Measuring Defined Benefit Plan Replacement Rates Using PenSync

James H. Moore, Jr.

Office of Research, Evaluation, and Statistics, Division of Policy Evaluation, Social Security Administration
500 E. Street SW, Suite 910, Washington, DC 20254
james.h.moore@ssa.gov

Abstract

This paper creates a synthetic pension data set through the use of regression and statistical matching procedures, and as a by-product, updates income replacement rates from defined benefit plans. This new data set contains detailed socioeconomic variables along with in-depth, employer-provided pension data. Through the use of data from the Internal Revenue Service (IRS) Form W-2, two earnings-based measures of replacement rates are calculated to evaluate the effectiveness of a defined benefit pension plan in meeting the income needs of retirees. The findings suggest that variation in pension replacement rates stem from differences in the types of benefit formulas, individual earnings, years of participation in the pension plan, and employment characteristics.

I. Introduction

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

Version 1 of the MINT used replacement rates calculated by the Bureau of Labor Statistics (BLS) to estimate retirement benefits from the private sector, as well as state and local government DB plans. Since BLS no longer publishes replacement rates [2] and there are no other sources to obtain replacement rates, SSA has developed an experimental replacement rate calculation requiring BLS data on pension plans, and data from the Survey of Income and Program Participation (SIPP) linked to earnings histories. Work was done under a Memorandum of Understanding between BLS and SSA under which BLS data would be analyzed at BLS and only results of statistical equations could be taken off -site.

Two key components-pension plan characteristics and pre-retirement earnings- are used to calculate replacement rates. The statistical equations developed at BLS are used to estimate pension plan characteristics as a function of job characteristics. These are statistically matched to SIPP individuals. SSA administrative data on earnings are used to develop two measures of earnings and to calculate DB benefit

[1] The MINT model is a micro-simulation model developed to estimate the distributional effects of proposed social security policy alternatives on current and future beneficiaries' retirement income. The MINT 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 RAND (Panis and Lillard, 1999), the Urban Institute (Toder et al., 1999), and Social Security Administration, Office of Research, Evaluation and Statistics (Butrica et al., 2001).

[2] 1993 was the last year BLS published replacement rates for full-time employees in medium and large private establishments and 1994 for State and local government employees.

[3] The Pension Insurance Modeling System model is a simulation model constructed by the Pension Guaranty Benefit Corporation of its pension insurance program.

amounts. DB benefit amounts and pre-retirement 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 have changed drastically as well as the type of pension plan providing the coverage. As recently as the mid-1990s the majority of individuals covered by a pension were covered by a DB plan. Currently, the majority are covered by a defined contribution (DC) plan. Not only has the type of pension changed but also the design of pension plans has changed (Mitchell 1998 and Papke 1999). Anew type of pension plan has evolved as well. Cash balance plans have gained popularity over the past few years (Elliott and Moore 2000). According to data recently released by BLS, cash balance plan participation has increased nearly fourfold between 1997 and 2000, from 6 percent to 23 percent.

Currently no dataset collects enough information to analyze these changes in pension plan coverage and design. The methodology in this paper, through a statistical match, brings together detailed information on pension plans and plan providers with survey data on plan participants and administrative data on earnings histories in order to improve the estimation of pension income for future retirees.

The arrangement of this paper is as follows: Section II presents the methodology, wherein a brief description of the key components of a DB plan is given and a description of the models used to replicate the EBS data are discussed. The subsequent section describes the data. Section IV discusses the statistical matching procedure and the assumptions. Section V presents results and section VI provides the conclusion.

II. Data

One of the major sources of data used in this study comes from the 1995 EBS - an employer-based survey. The EBS provides representative data on the incidence and detailed provisions of the nation's DB pension plans in all nonagricultural private-sector establishments employing 100 or more full- and part-time employees in all 50 states and the District of Columbia. The EBS sample used in the study contains 4,925 observations. For the average person, DB plan provisions are very difficult to interpret. [3] In light of this fact, Appendix A briefly describes some of the key provisions found in a DB plan, including the benefit formulas and some of their key components as well as eligibility requirements.

The SIPP provides data on representative samples of the nation's households. The SIPP collects data on sources and amounts of income, labor force information, program participation, eligibility data, and general demographic characteristics. This study focused on the data collected in the Retirement Expectations Pension Plan Coverage Topical Module and the Work History Topical Module. To make the SIPP analogous to the EBS, the SIPP sample is restricted to nonagricultural private-sector wage and salary workers who worked at an establishment with 100 or more employees and were covered by a DB plan. The self-employed are not included in this sample and individuals must have at least 5 years of employment on their current job. The sample also consists of individuals who were born between 1930 and 1955 who ranged in age from 40 to 65 in 1995. In accordance with the above restrictions, this sample has 2,508 observations for analysis.

