Measuring defined benefit plan replacement rates with PenSync

A synthetic pension data set created with regression and statistical matching procedures utilizes IRS data to evaluate the effectiveness of a defined benefit pension plan in meeting the income needs of retirees; the findings suggest that variations in replacement rates stem from differences in benefit formulas, earnings, years in the plan, and employment characteristics

James H. Moore, Jr.

James H. Moore, Jr., is an economist in the Office of Research, Evaluation, and Statistics, Division of Policy Evaluation, Social Security Administration, Washington, DC.

Will future generations of retirees have adequate retirement income to maintain their preretirement standard of living? In an effort to better understand retirement income security, the Social Security Administration (SSA) developed a microsimulation model, called Modeling Income in the Near Term (MINT),[1] to project the retirement income of persons born between 1926 and 1965. There are three main sources of retirement income: Social Security, employer pension benefits (from both defined benefit and defined contribution pension plans), and personal savings. This article focuses on a method for projecting income from defined benefit pension plans.

Version 1 of MINT used replacement rates calculated by the Bureau of Labor Statistics (BLS, the Bureau) to estimate retirement benefits from the private sector, as well as from State and local government defined benefit plans. Because the Bureau no longer publishes replacement rates,[2] and because there are no other sources from which to obtain such rates, SSA has developed an experimental replacement rate calculation requiring BLS data on pension plans. A file containing both the statistically re-created BLS data and data from the Survey of Income and Program Participation (SIPP) is linked to earnings histories. Work was done under a memorandum of understanding between the Bureau and the SSA such that BLS data would be analyzed at the Bureau and only results of statistical equations could be taken offsite.

Under the MINT, two key components—pension plan characteristics and preretirement earnings—are used to calculate replacement rates. The statistical equations developed at the Bureau are used to estimate pension plan characteristics as a function of job characteristics, which are statistically matched to SIPP individuals. SSA administrative data on earnings are used to develop two measures of earnings and to calculate defined benefit amounts. These amounts, together with preretirement earnings, are then used to calculate replacement rates. The resulting dataset is called *PenSync*.

Estimating future pension income is especially problematic in light of the major changes that have occurred in the world of pensions. For example, over the last two decades, the demographics of individuals covered by a pension, as well as the type of pension plan providing the coverage, have changed drastically. As recently as the mid-1990s, the majority of full-time employees in medium-sized and large private establishments who were covered by a pension plan were covered by a defined benefit plan.[3] Currently, the majority of all employees (full time and part time) in private industry are covered by a defined contribution plan.[4] Not only has the type of pension plan changed, but so has the design of the plan.[5] A new type of pension plan has evolved as well: the cash balance plan, which has gained popularity over the past few years.[6] According to data recently released by the Bureau, participation in cash balance plans increased nearly fourfold between 1997 and 2000, from 6 percent to 23 percent.

Currently, no data set collects enough information to analyze these changes in pension plan coverage and design. Through a statistical match, the methodology in this article brings together (1) detailed information on pension plans and plan providers, (2) survey data on plan participants, and (3) administrative data on earnings histories, in order to improve the estimation of pension income for future retirees.

The article begins with a presentation of the methodology, including a brief description of the key components of a defined benefit plan and the models used to replicate the employer-based survey (EBS) data. Next, the data are described, after which the statistical matching procedure and the assumptions are discussed. Finally, results are given and a conclusion proffered.

Data

One of the major sources of data used in this study was the 1995 EBS. Because the 1993 SIPP data and the 1995 EBS data were collected the same year, comparability of the two data sets is facilitated. The EBS provides representative data on the incidence and detailed provisions of the Nation's defined benefit pension plans in all nonagricultural private-sector establishments employing 100 or more full- and part-time workers in all 50 States and the District of Columbia. The sample used in the study contains 4,925 observations. Because defined benefit plan provisions are difficult for the average person to interpret, the appendix to this article briefly describes some of the major provisions found in such a plan, including the benefit formulas and some of their key components, as well as eligibility requirements.[7]

Using representative samples of the Nation's households, the SIPP collects data on sources and amounts of income, various characteristics of the labor force, participation in government programs, eligibility data, and general demographic characteristics. The study presented in this article focused on the data collected in the Retirement Expectations Pension Plan Coverage Topical Module and the Work History Topical Module. To make the SIPP more comparable to the EBS, the SIPP sample was restricted to nonagricultural private-sector wage and salary workers who worked at an establishment with 100 or more employees and who were covered by a defined benefit plan. The self-employed are not included in the sample, and individuals must have had at least 5 years of employment in their current job. The sample consists of individuals who were born between 1930 and 1955 and who thus ranged in age from 40 to 65 in 1995. All told, the sample has 2,508 observations for analysis.

Two sources of administrative earnings data were used for the construction of the earnings measures: the Detailed Earnings Record and the Summary Earnings Record, both maintained by the Social Security Administration. The Detailed Earnings Record contains information on wages, tips, other compensation, and deferred wages from 1981 through 2001. These data are provided to the Internal Revenue Service on Form W-2 from employers; the form reports on all persons with wages, including nonfilers and other noncovered employees. The Summary Earnings Record contains Social Security-covered earnings derived from payroll tax records for the years 1951 through 1999 (up to the taxable wage ceiling). After a review of both data sets, it was determined that the Detailed Earnings Record had significant advantages over the Summary Earnings Record. One major advantage to using the Detailed Earnings Record is that it has earnings data for each job in each year, whereas the Summary Earnings Record's earnings data is a sum of all earnings from all jobs in each year. By using the Detailed Earnings Record, it is possible to separate earnings out by job, which in turn makes it possible to isolate one defined benefit plan with the earnings from one job, instead of having a sum of earnings from multiple jobs.

Methodology

Chart 1 shows the flow of the systematic procedures applied to create PenSync and to calculate replacement rates. The first step is to determine the structure of the data and to select the proper econometric technique that best fits the data. Ordinary least-squares (OLS) regression is used to fit continuous explanatory and dependent variables. However, because the dependent variable that represents the type of formula is categorical, the traditional OLS multiple regression analysis is not appropriate. A discrete dependent-variable model fits the data substantially better than least-square methodology.[8] Therefore, the study used a multinomial logit (MNL) model to fit the categorical dependent variable.

