	53.4	649,650	46.6	256	a

(Continued)

Table A1.
Number and percentage distribution of individuals with Social Security taxable earnings records in EPUF, by sex, 1951–2006—*Continued*

		Men		Women		Sex unknown	
Year	All workers	Number	Percentage of workers	Number	Percentage of workers	Number	Percentage of workers
1996	1,417,938	755,129	53.3	662,564	46.7	245	a
1997	1,444,475	766,814	53.1	677,412	46.9	249	a
1998	1,472,473	779,589	52.9	692,640	47.0	244	a
1999	1,496,574	791,384	52.9	704,947	47.1	243	a
2000	1,521,937	802,776	52.7	718,923	47.2	238	a
2001	1,524,651	803,891	52.7	720,525	47.3	235	a
2002	1,519,561	799,527	52.6	719,799	47.4	235	a
2003	1,520,638	798,428	52.5	721,985	47.5	225	a
2004	1,535,509	805,264	52.4	730,008	47.5	237	a
2005	1,550,602	812,364	52.4	738,007	47.6	231	a
2006	1,562,797	815,763	52.2	746,806	47.8	228	a
Total	60,326,474	34,553,056	57.3	25,756,465	42.7	16,953	a

SOURCE: Author's calculations based on the 2006 EPUF.

NOTE: Rounded components of percentage distributions do not necessarily sum to 100.

a. Less than 0.05 percent

Table A2.
Dollar amount and percentage distribution of Social Security taxable earnings in EPUF, by sex of earner, 1951–2006

		Men		Women		Sex unknown	
Year	Total Social Security taxable earnings ($)	Dollar amount	Percentage of earnings	Dollar amount	Percentage of earnings	Dollar amount	Percentage of earnings
1951	1,176,121,621	915,224,528	77.8	260,228,626	22.1	668,467	0.1
1952	1,250,218,697	960,951,736	76.9	288,570,223	23.1	696,739	0.1
1953	1,315,308,988	1,002,401,906	76.2	312,204,394	23.7	702,689	0.1
1954	1,295,436,078	984,354,316	76.0	310,478,153	24.0	603,609	a
1955	1,533,057,873	1,186,138,562	77.4	346,171,624	22.6	747,687	a
1956	1,659,358,545	1,274,501,991	76.8	384,048,272	23.1	808,282	a
1957	1,766,986,216	1,358,130,591	76.9	408,072,212	23.1	783,413	a
1958	1,752,916,336	1,337,517,626	76.3	414,707,883	23.7	690,827	a
1959	1,965,128,948	1,509,520,746	76.8	454,914,691	23.1	693,511	a
1960	2,014,641,300	1,537,199,839	76.3	476,762,888	23.7	678,572	a
1961	2,037,456,281	1,544,306,338	75.8	492,486,504	24.2	663,439	a
1962	2,124,909,855	1,598,792,149	75.2	525,412,919	24.7	704,788	a
1963	2,185,708,897	1,635,775,204	74.8	549,211,363	25.1	722,331	a
1964	2,283,413,867	1,698,087,380	74.4	584,552,087	25.6	774,400	a
1965	2,408,907,420	1,779,058,958	73.9	629,033,153	26.1	815,308	a
1966	3,040,762,112	2,293,932,086	75.4	745,826,027	24.5	1,003,999	a
1967	3,188,408,570	2,365,472,074	74.2	821,894,805	25.8	1,041,691	a
1968	3,662,694,039	2,728,439,824	74.5	933,045,098	25.5	1,209,116	a
1969	3,909,791,660	2,873,795,305	73.5	1,034,695,870	26.5	1,300,485	a
1970	4,031,955,717	2,926,141,698	72.6	1,104,448,529	27.4	1,365,490	a

(Continued)

Table A2.
Dollar amount and percentage distribution of Social Security taxable earnings in EPUF, by sex of earner, 1951–2006—*Continued*