Two sources of administrative earnings data are considered for the construction of the earnings measures – the detailed earnings record (DER) and the Summary Earnings Records (SER) maintained by the Social Security Administration. The DER contains wages, tips, other compensation, and deferred wage data from 1981 through 2001. These data are provided to the IRS on the Form W-2 from employers on all persons with wages including non-filers and other non-covered employees. The SER contains Social Security-covered earnings derived from payroll tax records for the years 1951 through 1999 (up to the taxable wage ceiling). After reviewing both datasets it was determined that the DER had significant advantages over the SER. One major advantage of using the DER is that there are earnings data for each job in each year,

[3] To learn more about DB plans and their features see work by Gerald Cole.

whereas the SER's earnings data is a sum of all earnings from all jobs in each year. By using the DER, it is possible to separate earnings out by job, which makes it possible to isolate one DB plan with the earnings from one job instead of having a sum of earnings from multiple jobs.

III. Methodology

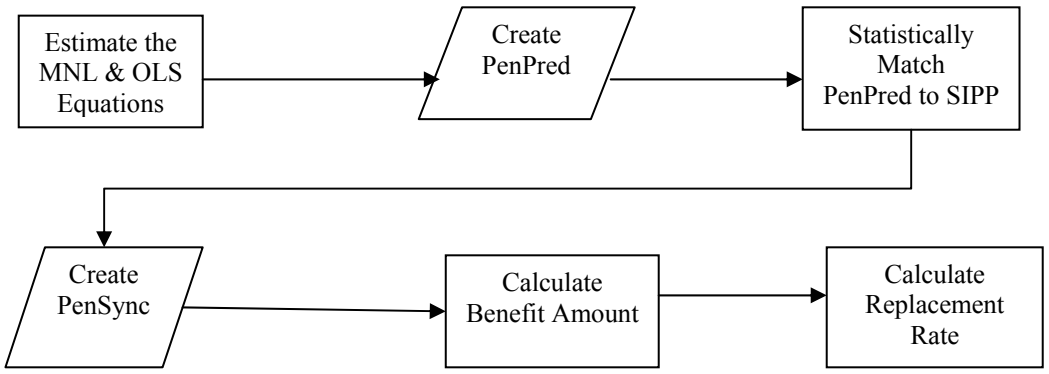
Overview

Diagram 1 shows the flow of the systematic procedures applied to create PenSync and 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. But, since the dependent variable that represents formula type 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 (Agresti, 1990). Therefore this study will use a multinomial logit (MNL) model to fit the categorical dependent variable.

The next step involves estimating the MNL and the OLS models to obtain coefficient estimates. The resulting coefficient 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 assigns a DB plan with detailed plan characteristics to the analytical sample of workers in the SIPP who reported being covered by a DB pension plan. The resulting dataset is called PenSync. The final two steps involve constructing an algorithm to calculate benefit amounts and calculating the replacement rate for each individual in the sample.

Diagram 1. The creation of PenSync and replacement rates



The following paragraphs provide more detailed explanations of the methodology applied in this study.

Model Specification

MNL Model Specification. The employer's choice of pension formula is modeled using MacFadden's random utility framework. Nine alternatives are identified: two flat dollar amount; four types of terminal earnings; two types of percent of career average; or a cash balance plan.[4] In choosing which type of formula to provide, employers may consider a variety of job characteristics of their employees, such as occupation and work schedule. The decision may also be affected by the characteristics of the employers, themselves, such as type of industry, number of employees, and presence of a union (see Table A.1 in Appendix A for the descriptive statistics of job characteristics variables used to model the choice of benefit formula). For any employer, the utility of choice j to employer i is expressed as:

[4] See Appendix A for a brief description of these alternatives.

$$U_{ij}=V_{ij}(E_i, W_i) + \varepsilon_{ij} \quad (1)$$

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 employee characteristics within the firm,
ε is a vector of unobserved components, and
j denotes pension formula alternatives.