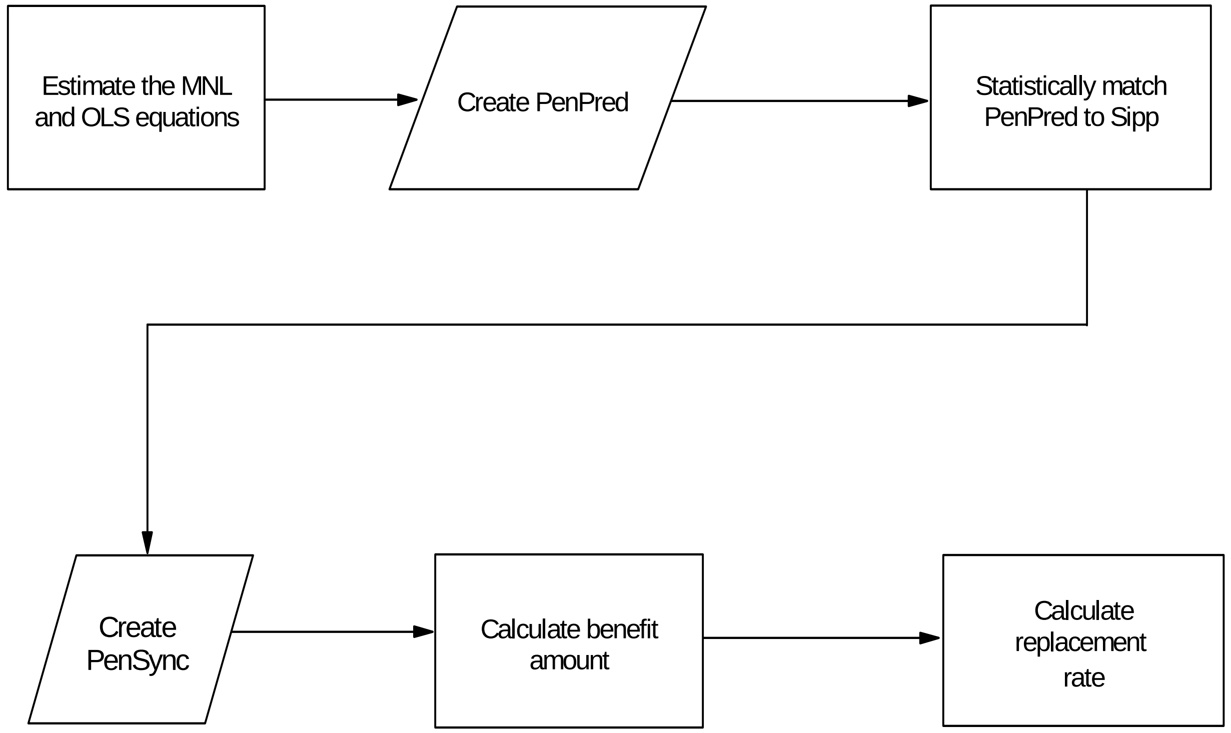
The next step involves estimating the MNL and the OLS models to obtain estimates of the coefficients. The resulting estimates are used to produce predicted values by a process of multiplying the estimated coefficients by the observed EBS data. The end product is a database called PenPred.

The next step in the process is to statistically match the predicted pension plan characteristics (PenPred) to the SIPP by job characteristics. This procedure assigns a defined benefit pension plan with detailed characteristics to the analytical sample of workers in the SIPP who reported being covered by such a plan. The resulting dataset is called PenSync. The final two steps involve constructing an algorithm to calculate benefit amounts and then calculating the replacement rate for each individual in the sample.

Model specification

MNL model specification. The employer's choice of pension formula is modeled with McFadden's random utility framework.[9] Nine alternatives are identified: two flat-dollar formulas; four types of terminal-earnings formulas; two types based on a

Chart 1. The creation of PenSync and replacement rates


Estimate the MNL and OLS equations
Create PenPred
Statistically match PenPred to Sipp
Create PenSync
Calculate benefit amount
Calculate replacement rate

percentage of the worker's career average earnings; and a cash balance plan.[10] In choosing which type of formula to provide, employers may consider a variety of job characteristics, such as their employees' occupations and work schedules. The decision may also be affected by the characteristics of the employers themselves, such as the type of industry in which the establishment operates, the number of employees in the firm, and the presence or absence of a union. (See table 1 for the descriptive statistics of job characteristics variables used to model the employer's choice of benefit formula.) For any employer *i*, the utility of choice *j* to that employer is expressed as

$$U_{ij} = V_{ij}(E_i, W_i) + å_{ij}\,, \qquad (1)$$

where

U_{ij} is the overall utility of choice *j* for employer *i*,
$V(E,W)$ represents utility determined by the observed data,
E is a vector of employer characteristics,
W is a vector of characteristics of employees within the firm,
å is a vector of unobserved components, and
j denotes pension formula alternatives.

Utility-maximizing behavior implies that employer *i* will choose a particular alternative *j* only if $U_{ij} > U_{ik}$ for all *k* not equal to *j*. The error term å is assumed to be a random variable and includes idiosyncrasies and measurement errors. Employer *i* chooses the alternative that produces the greatest utility. The decision is random.

The probability of any given alternative *j* being chosen by an employer can be expressed as

$$P = P(U_{ij} > U_{ik}), \text{ for all } k \neq j. \qquad (2)$$

By substitution of equation (1),

$$P = P(V_{ij} + å_{ij} > V_{ik} + å_{ik}, \text{ for all } k \neq j).$$

Rearranging terms yields

$$P = P[(å_{ij} – å_{ik}) > (V_{ij} – V_{ik}), \text{ for all } k \neq j]. \qquad (3)$$

If the distribution of the random å's is known, the distribution of each difference $å_{ij} – å_{ik,}$ for all j, $j \neq k$, can be derived. Then, from equation (3), the probability that the employer will choose alternative *j* can be calculated.

Table 1. Descriptive statistics for job characteristics variables

Category	Number	Percent
Industry		
Mining	56	1.14
Construction	49	.99
Manufacturing	1,330	27.01
Transportation	804	16.32
Wholesale	154	3.13
Retail	444	9.02
Finance	1,106	22.46
Service	982	19.94
Occupational groups		
Professional	1,564	31.76
Blue collar	1,652	33.54
Clerical	1,709	34.70
Union status		
Not a union member	3,547	72.02
Union member	1,378	27.98
Work Schedule		
Part time	308	6.25
Full time	4,617	93.75
Employment		
Less than 250	922	18.72
250–499	754	15.31
500–999	886	17.99
1,000 or more	2,363	47.98
Number of observations	4,925	100.00

SOURCE: Author's calculation using EBS data.