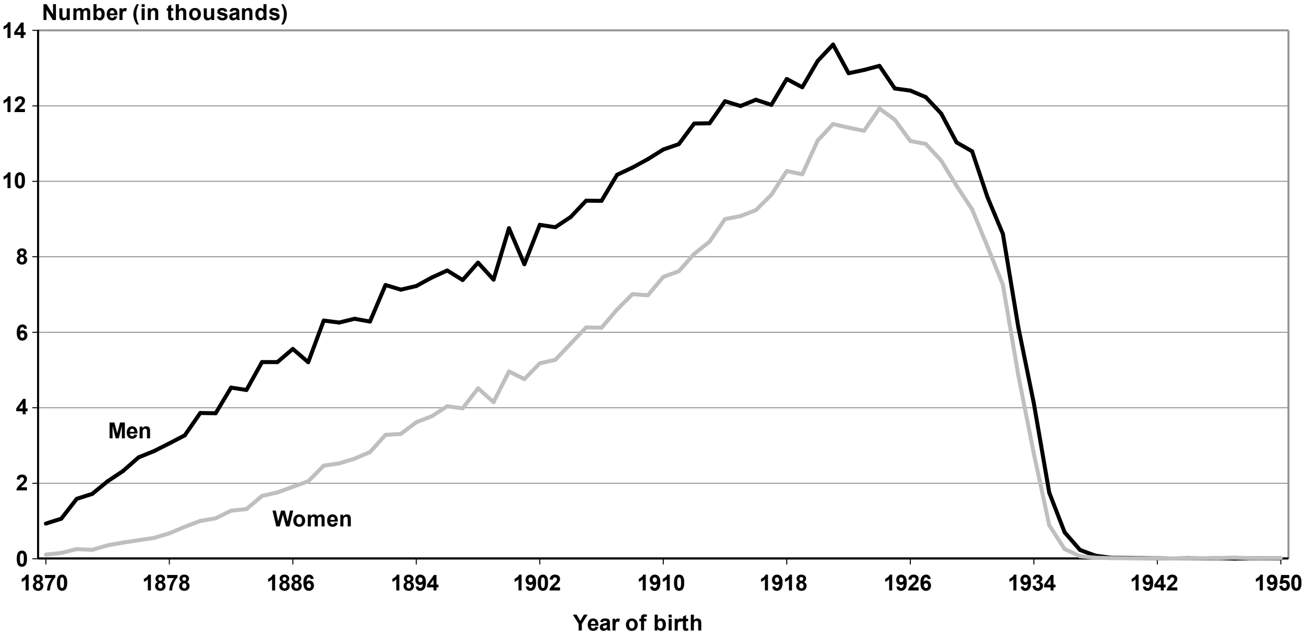
Year	Total Social Security taxable earnings ($)	Men		Women		Sex unknown	
		Dollar amount	Percentage of earnings	Dollar amount	Percentage of earnings	Dollar amount	Percentage of earnings
1971	4,138,931,362	2,975,093,757	71.9	1,162,426,686	28.1	1,410,920	a
1972	4,707,580,541	3,407,572,244	72.4	1,298,346,419	27.6	1,661,877	a
1973	5,479,673,083	4,002,814,306	73.0	1,474,814,111	26.9	2,044,666	a
1974	6,243,556,827	4,573,069,433	73.2	1,668,173,966	26.7	2,313,428	a
1975	6,537,771,640	4,726,502,691	72.3	1,808,883,849	27.7	2,385,100	a
1976	7,247,765,424	5,200,093,565	71.7	2,045,033,268	28.2	2,638,591	a
1977	8,007,612,706	5,711,058,117	71.3	2,293,712,581	28.6	2,842,008	a
1978	8,974,444,824	6,336,610,720	70.6	2,634,827,857	29.4	3,006,247	a
1979	10,535,550,984	7,488,124,641	71.1	3,044,030,994	28.9	3,395,348	a
1980	11,553,996,366	8,121,707,068	70.3	3,428,675,206	29.7	3,614,092	a
1981	12,774,215,295	8,926,653,455	69.9	3,843,742,790	30.1	3,819,050	a
1982	13,413,406,844	9,244,523,741	68.9	4,164,794,202	31.0	4,088,900	a
1983	14,285,581,480	9,768,685,099	68.4	4,512,623,779	31.6	4,272,602	a
1984	15,697,179,349	10,682,698,768	68.1	5,009,926,755	31.9	4,553,826	a
1985	16,916,577,414	11,405,063,900	67.4	5,506,867,114	32.6	4,646,400	a
1986	18,034,475,665	12,031,058,617	66.7	5,998,490,972	33.3	4,926,076	a
1987	19,239,275,056	12,704,633,159	66.0	6,529,650,970	33.9	4,990,927	a
1988	20,661,907,215	13,531,532,531	65.5	7,125,088,330	34.5	5,286,353	a
1989	22,075,919,050	14,350,553,706	65.0	7,719,806,564	35.0	5,558,780	a
1990	23,282,326,410	15,001,089,862	64.4	8,275,514,770	35.5	5,721,778	a
1991	23,911,705,385	15,225,958,913	63.7	8,679,790,242	36.3	5,956,230	a
1992	25,000,961,124	15,797,304,161	63.2	9,197,663,036	36.8	5,993,927	a
1993	25,982,355,871	16,362,219,545	63.0	9,614,015,049	37.0	6,121,278	a
1994	27,480,153,319	17,312,328,296	63.0	10,161,670,536	37.0	6,154,487	a
1995	28,777,048,662	18,048,809,034	62.7	10,722,137,749	37.3	6,101,879	a
1996	30,283,145,482	18,928,028,662	62.5	11,349,003,953	37.5	6,112,867	a
1997	32,336,383,309	20,147,226,145	62.3	12,182,955,785	37.7	6,201,379	a
1998	34,639,656,847	21,502,279,695	62.1	13,130,962,491	37.9	6,414,661	a
1999	36,786,136,937	22,793,062,944	62.0	13,986,108,356	38.0	6,965,637	a
2000	39,200,496,095	24,202,981,172	61.7	14,990,185,170	38.2	7,329,753	a
2001	40,767,753,758	25,034,170,600	61.4	15,725,846,857	38.6	7,736,301	a
2002	41,581,840,812	25,379,005,293	61.0	16,194,919,207	38.9	7,916,312	a
2003	42,590,915,589	25,866,065,725	60.7	16,717,169,849	39.3	7,680,015	a
2004	44,395,547,826	26,892,533,184	60.6	17,494,828,170	39.4	8,186,472	a
2005	46,123,343,358	27,897,078,736	60.5	18,217,892,871	39.5	8,371,751	a
2006	48,373,174,994	29,174,561,654	60.3	19,190,223,246	39.7	8,390,094	a
Total	862,641,549,921	554,262,495,993	64.3	308,177,569,073	35.7	201,484,854	a