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

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

expressed as:
$$P = P(U_{ij} > U_{ik}) \quad \text{for all } k<>j \quad (2)$$

By substitution of equation 1,
$$P = P(V_{ij} + \varepsilon_{ij} > V_{ik} + \varepsilon_{ik}, \text{ for all } k<>j)$$

Rearranging,
$$P = P[(\varepsilon_{ij} - \varepsilon_{ik}) > (V_{ij} - V_{ik}), \text{ for all } k<>j] \quad (3)$$

By knowing the distribution of the random ε's, we can derive the distribution of each difference ε_{ij}-ε_{ik} for all j, j<>k, and by using equation 3 we can calculate the probability that the employer will choose alternative j.
Letting $X_{ij}=(E_i, W_i)$ and assuming V is a linear function of components of X, we operationalize Equation 2 as:

$$U_{ij} = \beta_j X_{ij} + \varepsilon_{ij} \quad (4)$$

where β_j is a vector of coefficient values indicating the effect of the various X_{ij}'s on employer i's utility for option j. Note that β_j is subscript by the choice index j. This means that in the analysis a given X_{ij} is allowed to "interact" with each choice option. For example, union status may have one effect on the utility of choosing a dollar formula and another effect on the utility of choosing a cash balance plan.

As mentioned earlier, a 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 with a fixed dollar amount times years of service;
(2) Flat Dollar Amount times years of service with a varying dollar amount times years of service;
(3) Percent of terminal earnings with a fixed percent of earnings averaged over the last few year of employment;
(4) Percent of terminal earnings with varying percent of earnings averaged over a specified period of consecutive years of employment;
(5) Percent of terminal earnings with varying percent of earnings averaged over the last few year of employment;
(6) Percent of terminal earnings with a fixed percent of earnings averaged over a specified period of consecutive years of employment;
(7) Percent of terminal earnings with a fixed percent of earnings averaged over the employee career;
(8) Percent of terminal earnings with varying percents of earnings averaged over the employee career;
(9) Cash balance plan.

The MNL model is frequently used to analyze situations in which there are multiple choice alternatives. However, it is widely known that a potentially important drawback of the MNL model is the Independence from Irrelevant Alternatives (IIA) property; that is, the model can only be applied to situations where alternatives from which you can choose are totally independent.

To test for the existence of IIA, a model is constructed where 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, when the model is re-estimated it will not show a significant difference in explanatory power from the original model. The model used in this paper passed the IIA assumption.

This finding is not entirely surprising, given that there are many incentives embedded in the different types of pension formulas offered by employers. Some pension formula types are geared towards retaining employees, while others encourage retirement. Therefore, depending upon the incentive sought by the employer, his decision to offer a particular type of pension formula is independent. Again, the purpose of the IIA test is to ensure that the alternatives presented to employers are indeed viewed as independent. Interested readers should refer to Greene (1990), Train (1986), and Ben-Akiva and Lerman (1985). Therefore, 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 following MNL model.

$$BF_{ij} = \frac{e^{v_{ij}}}{\sum_{k=1}^{K} e^{v_{ijk}}} \quad \text{, where} \qquad (5)$$

BF_{ij} = the probability that the employer i choose formula j,
v_{ij} = $\sum \beta_m X_{ijm}$ = the deterministic component of utility of formula j to employer i,
X_{ijm} = the m-th explanatory variable for formula j and employer i, m= 1….M, and
β_m = coefficient to be estimated.

The MNL model includes information on characteristics of the employer, his employees and pension plan characteristics. For a description of the values of the dependent variable see Table A.2. In addition to predicting the formula type, the quantitative values common to each formula type is estimated using OLS.

OLS Model Specification. The quantitative variables for employer i and formula j can be written as:
$QVij = \beta_{0ij} + \beta_{1ij} X + \varepsilon_{ij}$; (6)

where:
QV_{ij} is a set of quantitative pension provision variables used in the pension benefit calculation and i denotes the ith employer. The coefficients are estimated by linear least squares-multiple regression. β_{0i} is a constant, X is a vector of job characteristics of the employer and his employees, pension plan characteristics, and ε_i is an error term. See Table A.3 for a listing and definition of the quantitative pension variables.

IV. Creating the Synthetic Pension File

As shown in diagram 1, the first two steps in creating PenSync involve fitting the MNL and OLS models to the EBS data set to score[5] a new data set of predicted observations (SAS, 2001). Table A.4 gives an overview of the accuracy of the MNL model. The model predicted the correct formula on average 71 percent of the time. Many of the incorrect predictions were among similar type of formulas. For example,

[5] For a description of SAS Proc Score procedure visit the following web address:
http://ftp.sas.com/techsup/download/stat/scorenew.html

the model predicted flat dollar formula with a fixed dollar amount (FDF) with a 95.77 percent accuracy rate, while only predicting flat dollar formula with a varying dollar amount (FVD) correctly 20.45 percent of the time. However, when the model incorrectly predicted FVD it predicted that it would be a FDF 50 percent of the time. FVD and FDF are very similar in their design and any attempts that were made to increase the prediction accuracy flawed the model with multicollinearity and over specification. The results from the OLS models are found in Table A.5.