Letting $X_{ij} = (E_i, W_i)$ and assuming that V is a linear function of components of X operationalizes equation 2 as

$$U_{ij} = \hat{a}_j X_{ij} + \mathring{a}_{ij}, \tag{4}$$

where $\hat{a}_j$ is a vector of coefficients indicating the effect of the various X_{ij}'s on employer *i*'s utility derived from option *j*. Note that $\hat{a}_j$ is subscripted by the choice index *j*. This means that, in the analysis, a given X_{ij} is allowed to "interact" with each option. For example, union status may have one effect on the utility of choosing a flat-dollar formula and another on the utility of choosing a cash balance plan.

As mentioned earlier, an MNL approach is used to determine the probability that an employer will choose one of nine mutually exclusive benefit formulas:

1. flat dollar amount times years of service, together with a fixed dollar amount times years of service;
2. flat dollar amount times years of service, together with a varying dollar amount times years of service;
3. percentage of terminal earnings, together with a fixed percentage of earnings, averaged over the last few years of employment;
4. percentage of terminal earnings, together with a varying percentage of earnings, averaged over a specified period of consecutive years of employment;
5. percentage of terminal earnings, together with a varying percentage of earnings, averaged over the last few years of employment;
6. percentage of terminal earnings, together with a fixed percentage of earnings, averaged over a specified period of consecutive years of employment;
7. percentage of terminal earnings, together with a fixed percentage of earnings, averaged over the employee's career;
8. percentage of terminal earnings, together with varying percentages of earnings, averaged over the employee's career;
9. cash balance plan.

(Yet a 10th formula is a pension equity plan, based on terminal earnings and to which interest rates do not apply. However, the incidence of such plans is too scarce to estimate with any precision.)

The MNL model is frequently used to analyze situations in which there are a number of alternatives. However, it is widely known that a potentially important drawback of the model is the property called "independence from irrelevant alternatives" (IIA); that is, the model can be applied only to situations in which the alternatives from which one chooses are totally independent.

To test for the existence of IIA, a model is constructed such that the alternatives include choosing one type of benefit formula over a different type of benefit formula. If the employer views the alternatives as differing only along irrelevant dimensions, then, when the model is reestimated, it will not show a significant difference in explanatory power from that of the original model. The model used in this article passed the IIA assumption.

That the model passed the IIA assumption is not entirely surprising, given that there are many incentives embedded in the different types of pension formulas offered by employers. Some types of pension formula are geared toward retaining employees, while others encourage retirement. Therefore, depending upon the incentive sought by the employer, his or her decision to offer a particular type of pension formula is IIA. Again, the purpose of the IIA test is to ensure that the alternatives presented to employers are indeed viewed as independent.[11] Consequently, in this context, for a given employer *i* with characteristic x_i, the probability of choosing a given benefit formula can be estimated with the MNL model

$$\mathrm{BF}_{ij} = \frac{e^{v_{ij}}}{\sum_{k=1}^{K} e^{v_{ijk}}}, \tag{5}$$

where

BF_{ij} = the probability that employer *i* chose formula *j*,

$v_{ij} = \sum\beta_m X_{ijm}$ = the deterministic component of the utility of formula j to employer i,
X_{ijm} = the mth explanatory variable for formula j and employer i, in which $m = 1...M$, and
β_m = coefficient to be estimated.

The MNL model includes information on characteristics of the employer, of his or her employees, and of the pension plan the employer is offering. (For a description of the values of the dependent variable, see exhibit 1.) In addition to predicting the type of formula, the model estimates the quantitative values common to each type, using OLS.

OLS model specification. The quantitative variables for employer i and formula j can be written as

$$\mathrm{QV}_{ij} = \beta_{0ij}...\beta_{1ij}X + \mathring{a}_{ij}, \qquad (6)$$

where QV_{ij} is a set of quantitative pension provision variables used in the pension benefit calculation and i denotes the ith employer. In this model, the coefficients are estimated by a linear least-squares multiple regression, β_{0i} is a constant, X is a vector of job characteristics of the employer and his or her employees and pension plan characteristics, and $\mathring{a}_i$ is an error term. (See exhibit 2 for a listing and definition of the quantitative pension variables.)

Creating the synthetic pension file

As shown in chart 1, the first two steps in creating PenSync involve fitting the MNL and OLS models to the EBS data set to score a new data set of predicted observations.[12] Table 2 gives an overview of the accuracy of the MNL model. The model predicted the correct formula 71 percent of the time, on average, and many of the incorrect predictions were among similar types of formulas. For example, the model predicted a flat-dollar formula with a fixed dollar amount with a 95.77-percent accuracy rate, while predicting a flat-dollar formula with a varying dollar amount 20.45 percent of the time. However, when the model incorrectly predicted a flat-dollar formula with a varying dollar amount, it predicted that that formula would be a flat-dollar formula with a fixed dollar amount 50 percent of the time. Both types of formula are similar in their design, and any attempts that were made to increase the accuracy of the prediction flawed the model with multicollinearity and overspecification. The results from the OLS models are found in table 3.

To summarize the procedure, the first step involved estimating equations 5 and 6 to generate a set of coefficient estimates, which are used to replicate the EBS data. The resulting estimates of the coefficients are used to produce predicted values by multiplying each estimated coefficient by the corresponding observed EBS data. This multiplication process is repeated for each variable in the equations specified. The end product is a database containing the predicted values for each observation required to compute a pension benefit amount, along with the related explanatory variables. The database is called PenPred. To assess the quality of PenPred, the resulting means and standard deviations are compared with those of the EBS. (See table 4.)