SOURCE: Author's calculations based on the 2006 EPUF.

NOTE: Rounded components of percentage distributions do not necessarily sum to 100.

a. Less than 0.05 percent

Chart A1.
Number of individuals in EPUF with capped Social Security taxable earnings during 1937–1950, by year of birth and sex



SOURCE: Author's calculations based on the 2006 EPUF.

Notes

Acknowledgments: The author gratefully acknowledges the assistance of many individuals in the process of creating the 2006 Earnings Public-Use File and this article: John Hennessey, for graciously sharing his programming and methodological expertise; Russell Hudson, for his programming expertise and sharing his vast knowledge of the earnings data; Paul Davies, for his guidance and support throughout the project; Scott Muller, Greg Diez, and Bill Piet, for sharing programmatic and earnings knowledge; Sirisha Anne, Brenda South, Stu Friedrich, and Randall Miles, for their assistance in providing the data extracts used in the process of creating EPUF; Bill Davis and Justin Ronca, for their statistical expertise; and Susan Grad, Howard Iams, Hilary Waldron, and Anya Olsen, for their comments on previous drafts of the article.

[1] The MEF contains all of the earnings data collected to administer the Social Security programs.

[2] Noncovered earnings are wage and salary income not covered under the Social Security programs.

[3] For a discussion of SSA earnings data, see Olsen and Hudson (2009).

[4] This limitation is discussed later in the article.

[5] For historical changes in coverage, see SSA (2009, Table 2.A1).

[6] SSA's Office of Research, Evaluation, and Statistics uses this measure to generate its published estimates of earnings.

[7] Technically, this is not always correct because some earnings are reported on the Earnings Suspense File and not posted on the MEF. For a detailed discussion, see GAO (2005).

[8] The average wage index is calculated annually using wages subject to federal income taxes and contributions to deferred compensation plans. The index is used in determining an individual's retirement benefit amount as well as to determine several other key dollar amounts in the administration of the Social Security programs. For more detail, see SSA (2010).

[9] This process is done because of the prohibitive costs associated with going back to the microfilm to determine the exact number of QCs earned by individuals with earnings during the 1937–1946 period.

[10] For individuals with earnings during this period who did not meet program criteria for benefits or coverage (using this technique to estimate QCs), a detailed manual search of microfilm records determines if the individual was eligible for benefits and, if so, the benefit amount.

[11] Including these flags would have created serious data disclosure problems because they provide much more individually identifiable information.

[12] For a detailed discussion of deferred earnings in SSA data, see Pattison and Waldron (2008).

[13] For a description of the three components of the SSN (area, group, and serial number), see Puckett (2009).

[14] Nonoverlapping samples are important from a data disclosure perspective if SSA decides to release any additional public-use data files.