To summarize the procedure, the step involved estimating equations 5 and 6 to generate a set of coefficient estimates, which are used to replicate the EBS data. The resulting coefficients estimates are used to produce predicted values by a process of multiplying the estimated coefficient by the observed EBS data. This multiplication process is repeated for each variable in the equations specified above. 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 to those of the EBS (See Appendix A.6).

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

There exists a substantial amount of research concerning the validity of using statistically matched data for analysis. Earlier work by Okner (1972 and 1974), Alter (1974), Radner, et al (1980), and Barry (1988) carefully documented some of the shortcomings of statistical matching. In particular, Benjamin Okner points out some of the common problems with statistical matching, which are data comparability, handling of missing data, specific techniques for matching, and the definition and evaluation of goodness of a match. The next few paragraphs briefly discuss the 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 criterions are verified. These criterions are well discussed in literature[6]:
(1) Unit harmonization: it is necessary that records of the different sources refer to the same unit. The unit of analysis for this study is workers.
(2) Target population harmonization: if the data set refers to different target populations, it is important to select just those records that refer to the population of interests. Both datasets are a sample of workers employed in private nonagricultural industries and occupations and participate in a DB plan.
(3) Variable harmonization: the common variables should be defined in the same way. Both datasets 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 cases with missing data and records without missing data, the more conservative approach to handling missing data is used. Hence, missing values are replaced with means for each variable (Little and Rubin, 1987, Kim and Curry, 1977, Roth, 1994).

[6] Statistical Matching: a tool for integrating data in National Statistical Institutes by Marcello D'Orazio, Marco Dizio and Mauro Scanu can be found on the following webpage: http://webfarm.jrc.cec.eu.int/ETK-NTTS/Papers/final_papers/43.pdf.

Selection of the Matching Variables

Let's begin with PenPred, henceforward called the universe "U". U consists of a set of N records. For each record there are values for R variables. U is represented by an N x 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 records. For each record there are values for the S variables that are represented by an M x 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, X = $(x_1,\ldots,x_p)$. The ith record in U, are denoted as U_i, which contains j observed variables, as shown below:

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

Similarly the ith, record in the SIPP contains h observed variables:

$SIPP_i = (SIPP_{il}\ SIPP_{i2}...\ SIPP_{ih})$ (8)

Where: x_1 = the worker 2 digit standard industry classification;[7]
x_2 = the worker 3 digit standard occupational classification;[8] and
x_3 = the worker union status.

The remaining variables in each of the files are referred to as Y on the PenPred file and Z on the SIPP file . Where Y = $(Y_1\ldots Y_q)$ and y_1 is a vector of predicted values of all pension provisions and Z= $(Z_1\ldots Z_r)$ and z_1 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. The distance function is defined as the absolute difference in the numerical value of the occupations and the union status of two cases: The distance between the ith worker in the U and the jth worker in the SIPP is defined by

$$Dij = \sum_{n=1}^{k} \left(I_{in} - I_{jn}\right) + \left(O_{in} - O_{jn}\right) + \left(U_{in} - U_{jn}\right) \quad (9)$$

Where: (n = 1,…,k) and

D_{ij} = the distance between the ith U record and the jth SIPP record.

$(I_{in} - I_{jn})$ = the distance between the values of the nth pair of industry variables in the ith record. (Since it is the cohort variable if must always equal 0)

$(O_{in} - O_{jn})$ = the distance between the values of the nth pair of occupation code variables in the ith record.

$(U_{in} - U_{jn})$ = the distance between the values of the nth pair of union status variables in the ith record.

Certain X variables may be treated as cohort variables. A cohort variable establishes subclasses of the records in each 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 on the SIPP file can only be matched to another worker in the mining industry in the U file.

Assumptions

This section presents assumptions relevant for the statistical match procedures.

[7] All workers are classified into one of over 82 industries according to their industrial classification.
[8] All workers are classified into one of over 820 occupations according to their occupational definition.

Assumption 1. No unobserved heterogeneity exists between the predicted data and the observed data. Stated differently, the probability associated with being covered by a given pension formula and having a particular set of job characteristics are analogous across the three datasets. This is an identifying assumption.

$$\pi(x,y|\ X, Data_{BLS}) - \pi\ (x,y,|\ X, Data_{SIPP}) - \pi\ (x,y,|\ X, Data_{PenSync}) = 0 \quad (10)$$

where x= type of pension plan, y = formula type, and X is a vector of individual job characteristics (e.g., 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 very similar to the predicted data. Cross tabulations also revealed similarities between the three datasets.