Statistical matching. Statistical matching is a process of linking data from multiple data sets on the basis of similar characteristics rather than unique identifying information. In a

Exhibit 1. Description of the values for the multinomial logit dependent variable

Value	Type of formula
1	Flat dollar amount times years of service, together with a fixed dollar amount times years of service
2	Flat dollar amount times years of service, together with a varying dollar amount times years of service
3	Percentage of terminal earnings, together with a fixed percentage of earnings, averaged over the last few years of employment
4	Percentage of terminal earnings, together with a varying percentage of earnings, averaged over a specified period of consecutive years of employment
5	Percentage of terminal earnings, together with a varying percentage of earnings, averaged over the last few years of employment
6	Percentage of terminal earnings, together with a fixed percentage of earnings, averaged over a specified period of consecutive years of employment
7	Percentage of terminal earnings, together with a fixed percentage of earnings, averaged over the employee's career
8	Percentage of terminal earnings, together with varying percentages of earnings, averaged over the employee's career
9	Cash balance plan

Exhibit 2. Definitions of quantitative variables

DOL_DOL1	First dollar-amount breakpoint used to calculate a flat-dollar formula
DOL_DOL2	Second dollar-amount breakpoint used to calculate a flat-dollar formula
DOL_DOL3	Third dollar-amount breakpoint used to calculate a flat-dollar formula
DOL_YRS1	First years-of-service breakpoint used to calculate a flat-dollar formula
DOL_YRS2	Second years-of-service breakpoint used to calculate a flat-dollar formula
NORM_AAS	Sum of normal retirement age and years of service
NORM_AGE	Normal retirement age
NORM_SRV	Normal retirement service requirement
NR_PAY	Percentage of earnings contributed to a cash balance plan
NR_INT	Interest rate
EBASEYR1	First breakpoint for number of years to be included in the calculation of benefits
EBASEYR2	Second breakpoint for number of years to be included in the calculation of benefits
POE_DOL1	First dollar-amount breakpoint used to calculate a percentage-of-earnings formula
POE_DOL2	Second dollar-amount breakpoint used to calculate a percentage-of-earnings formula
POE_PCT1	First percentage-of-earnings breakpoint used to calculate a percentage-of-earnings formula
POE_PCT2	Second percentage-of-earnings breakpoint used to calculate a percentage-of-earnings formula
POE_PCT3	Third percentage-of-earnings breakpoint used to calculate a percentage-of-earnings formula
POE_PCT4	Fourth percentage-of-earnings breakpoint used to calculate a percentage-of-earnings formula
POE_PCT5	Fifth percentage-of-earnings breakpoint used to calculate a percentage-of-earnings formula
POE_YRS1	First breakpoint for number of years of service to be included in the calculation of benefits
POE_YRS2	Second breakpoint for number of years of service to be included in the calculation of benef

statistical match, each observation in one microdata set (a *base* database) is assigned one or more observations from another microdata set (a *secondary* database). The assignment is made on the basis of similar characteristics because the files lacked the same unique identifier.

A substantial amount of research has been carried out concerning the validity of using statistically matched data for analysis. A number of the early researchers in the field carefully documented some of the shortcomings of statistical matching.[13] In particular, Benjamin Okner pointed out some of the common problems with statistical matching, including comparability of the data, the handling of missing data, specific techniques for matching, and the definition and evaluation of the goodness of a match. The next subsection briefly discusses some steps taken to address Okner's concerns.

Data comparability. In an effort to make the PenPred data and the SIPP data compatible, the following harmonization criteria, well discussed in the literature, were used: [14]

1. *Harmonization of units*. It is necessary that records from the different sources refer to the same unit. The unit of analysis for this study is workers.

2. *Harmonization of target population*. If the data sets refer to different target populations, it is important to select just those records which refer to the population of interest. Both data sets comprise a sample of workers employed in private nonagricultural industries and occupations and who participate in a defined benefit plan.

3. *Harmonization of variables*. The common variables should be defined in the same way. Both data sets use Standard Industry Codes and Census Occupation Codes to categorize the industry and occupation, respectively.

Missing data. There are three common approaches to handling missing data: impute the missing data, model the probability of "missingness," or ignore the missing data. After testing to make sure that there were no significant differences on the key variables between records with missing data and records without missing data, the more conservative approach to handling missing data was adopted. Hence, missing values are replaced with means for each variable.[15]

Selection of the matching variables. Consider first PenPred, henceforward called the universe U, consisting of a set of N records. For each record, there are values for R variables. U is represented by an N-by-R matrix, in which each of the N rows contains the values of the R variables for one record. The R variables represent the industry code, the occupation code, and the union status, all of which are considered key variables for matching based on analysis performed on the EBS data. The SIPP consists of a set of M

Table 2. Accuracy of multinomial logit model

Frequency and percent	Observed formula value	Predicted formula value									Observed total
		Flat dollar		terminal earnings				Career average		Cash balance	
		1	2	3	4	5	6	7	8	9	
Predicted total		873	20	147	1,683	358	1,446	21	95	282	4,925
Frequency	1	816	6	0	14	0	1	2	1	12	852
Percent		95.77	.70	.00	1.64	.00	.12	.23	.12	1.41	…
Frequency	2	22	9	0	13	0	0	0	0	0	44
Percent		50.00	20.45	.00	29.55	.00	.00	.00	.00	.00	…
Frequency	3	0	0	112	0	43	0	0	0	0	155
Percent		.00	.00	72.26	.00	27.74	.00	.00	.00	.00	…
Frequency	4	1	1	2	1,182	0	207	1	1	0	1,395
Percent		.07	.07	.14	84.73	.00	14.84	.07	.07	.00	…
Frequency	5	0	1	29	1	315	1	0	0	0	347
Percent		.00	.29	8.36	.29	90.78	.29	.00	.00	.00	…
Frequency	6	0	3	4	473	0	1,099	6	10	0	1,595
Percent		.00	.19	.25	29.66	.00	68.90	.38	.63	.00	…
Frequency	7	0	0	0	0	0	6	11	0	0	17
Percent		.00	.00	.00	.00	.00	35.29	64.71	.00	.00	…
Frequency	8	0	0	0	0	0	132	0	83	0	215
Percent		.00	.00	.00	.00	.00	61.40	.00	38.60	.00	…

SOURCE: Author's calculation using EBS and PenSync data.

records. For each record, there are values for the *S* variables that are represented by an *M*-by-*S* matrix, in which each of the *M* rows contains the values of the *S* variables for one record. The *S* variables represent the industry code, the occupational code, and the union status.