[15] The sample design is equal to the ratio of the variance of the systematic random sample for EPUF and the variance assuming a simple random sample without replacement.

[16] The Numident is a master file of all SSNs ever assigned. It contains the identifying information given when an individual applies for an SSN.

[17] This includes 319 individuals who were ultimately removed from the underlying EPUF sample because they were also in the New Beneficiary Data Systems (discussed in the data disclosure section of the article).

[18] The source for YOB data in EPUF is the MEF summary record, which may not contain the same value that appears in the Numident or Master Beneficiary Record files.

[19] See the text box for a brief description of the other public-use data files that contain earnings data from Social Security administrative files. To evaluate the disclosure risk for individuals in EPUF who are included in other publicly available data files, SSA considers four key points: the potential magnitude of the overlap between files, the possibility of matching records across files with any certainty, the additional information that would be revealed in the unlikely event that records could be matched with any certainty, and the ability to reidentify someone in EPUF based on publicly available data.

[20] Thus, the total number of individuals removed from the underlying EPUF sample because of data cleaning and data disclosure is 28,770.

[21] The SSB is a set of files containing individual-level data synthesized from Census Bureau's Survey of Income Program Participation (SIPP) results linked to various Social Security administrative files. The Census Bureau produces the SSB, which is the result of an interagency project that also includes SSA and IRS.

[22] Under random rounding, a multiple of the rounding base will not change, while a number that is not a multiple of the base will round to either of the two closest multiples of the base. For example, when random-rounding to a base of $25, the value $550 will not change. However, a value of $562 may round to either $550 or $575. The random-rounding process provides some uncertainty about the actual number reported on the individual's SSA earnings record. For example, if the earnings contained in EPUF are $550 we know the actual amount reported to SSA was between $526 and $574. The interval of uncertainty increases with the amount of earnings reported.

[23] Unless otherwise noted, the numbers of records and the amounts of earnings shown in the charts and tables are unweighted.

[24] Additionally, in many years, the percentage of individuals with earnings at or above the taxable maximum differs substantially by sex.

[25] SSA cannot determine married-couple or parent-child relationships in the file based on the information derived from the MEF. SSA establishes such linkages after an individual applies for benefits. In any event, linking currently or previously married individuals or indicating a familial relationship in EPUF would create serious data disclosure risks.

[26] An electronic folder (created when an individual applies for benefits) contains the certified earnings record, which summarizes all the earnings records from the MEF and provides the basis for computing an individual's benefits.

[27] Enumeration is the process by which SSA assigns a unique SSN for every person in order to create a work and benefit record for the Social Security program. SSA verifies all of the information on the SSN application.

[28] Earnings that cannot be properly assigned to an individual's earnings records on the MEF are placed on the Earnings Suspense File. The amount of earnings assigned to the Earnings Suspense File has grown dramatically over the past 20 years (GAO 2005).

[29] In March 2011, budget constraints led the SSA to suspend the production and mailing of printed statements. The agency is working toward developing an online alternative.

[30] This chart omits individuals whose sex is unknown. Appendix Table A-2 shows distributions by sex, including individuals of unknown sex.

[31] Recall that any earnings reported before the individual was 15 years old were assigned a value of zero for data disclosure reasons.

[32] Appendix Tables A1–A2 present the data underlying Charts 8–9.

References

[Board of Trustees] Board of Trustees of the Old-Age, Survivors, and Disability Insurance Trust Funds. 2010. *Annual Report of the Board of Trustees of the Old-Age, Survivors, and Disability Insurance Trust Funds, 2010.* Washington, DC: SSA.

[GAO] Government Accountability Office. 2005. *Better Coordination Among Federal Agencies Could Reduce Unidentified Earnings Reports.* Report no. GAO-05-154. Washington, DC: Government Printing Office.

Olsen, Anya, and Russell Hudson. 2009. "Social Security Administration's Master Earnings File: Background Information." *Social Security Bulletin* 69(3): 29–45.

Pattison, David, and Hilary Waldron. 2008. “Trends in Elective Deferrals of Earnings from 1990–2001 in Social Security Administrative Data.” Research and Statistical Note No. 2008-03. Washington, DC: SSA.

Puckett, Carolyn. 2009. “The Story of the Social Security Number.” *Social Security Bulletin* 69(2): 55–74.

[SSA] Social Security Administration. 2009. *Annual Statistical Supplement to the Social Security Bulletin, 2008.* Washington, DC: SSA.

———. 2010. “Automatic Increases: National Average Wage Index.” http://www.socialsecurity.gov/OACT/COLA/AWI.html.