Assumption 2. Workers will remain on their current job until they reach the normal retirement age.

$$\pi(x,y|X_t\ , Data_{SIPP}) - \pi(x,y|X_{t+i}, Data_{SIPP}) = 0, \quad (11)$$

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

Many DB 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 DB plan. By asserting that workers will remain on their current job, an obvious assumption is that 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 DER data were used to measure tenure on the current job and the frequency of job change. The SIPP data reveals that the average tenure on the current DB pension job was 18 years and the DER data reveals that between the start year of the current job and 2003, 63 percent of the workers in the sample remained with their same employer. To further test these assumptions, the SIPP data is used to check how often a worker reports changing industry and/or occupations. When analyzing the full panel of the SIPP, 92 percent and 90 percent of the workers report remaining in the same industry and occupation, respectively.

Assumption 3. The SIPP reported pension job for employer 1 is the highest earnings job in the W-2 file in each year.

$$\pi(x,y|\ X_t, Data_{DER}) - \pi(x,y|\ X_t, Data_{SIPP}) = 0 \quad (12)$$

where X = earnings in a given year and t= 1951…2002.

This assumption assumes that the pension module job 1[9] on the SIPP is the same as the DER job reporting the highest wage. SIPP respondents are asked the question about calendar-year wage and salaries twice per panel. Respondents are encouraged to refer to their respective W-2 form or other documents to ensure accuracy. To test the validity of this assumption, the earnings total reported in the SIPP for the pension job is compared to the highest wage job in the DER for the same year. The SIPP reported earnings are very similar to the DER highest earnings, varying by plus or minus $2,000 annually. Respondents in the SIPP can also report earnings and pension coverage from two employers; therefore, to further ensure that the probability that the pension job reported for employer 1 is indeed the DER highest wage job, analysis is conducted on the second job reported in the SIPP. The analysis reveals that less than 3 percent of the unweighted individuals who reported having a DB type pension reported having a DB pension on their second job.

The Matching Algorithm

The match procedure is unconstrained. This 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 MOU confidentially provision, particular attention is given to tabulations based on

[9] The SIPP asks respondents about two jobs.

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 (i.e. industry, occupation, and union status), designate the pair as a Level 1 Match, else if the pair agrees on two of the characteristics (i.e. industry and occupation), designate the pair as a Level 2 Match, else if the pair agrees on two of the characteristics (i.e. industry and major occupational group (MOG)), designate the pair as a Level 3 Match, else if the pair agrees on one of the characteristics (i.e. industry), designate the pair as a Level 4 Match, else designate the pair as a non-match. As shown in Table 1, the final data file for analysis consists of 2508 observations, which contains detailed socioeconomic variables along with in-depth employer-provided pension data. This database is called PenSync.

Table 1. Number and percent of matches by the number of levels required for the match.

Levels	Number of matches	Match rate
Level 1	1876	75%
Level 2	192	8%
Level 3	430	17%
Level 4	10	.004%
Total	2508	100%

Source: Author's calculation

Benefit Algorithm

The final procedure in this study involves constructing an algorithm to calculate benefit amounts and the replacement rates for each individual in PenSync. The algorithm starts by determining the type of formula assigned to an individual (e.g. career average earnings, terminal earnings, cash balance, or flat dollar). For individuals covered by a percent of earnings times years of service formula, a subroutine is initiated to determine whether the earnings are based upon a career average or terminal earnings. For individuals covered by a career average arrangement, the benefit amount is determined by multiplying a proportion of the average DER earnings by the workers' total number of credited years of service.[10] For individuals whose benefit amounts are based upon a terminal earnings arrangement, the algorithm multiplies a proportion of the average DER earnings in a specified period of time, typically near the individuals' retirement age.

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

The final step in the algorithm produces a set of pension benefits and replacement rate ratios for the two measures of earnings, the last ten years of earnings (L10yr) and the last five years of earnings (L5yr). L10yr is the average of the five highest years of earnings 10 years prior to the pension plan normal retirement date. L5yr is the average of the three highest years of earnings 5 years prior to the pension plan normal retirement date. The pension plan normal retirement date is the year in which the worker satisfies provisions specified in their pension plan in order to receive an unreduced retirement benefit. The year

[10] For all individuals regardless of formula type, the number of credited years of service is determined by subtracting the pension plan normal retirement year from the year the workers reported starting his/her current job. For years of earnings that are outside the scope of the DER, the SER is used to supplement the missing data.

2003 is used to verify whether an individual has satisfied the pension plan normal retirement requirement. All earnings and benefit amounts are measured in 2003 dollars.