As mentioned earlier, to enable two or more data sources to be statistically matched, a set of variables common to all data sets must be found. These common characteristics are referred to as *X* variables, where $X = (x_1,..., x_p)$. In this equation,

x_1 = the worker's two-digit standard industry classification;[16]

x_2 = the worker's three-digit standard occupation classification;[17] and

x_3 = the worker's union status.The *i*th record in *U* is denoted

$$U_i = (u_{il}\, u_{i2}...u_{ij}) \tag{7}$$

and, as indicated, contains *j* observed variables. Similarly, the *i*th record in the SIPP,

$$\text{SIPP}_i = (\text{SIPP}_{il}\ \text{SIPP}_{i2}...\ \text{SIPP}_{ih}) \tag{8}$$

contains *h* observed variables. The remaining variables in each of the files are referred to as *Y* on the PenPred file and *Z* on the SIPP file. $Y = (y_1..y_q)$, where y_i is a vector of predicted values of all pension provisions; and $Z = (z_1...z_r)$, where z_i is a vector of socioeconomic and work history variables.

Specification of the distance function. The statistical matching procedure is carried out by minimizing a distance function, defined as the absolute difference of the numerical values of the occupations and the union statuses in two cases: the distance between the *i*th worker in the *U* and the *j*th worker in the SIPP is defined by

$$D_{ij} = \sum_{n=1}^{k} \left(I_{in} - I_{jn}\right) + \left(O_{in} - O_{jn}\right) + \left(U_{in} - U_{jn}\right) \ , \tag{9}$$

where

$n = 1,..., k$,

D_{ij} = the distance between the *i*th *U* record and the *j*th SIPP record,

$I_{in} - I_{jn}$ = the distance between the values of the *n*th pair of industry variables in the *i*th record,

$O_{in} - O_{jn}$ = the distance between the values of the *n*th pair of occupation code variables in the *i*th record, and

$U_{in} - U_{jn}$ = the distance between the values of the *n*th pair of union status variables in the *i*th record.

Certain *X* variables may be treated as cohort variables. A cohort variable establishes subclasses of the records in each

Table 3. Regression results for selected quantitative variables ordinary least squares model

Variable	Constant	Size	Industry	Work schedule	Occupation	Union status	Dollar formula	Career average	R^2
DOL_DOL1	5.0851	–0.0005	–2.862	–2.0372	1.2767	0.3024	31.8015	0.7117	.74
	[1](.80890)	[1](.00001)	[1](.3666)	[1](.4234)	[1](.2336)	(.2616)	[1](.5091)	[1](.4262)	…
CB PERCENT	4.5894	.0001	.164	–.0600	–.0032	–.0346	–4.8377	–4.8791	.79
	[1](.0735)	[1](.00001)	[1](.0322)	(.0372)	(.0205)	(.023)	[1](.0447)	[1](.0375)	…
CB INTEREST	5.26057	–.0001	.0044	.043	.0502	.016	–5.2488	–5.2148	.79
	[1](.076)	(.00001)	(.0333)	(.0385)	(.0212)	(.0238)	[1](.0462)	[1](.0387)	…
POE 1	–2.6099	.0002	–.3918	1.8657	.6683	.8312	–.3176	12.9813	.67
	[1](.480)	[2](.00005)	(.2103)	[1](.2429)	[1](.1340)	[1](.1501)	(.2921)	[1](.2445)	…
POE 2	.2800	.00002	.1202	–.054	–.0807	–.2721	–.1862	5662	.18
	[2](.0911)	(.000009)	(.0399)	[2](.0461)	(.0254)	[1](.0285)	[2](.0554)	[1](.0464)	…
YEARS 1	–3143	.0001	.3194	.0678	–.062	.0314	–.3266	3.3456	.41
	(.2185)	[1](.000002)	[2](.0957)	(.1106)	(.0610)	(.0683)	(.133)	[1](.1113)	…
YEARS 2	–4.3253	–.0006	4.3718	8.346	–1.8145	3.6991	–6.4945	26.0477	.12
	(3.9373)	(.0004)	(1.7254)	[1](1.993)	(1.1)	(1.2312)	(2.3964)	[1](2.0059)	…
NORM_AGE	46.606	.001564	5.454	–3.20707	–2	–2.98348	–2.8452	7.651	.09
	[1](2.01)	[1](.0002)	[1](.88)	[2](1.01)	[2](.56)	[1](2.98)	(1.22)	[1](1.02)	…
NORM_SRV	10.629	–.00152	–6.373	3.71762	1.3416	2.67692	6.3605	1.856	.10
	[1](1.94)	[1](.0001)	[1](.523)	[1](.604)	[1](.333)	[1](.7)	[1](.723)	(.61)	…

[1] Significant at 1-percent statistical level.
[2] Significant at 5-percent statistical level.

of the two files, with matching permitted only between a pair of cases in the same subclass. In this study, x_1, "industry," is the cohort variable. For example, a worker in the mining industry in the SIPP file can be matched only to another worker in the mining industry in the *U* file.

Assumptions. Three assumptions are relevant to the statistical matching procedures:

1. *No unobserved heterogeneity exists between the predicted data and the observed data*. Stated differently, the probabilities associated with being covered by a given pension formula and having a particular set of job characteristics are analogous across the three data sets. Mathematically, this identifying assumption is captured in the formula

$$\pi(x,y|\, X, \mathrm{Data}_{\mathrm{BLS}}) - \pi(x,y,|\, X, \mathrm{Data}_{\mathrm{SIPP}}) - \pi(x,y,|\, X, \mathrm{Data}_{\mathrm{PenSync}}) = 0 \qquad (10)$$

where
x = type of pension plan,
y = type of formula,
and X is a vector of individual job characteristics (for example, industry, occupation, and union status).

Sensitivity analysis was conducted to check the validity of this assumption. Basic descriptive analysis revealed that the mean values of the observed data are similar to the mean values of the predicted data. Cross tabulations also revealed similarities between the three data sets.

2. *Workers will remain on their current job until they reach the normal retirement age.* This assumption is rendered mathematically as

$$\pi(x,y|X_t, \mathrm{Data}_{\mathrm{SIPP}}) - \pi(x,y|X_{t+i}, \mathrm{Data}_{\mathrm{SIPP}}) = 0, \qquad (11)$$

where
i = start year of current job,..., retirement year.

Many defined benefit plans allow workers to retire prior to the normal retirement date, but the worker's benefit is reduced by an actuarial reduction factor. The current version of PenSync does not have the capability to model early retirement; therefore, it is assumed that workers will remain on their current job until they satisfy the normal retirement provision specified in their defined benefit plan. Note that the assertion that workers will remain on their current job obviously presupposes that those workers will continue to work in the same industry and occupation. To test the feasibility of remaining on the current job, the SIPP and the data from the Detailed Earnings Record were used to measure tenure on the current job and the frequency of job change. The SIPP data reveal that the average tenure on the current defined benefit pension job was 18 years, and the Detailed Earnings Record data indicate that, between the starting year (reported in the work history topical module of the SIPP) of the current job and 2003, 63 percent of the workers in the sample remained with their same employer. To test these assumptions further, the SIPP data are used to check how often a worker reports changing industry or occupation. When the full panel of the SIPP is analyzed, it is found that 92 percent and 90 percent of the workers report remaining in the same industry and occupation, respectively. (Recent growth

Table 4. Mean and standard deviation for predicted and observed quantitative variables

Variables	Mean			Standard deviation		
	Predicted	Observed	Difference	Predicted	Observed	Difference
DOL_DOL1	6.40	6.33	0.06	11.81	13.83	–2.02
DOL_DOL2	.04	.09	–.05	.20	1.44	–1.25
DOL_DOL3	.66	.46	.19	1.10	5.20	–4.10
DOL_YRS1	.15	.11	.04	.36	1.14	–.78
DOL_YRS2	.05	.11	–.06	.22	1.81	–1.59
NORM_AAS	5.32	5.30	.02	2.03	20.10	–18.07
NORM_AGE	57.38	57.33	.04	5.29	17.77	–12.49
NORM_SRV	7.89	7.91	–.02	3.23	10.59	–7.36
NR_PAY	.31	.30	.01	1.21	1.34	–.13
NR_INT	.31	.32	–.01	1.21	1.41	–.20
EBASEYR1	2.97	2.79	.18	1.70	2.40	–.71
EBASEYR2	21.24	20.76	.48	11.67	35.52	–23.85
POE_DOL1	243.58	234.11	9.47	146.37	1,877.95	–1,731.58
POE_DOL2	.00	.00	.00	.00	.00	.00
POE_PCT1	10.19	10.24	–.04	5.64	7.03	–1.39
POE_PCT2	.76	.67	.09	.43	.85	–.42
POE_PCT3	.00	.18	–.18	.00	.43	–.43
POE_PCT4	.00	.02	–.02	.00	.14	–.14
POE_PCT5	.00	.04	–.04	.00	.21	–.21
POE_YRS1	5.40	5.22	.18	2.91	11.30	–8.39
POE_YRS2	.50	.43	.06	.50	2.28	–1.78

SOURCE: Author's calculation using EBS and PenSync data.

in cash balance plans may have affected the length of time people stay in their jobs, but the timeframe of the data is years before that growth.)

3. *The SIPP-reported pension job for employer 1 is the job with the highest earnings in the W-2 file in each year.* Again, mathematically, this assumption can be stated as

$$\pi(x,y|\, X_t, \text{Data}_{\text{DER}}) - \pi(x,y|\, X_t, \text{Data}_{\text{SIPP}}) = 0, \qquad (12)$$

where X = earnings in a given year and t = 1951...2002. This assumption assumes that the pension module job 1 in the SIPP[18] is the same as the job reporting the highest wage on the Detailed Employment Record. SIPP respondents are asked the question about calendar-year wages and salaries twice per panel and are encouraged to refer to their respective W-2 forms or other documents to ensure their accuracy.

To test the validity of the third assumption, the earnings total reported in the SIPP for the pension job is compared with the highest-wage job on the Detailed Employment Record for the same year. The SIPP earnings are similar to the highest earnings on the Detailed Employment Record, varying by plus or minus $2,000 annually. Respondents in the SIPP also can report earnings and pension coverage from two employers; therefore, to render it yet more likely that the probability that the pension job reported for employer 1 is indeed the highest-wage job on the Detailed Employment Record, the second job reported in the SIPP is analyzed. The analysis reveals that less than 3 percent of the unweighted individuals who reported having a defined benefit type of pension reported having the same type of pension on their second job.

The matching algorithm. The match procedure is unconstrained, which has the advantage of permitting the closest possible match for a *U* record, but at the cost of increasing the sample variance of estimators involving the *Y* and *Z* variables. To avoid violating the confidentially provision in the memorandum of understanding, particular attention is given to tabulations based on small cell sizes. To avoid the possibility of unauthorized disclosure, cells with three or fewer cases were dropped from the sample.

The matching algorithm also employs a decision rule: if the pair agrees on all three characteristics (that is, industry, occupation, and union status), designate the pair as a level-1 match; or else if the pair agrees on the two characteristics industry and occupation, designate the pair as a level-2 match; or else if the pair agrees on the two characteristics industry and major occupational group, designate the pair as a level-3 match; or else if the pair agrees on industry characteristics only, designate the pair as a level-4 match; or else designate the pair as a nonmatch. As shown in the following tabulation, the final data file for analysis consists of 2,508 observations containing detailed socioeconomic variables, along with in-depth employer-provided pension data:

Level	*Number of matches*	*Match rate (percent)*
Total	2,508	100
1	1,876	75
2	192	8
3	430	17
4	10	.004

This database is called PenSync.

Benefit algorithm. The final procedure used to create the synthetic pension file involves constructing an algorithm to calculate benefit amounts and replacement rates for each individual in PenSync. The algorithm starts by determining the type of formula assigned to an individual (for example, career average earnings, terminal earnings, cash balance, or a flat-dollar formula). For individuals covered by a formula based on a percentage of their earnings times years of service, a subroutine is initiated to determine whether the earnings are career average earnings or terminal earnings. For individuals covered by a career average arrangement, the benefit amount is determined by multiplying a proportion of the average earnings from the Detailed Earnings Record by the worker's total number of credited years of service.[19] For individuals whose benefit amounts are based upon a terminal earnings arrangement, the algorithm multiplies a proportion of the average earnings from the Detailed Earnings Record during a specified period, typically near the individual's retirement age.

For individuals who are covered by a cash balance plan, the benefit amounts are represented as an account balance equal to a percentage of the individual's earnings during each year of participation in the plan, credited with interest based on some index. At retirement, a participant in a cash balance plan typically receives his or her accumulated vested account as a lump sum. For purposes of the analysis carried out in this article, once the worker reaches the normal retirement age specified by the plan, the accumulated vested account is transformed into an annuity. Some benefits are associated, not with earnings, but rather, with a dollar amount per year of service. For those individuals, the benefit amount is determined by multiplying a fixed dollar amount by years of service in the plan.