V. Results

For workers who are eligible for normal retirement benefits prior to 2003, their DB plan is estimated to replace about 30 percent of last earnings. The average earnings are estimated to be about $35,000 and the average monthly pension benefit is $1012 (table 2). 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 the lowest for those in flat dollar or career average formulas and highest for those in terminal earnings formulas or cash balance formulas with a 16-17 percentage point differential. Replacement rates were considerably lower for those in administrative/clerical or production/service jobs compared to those in professional/technical jobs, and lower for those in goods producing industries than those in nongoods producing industries. Union members are estimated to have higher replacement rates than nonunion 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 over 60 percent of their earnings in the five years prior to retirement replaced by their pension plans compared with only a 9 percent replacement rate for those with fewer than 10 years of participation.

VI. Conclusion

Predicting retirement income from a pension plan is a difficult task. The lack of good data is a major contributor. Furthermore, the lack of comprehensive pension data sources poses limitations on pension research and policy decisions. The methodologies applied in this paper have been in existence for decades. However, they still remain more of an art than a science. However, many challenges are inherent when employing such a procedure. They include model specifications, harmonization, and probably most importantly, the quality of the data involved. Nevertheless, the methodology is a reasonable approach given constraints from two different restricted data sets.

Table 2. Pension Income and Replacement Rate for Workers Who Qualify for Normal Retirement Prior to 2003.

Items	Percent	Average Earnings		Monthly Benefit	Replacement Rates	
		High 3 of last 5	High 5 of last 10		High 3 of last 5	High 5 of last 10
		Dollars			Percent	
All	100	37,958	32,649	1,012	32	29
Formula Type						
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
NonGoods 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
Greater than 30	22	47,428	41,674	1,840	61	41
Union Status						
Nonunion 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 three highest years of earnings 5 years prior to the pension plan normal retirement date. High 5 of last 10 is the average of the five highest years of earnings 10 years prior to the pension plan normal retirement date. All earnings and benefit amounts are measured in 2003 dollars. Eligibility for retirement depends on a workers age or the 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 provision specified in their pension plan to receive 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.

References:

Alter, Horst E. 1974, "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.* 3(2): 373-394.

Agresti, A. 1990. *Categorical Data Analysis*. New York, NY: J. Wiley & Son.

Barry, J. T. 1988. "An Investigation of Statistical Matching." *Journal of Applied Statistics* (15): 275-283.

Ben-Akiva, Moshe, and Steven Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand.* Cambridge, MA: MIT Press 4th printing 1991.

Bureau of Labor Statistics, U.S. Department of Labor. 1993. *Employee Benefits in Medium and Large Private Establishments.* Washington, D.C.: US GPO.

Butrica, Barbara, A., Howard M. Iams, James Moore, and Mikki Waid. 2001. *Methods in Modeling Incomei n the Near Term (MINT)*. ORES Working Study No. 91. Office of Policy, Social Security Administration. May.

Cook R.D. 1979. "Influential Observations in Linear Regression." *Journal of the American Statistical Association*. 74:169-174.

Elliot, Kenneth R., and James H. Moore. 2000. "Cash Balance Pension Plans: the New Wave." *Compensation and Working Conditions.* 5(2): 3-12.

Green, W.H.1990. *Econometric Analysis*. New York, NY: Macmillan.

Little, R. J. A. and Rubin, D. B. 1987. *Statistical Analysis with Missing Data*. New York: J. Wiley and Sons

Kim, J. O., & Curry, J. 1977. "The treatment of missing data in multivariate analysis.
*Sociological Methods and Researc*h. 6: 215-240.

McFadden, D. 1973. "Conditional logit analysis of qualitative choice behavior." *Frontiers in Econometrics,* edited by P. Zarembka. New York, NY: Academic Press, Inc.

Mitchell, Olivia S. 1998. "Developments in Pensions," NBER Reporter, National Bureau of Economic Research.

Papke, Leslie E. 1999. "Are 401(k) Plans Replacing Other Employer-Provided Pensions? Evidence from Panel Data." *Journal of Human Resources.* 34(2): pp. 346-68.

Okner, B. A. 1972. "Constructing a New Data Base From Existing Microdata Sets: The 1966 Merge File." *Annuals of Economic and Social Measurement.* 1(3): 325-342.

Radner, D.B., Allen, R., Gonzalez, M E. Jabine, T. B., and Muller, H.J. 1980. "Report on Exact and Statistical Matching Techniques." Statistical Policy Working Paper, U.S. Dept. of Commerce, Washington, D.C.: U.S. GPO

Rogers, Willard L. 1984. "An Evaluation of Statistical Matching." *Journal of Business & Economic Statistics.* 2(1): 91-102.

Roth, P. L. 1994. "Missing data: A conceptual view for applied psychologists." *Personnel Psychology*. *4*7: 537-560.