Table 5. Pension income and replacement rate for workers who qualify for normal retirement prior to 2003

Category	Percent of workers	Average earnings (dollars)		Monthly benefit	Replacement rate (percent)	
		High 3 of last 5	High 5 of last 10		High 3 of last 5	High 5 of last 10
All workers	100	$37,958	$ 32,649	$1,012	32	29
Type of formula						
Dollar formula	19	35,858	30,068	818	21	24
Terminal earnings	54	38,921	34,381	1,144	38	30
Career average	10	32,233	28,192	781	21	20
Cash balance	17	40,600	32,614	960	32	36
Occupation						
Professional/technical	39	49,779	42,579	1,415	42	33
Administrative/clerical	18	25,148	22,607	579	24	25
Production/service	43	32,308	27,606	815	26	27
Industry						
Goods producing	40	37,828	32,999	913	26	27
Non-goods producing	60	38,044	32,417	1,079	36	31
Years in the plan						
0–10	16	28,015	23,711	256	9	11
11–15	15	31,144	27,315	502	18	20
16–20	10	33,406	29,080	845	28	31
21–25	12	29,837	26,122	955	30	34
26–30	26	45,759	38,206	1,178	33	33
More than 30	22	47,428	41,674	1,840	61	41
Union status						
Non-union member	66	39,594	33,930	917	25	27
Union member	35	34,852	30,219	1,202	46	32

NOTE: High 3 of last 5 is the average of the 3 highest years of earnings 5 years prior to the normal retirement date specified in the pension plan. High 5 of last 10 is the average of the 5 highest years of earnings 10 years prior to the normal retirement date specified in the pension plan. All earnings and benefit amounts are measured in 2003 dollars. Eligibility for retirement depends on a worker's age or number of years of credited service, or both. The mean normal retirement age in PenSync is 60, with an average of 25 years of service. The normal retirement date is the year in which the worker satisfies his or her pension plan provision which specifies that the worker is eligible to receive an unreduced retirement benefit. The year 2003 is used to verify whether an individual has satisfied the normal retirement requirement. The mean normal retirement year in PenSync is 1998.

SOURCE: Author's calculation using PenSync.

The final step in the algorithm produces a set of pension benefits and replacement rate ratios for the two measures of earnings: the last 10 years of earnings (L10YR) and the last 5 years of earnings (L5YR). L10YR is the average of the 5 highest years of earnings 10 years prior to the normal retirement date specified in the pension plan; L5YR is the average of the 3 highest years of earnings 5 years prior to the pension plan's normal retirement date. The latter is the year in which the worker satisfies provisions specified in the plan in order to receive an unreduced retirement benefit. The year 2003 is used to verify whether an individual has satisfied the pension plan's normal retirement requirement. All earnings and benefit amounts are measured in 2003 dollars.

Results

For workers who are eligible for normal retirement benefits prior to 2003, the defined benefit plan is estimated to replace about 30 percent of the last year of positive earnings. The average earnings are estimated to be about $35,000, and the average monthly pension benefit is $1,012. (See table 5.) Pension replacement rates are estimated to vary by the type of benefit formula, employment characteristics, and years of participation in the pension plan. Replacement rates were lowest for those in flat-dollar or career average formulas and highest for those in terminal earnings formulas or cash balance formulas, with a 16- to 17-percentage-point differential. Replacement rates were considerably lower for those in administrative/clerical or production/service jobs, compared with those in professional/technical jobs, and were lower for those in goods-producing industries than those in non-goods-producing industries. Union members are estimated to have higher replacement rates than non-union members, and more years of participation in a pension plan is associated with much higher replacement rates. Workers who remain in the same pension plan for more than 30 years have more than 60 percent of their earnings in the 5 years prior to retirement replaced by their plans, compared with only a 9-percent replacement rate for those with less than 10 years of participation.

PREDICTING RETIRMENT INCOME FROM A PENSION PLAN is a difficult task. The absence of good data is a major contributor to the difficulty involved. Furthermore, the lack of comprehensive data sources on pensions places limitations on pension research and policy decisions. The methodologies applied in this article have been in existence for decades, yet they remain more of an art than a science. However, many challenges are inherent in the employment of the procedure itself: the specification of an appropriate model, data harmonization, and, probably most important, the quality of the data. Nevertheless, the methodology set forth herein is a reasonable approach, given constraints from two different restricted data sets. □

Notes

[1] MINT was developed to estimate the distributional effects of proposed Social Security policy alternatives on current and future beneficiaries' retirement income. The model projects retirement income from Social Security, pensions, personal investments or savings, and partial retirement earnings. For a complete description of the MINT project, see the final reports prepared by the RAND Corporation (Constantijn Panis and Lee Lillard, "Near Term Model Development," draft final report, SSA contract no. 600-96-27335 (Santa Monica, CA, RAND, 1999); Constantijn Panis, Michael Hurd, David Loughran, Julie Zissimopoulos, Steven Haider, and Patricia St. Clair, "The Effect of Changing Social Security Administration's Early Entitlement Age and the Normal Retirement Age," draft report, SSA contract no. 600-96-27335 (Santa Monica, CA, RAND, 2002)); The Urban Institute (Eric Toder and others, "Modeling Income in the Near Term—Projections of Retirement Income through 2020 for the 1931–1960 Birth Cohorts," final report, SSA contract no. 600–96–27332 (Washington, DC, The Urban Institute, 1999)); and the Social Security Administration (Barbara A. Butrica, Howard M. Iams, James Moore, and Mikki Waid, *Methods in Modeling Income in the Near Term (MINT)*, ORES working study no. 91 (Social Security Administration, May 2001)).

[2] The last years the Bureau published replacement rates for full-time employees were 1993 for those in medium and large private establishments and 1994 for State and local government employees.

[3] See *Employee Benefits in Medium and Large Private Establishments, 1993*, Bulletin 2456 (Bureau of Labor Statistics, November 1994), especially table 1, p. 8.

[4] See *National Compensation Survey: Employee Benefits in Private Industry in the United States, 2000*, Bulletin 2555 (Bureau of Labor Statistics, January 2003), especially table 1, p. 4.

[5] See Olivia Mitchell, "Developments in Pensions," *NBER Reporter* (Washington, DC, National Bureau of Economic Research, 1998); and Leslie E. Papke, "Are 401(k) Plans Replacing Other Employer-Provided Pensions? Evidence from Panel Data," *Journal of Human Resources*, vol. 34, no. 2, spring 1999, pp. 346–68.