Rubin, D. B. 1986. "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. "*Journal of Business and Economic Statistics.* 4:87-94.

SAS Institute Inc. 2001. 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.

Toder, Eric and others. 1999. " 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.

Train, K. 1986. *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*. Cambridge, Mass.: MIT Press.

Appendix A.
Brief description of DB provisions

A DB plan provides employees with guaranteed retirement benefits based on a predetermined benefit formula. There are three basic types of DB formulas found in the EBS data: (1) a percent 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 DB plan are covered by a percent of earnings per year of service formula (BLS, 2000)[11]. 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 credited years of service may be based upon either a career average or final earnings. Under a career average arrangement the plan benefits are based on the average of the earnings paid over the entire period of the employee's participation in the plan. On the other hand, under a final pay arrangement, the plan benefits are based on an average of the employee's earnings during a short period of time, typically near the employee's retirement age. For example, the earnings may be averaged over the last three or five years of employment, or over the three or five 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 DB plan whereby the benefit formula takes into account the employee's income and number of credited years of service. Although a cash balance plan is structured to bear resemblance to a defined contribution plan where 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 not associated with income, but rather a dollar amount per year of service. In 2000, 14 percent of all workers in the private sector who were covered by a DB plan had this type of plan. A flat dollar amount per year of service formula 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/she must become vested, which means having a designated number of years of service with an employer. A five-year cliff-vesting requirement is the most prevalent provision. Therefore, this study assumes that an individual, upon satisfying the five-year vesting requirement, is entitled to receive a non-forfeitable accrued benefit upon separation or retirement.

Benefits under a DB plan are usually paid when the employee retires. All DB 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. However, many DB plans allow early 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.

[11] This data can be found at http://www.bls.gov/ncs/ebs/sp/ebrp0001.pdf.

A.1. Descriptive Statistics for Job Characteristics Variables

Category	Number	Percent
Industry		
Mining	56	1.14
Construction	49	0.99
Manufacturing	1330	27.01
Transportation	804	16.32
Wholesale	154	3.13
Retail	444	9.02
Finance	1106	22.46
Service	982	19.94
Occupational Groups		
Professional	1564	31.76
Blue Collar	1652	33.54
Clerical	1709	34.7
Union Status		
Non Union Member	3547	72.02
Union Member	1378	27.98
Work Schedule		
Part-time	308	6.25
Fulltime	4617	93.75
Employment		
Less than 250	922	18.72
250-499	754	15.31
500-999	886	17.99
1,000 or greater	2363	47.98
Number of Observations	4925	

Source: Author's calculation using EBS data

A.2. Description of the values for the MNL dependent variable

Value	Formula Type
1	Flat Dollar Amount times years of service with a fixed dollar amount times years of service
2	Flat Dollar Amount times years of service with a varying dollar amount times years of service
3	Percent of terminal earnings with a fixed percent of earnings averaged over the last few year of employment
4	Percent of terminal earnings with varying percent of earnings averaged over a specified period of consecutive years of employment
5	Percent of terminal earnings with varying percent of earnings averaged over the last few year of employment
6	Percent of terminal earnings with a fixed percent of earnings averaged over a specified period of consecutive years of employment
7	Percent of terminal earnings with a fixed percent of earnings averaged over the employee career
8	Percent of terminal earnings with varying percents of earnings averaged over the employee career
9	Cash balance plan

A.3 Definitions of Quantitative Variables

DOL_DOL1	1st dollar amount breakpoint use to calculate a flat dollar formula
DOL_DOL2	2nd dollar amount breakpoint use to calculate a flat dollar formula
DOL_DOL3	3rd dollar amount breakpoint use to calculate a flat dollar formula
DOL_YRS1	1st YOS breakpoint use to calculate a flat dollar formula
DOL_YRS2	2nd YOS breakpoint use to calculate a flat dollar formula
NORM_AAS	Sum of normal retirement age and service
NORM_AGE	Normal retirement age
NORM_SRV	Normal retirement service requirement
NR_PAY	Percent of earnings contribute to a Cash Balance plan
NR_INT	Interest rate
EBASEYR1	1st breakpoint for number of years to be included in the calculation of benefits
EBASEYR2	2nd breakpoint for number of years to be included in the calculation of benefits
POE_DOL1	1st breakpoint number of years to be included in the calculation of benefits
POE_DOL2	2nd dollar amount breakpoint use to calculate a percent of earnings formula
POE_PCT1	1st percent of earnings breakpoint use to calculate a percent of earnings formula
POE_PCT2	2nd percent of earnings breakpoint use to calculate a percent of earnings formula
POE_PCT3	3rd percent of earnings breakpoint use to calculate a percent of earnings formula
POE_PCT4	4th percent of earnings breakpoint use to calculate a percent of earnings formula
POE_PCT5	5th percent of earnings breakpoint use to calculate a percent of earnings formula
POE_YRS1	1st breakpoint number of years of service to be included in the calculation of benefits
POE_YRS2	2nd breakpoint number of years of service to be included in the calculation of benefits