[6] Kenneth R. Elliott and James H. Moore, "Cash Balance Pension Plans: The New Wave," *Compensation and Working Conditions*, vol. 5, no. 2, summer 2000, pp. 3–12.

[7] To learn more about defined benefit plans and their features, see Gerald E. Cole, "An Explanation of Pension Plans," *Employee Benefits Journal*, June 1999, pp. 3–13.

[8] A. Agresti, *Categorical Data Analysis* (New York, J. Wiley & Sons, 1990).

[9] D. McFadden, "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, ed., *Frontiers in Econometrics* (New York, Academic Press, 1974), pp. 105–42.

[10] See the appendix for a brief description of these alternatives.

[11] Interested readers should refer to W. H. Green, *Econometric Analysis* (New York, Macmillan, 1990); K. Train, *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand* (Cambridge, MA, MIT Press, 1986); and Moshe Ben-Akiva and Steven Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand* (Cambridge, MA, MIT Press, 1985; 4th printing, 1991).

[12] For a description of the SAS Proc Score procedure, visit the website **http://ftp.sas.com/techsup/download/stat/scorenew.html**. See also SAS Technical Support Documents 650e, *Multinomial Logit, Discrete Choice Modeling: An Introduction to Designing Choice Experiments*, and *Collecting, Processing, and Analyzing Choice Data with SAS* (Cary, NC, SAS Institute, Inc., 2001).

[13] See Benjamin A. Okner, "Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File," *Annals of Economic and Social Measurement*, July 1972, pp. 325–52, and "Data Matching and Merging: An Overview," *Annals of Economic and Social Measurement*, April 1974, pp. 347–52; Horst E. Alter, "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970," *Annals of Economic and Social Measurement*, vol. 3, no. 2, 1974, pp. 373–94; D. B. Radner, R. Allen, M. E. Gonzalez, T. B. Jabine, and H. J. Muller, *Report on Exact and Statistical Matching Techniques*, statistical policy working paper (U.S. Dept. of Commerce, 1980); and J. T. Barry, "An Investigation of Statistical Matching," *Journal of Applied Statistics*, vol. 15, 1988, pp. 275–83.

[14] The statistical matching criteria for integrating data were taken from Marcello D'Orazio, Marco Di Zio, and Mauro Scanu, "Statistical Matching: a tool for integrating data in National Statistical Institutes" (Rome, Italian National Statistical Institute, 2001); on the Internet at **http://webfarm.jrc.cec.eu.int/ETK-NTTS/Papers/final_papers/43.pdf**.

[15] See R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data* (New York, J. Wiley and Sons, 1978); J. O. Kim and J. Curry, "The treatment of missing data in multivariate analysis," *Sociological Methods and Research*, vol. 6, 1977, pp. 215–40; and P. L. Roth, "Missing data: A conceptual view for applied psychologists," *Personnel Psychology*, vol. 47, 1994, pp. 537–60.

[16] All workers are classified into one of more than 82 industries according to their Standard Industrial Classification.

[17] All workers are classified into one of more than 820 occupations according to their Standard Occupational Classification.

[18] The SIPP asks respondents about two jobs.

[19] For all individuals, regardless of type of formula, the number of credited years of service is determined by subtracting the normal retirement year specified in the pension plan from the year the worker reported starting his or her current job. For years of earnings that are outside the scope of the Detailed Earnings Record, the Summary Earnings Record is used to supplement the missing data.

APPENDIX: Brief description of defined benefit provisions

A defined benefit plan provides employees with guaranteed retirement benefits based on a predetermined formula. There are three basic types of defined benefit formulas found in the employer-based survey (EBS) data: (1) a percentage of earnings per year of service, (2) a cash balance arrangement, and (3) a flat amount per year of service.

According to the EBS data, the majority of workers who participate in a defined benefit plan are covered by a formula based on a percentage of their earnings per year of service.[1] In this type of arrangement, the employee benefit is based on a proportion of earnings per year of service for each year that an employee participates in the plan. The years of service credited may be based upon either a career average or final earnings. Under a career average arrangement, the plan benefits accrue in accordance with the average of the earnings paid over the entire period of the employee's participation in the plan. Under a final-pay arrangement, by contrast, the plan benefits are based on an average of the employee's earnings during a short period, typically near the employee's retirement age. For example, the earnings may be averaged over the last 3 or 5 years of employment or over the 3 or 5 consecutive years in the 10-year period immediately prior to retirement, during which the employee's earnings are typically the highest.

A cash balance plan is another type of defined benefit plan—one whereby the benefit formula takes into account the employee's income and the number of years of service credited. Although a cash balance plan is structured to bear a resemblance to a defined contribution plan, the benefits are represented as an account balance instead of as an annuity. The account balance is equal to a percentage of the employee's income during each year of participation in the plan, and it is also credited with interest. The interest rate is often based on an index, such as the rate of return on 30-year Treasury bonds.

Some benefits are associated, not with income, but rather, with a dollar amount per year of service. In 2000, 14 percent of all workers in the private sector who were covered by a defined benefit plan had this type of plan. A formula incorporating a flat dollar amount per year of service provides a benefit amount based on a fixed dollar amount multiplied by years of service in the plan. To illustrate, if a plan specifies a benefit of $40 a month for each year of service, an employee with 30 years of participation in the plan would receive a monthly benefit of $1,200.

Before an employee is entitled to benefits from the plan, he or she must become *vested*, which means having a designated number of years of service with an employer. A 5-year cliff-vesting requirement is the most prevalent provision. Therefore, the study presented in this article assumes that, upon satisfying the 5-year vesting requirement, an individual is entitled to receive a nonforfeitable accrued benefit upon separation or retirement.

Benefits under a defined benefit plan are usually paid when the employee retires. All defined benefit plans are required to specify an age, years of service, or some combination of the two whence an employee can receive unreduced benefits. The normal retirement age in most plans is 65 years. However, many defined benefit plans allow retirement after a stated age that is earlier than the declared normal retirement age, but the employee's benefit is reduced by an actuarial reduction factor. This provision is called *early retirement*.

Note to the appendix

[1] These data can be found at **http://www.bls.gov/ncs/ebs/sp/ebrp0001.pdf**.