Table A.4 Prediction Accuracy Predicted Compared to Observed Formula Types

	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	
Frequency	1	816	6	0	14	0	1	2	1	12	852
Percent		95.77	0.7	0	1.64	0	0.12	0.23	0.12	1.41	
	2	22	9	0	13	0	0	0	0	0	44
		50	20.45	0	29.55	0	0	0	0	0	
	3	0	0	112	0	43	0	0	0	0	155
		0	0	72.26	0	27.74	0	0	0	0	
	4	1	1	2	1182	0	207	1	1	0	1395
		0.07	0.07	0.14	84.73	0	14.84	0.07	0.07	0	
	5	0	1	29	1	315	1	0	0	0	347
		0	0.29	8.36	0.29	90.78	0.29	0	0	0	
	6	0	3	4	473	0	1099	6	10	0	1595
		0	0.19	0.25	29.66	0	68.9	0.38	0.63	0	
	7	0	0	0	0	0	6	11	0	0	17
		0	0	0	0	0	35.29	64.71	0	0	
	8	0	0	0	0	0	132	0	83	0	215
		0	0	0	0	0	61.4	0	38.6	0	
	9	34	0	0	0	0	0	1	0	270	305
		11.15	0	0	0	0	0	0.33	0	88.52	
Predicted Total		873	20	147	1683	358	1446	21	95	282	4925

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

Table A.5. Regression Results for Selected Quantitative Variables

	Constant	Size	Industry	Work Schedule	Occupation	Union Status	Dollar Formula	Career Average	R^2
dol_dol1	5.0851 (.80890)**	-0.0005 (.00001)**	-2.862 (.3666)**	-2.0372 (.4234)**	1.2767 (.2336)**	0.3024 (.2616)	31.8015 (.5091)**	0.7117 (.4262)**	.74
CB percent	4.5894 (.0735)**	0.0001 (.00001)**	0.164 (.0322)**	-0.0600 (.0372)	-0.0032 (.0205)	-0.0346 (.023)	-4.8377 (.0447)**	-4.8791 (.0375)**	.79
CB interest	5.26057 (.076)**	-0.0001 (.00001)	0.0044 (.0333)	0.043 (.0385)	0.0502 (.0212)	0.016 (.0238)	-5.2488 (.0462)**	-5.2148 (.0387)**	.79
POE 1	-2.6099 (.480)**	0.0002 (.00005)*	-0.3918 (.2103)	1.8657 (.2429)**	0.6683 (.1340)**	0.8312 (.1501)**	-0.3176 (.2921)	12.9813 (.2445)**	.67
POE 2	0.2800 (.0911)*	0.00002 (.000009)	0.1202 (.0399)	-0.054 (.0461)*	-0.0807 (.0254)	-0.2721 (.0285)**	-0.1862 (.0554)*	0.5662 (.0464)**	.18
Years 1	-0.3143 (.2185)	0.0001 (.000002)**	0.3194 (.0957)*	0.0678 (.1106)	-0.062 (.0610)	0.0314 (.0683)	-0.3266 (.133)	3.3456 (.1113)**	.41
Years 2	-4.3253 (3.9373)	-0.0006 (.0004)	4.3718 (1.7254)	8.346 (1.993)**	-1.8145 (1.1)	3.6991 (1.2312)	-6.4945 (2.3964)	26.0477 (2.0059)**	.12
Norm_age	46.606 (2.01)**	0.001564 (.0002)**	5.454 (.88)**	-3.20707 (1.01)*	-2 (.56)*	-2.98348 (2.98)**	-2.8452 (1.22)	7.651 (1.02)**	.09
Norm_srv	10.629 1.94)**	-0.00152 (.0001)**	-6.373 (.523)**	3.71762 (.604)**	1.3416 (.333)**	2.67692 (.7.)**	6.3605 (.723)**	1.856 (.61)	.10

** Significant at 1% statistical level and * Significant at 5% statistical level.

A.6. 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	0.04	0.09	-0.05	0.20	1.44	-1.25
Dol_dol3	0.66	0.46	0.19	1.10	5.20	-4.10
Dol_yrs1	0.15	0.11	0.04	0.36	1.14	-0.78
Dol_yrs2	0.05	0.11	-0.06	0.22	1.81	-1.59
Norm_aas	5.32	5.30	0.02	2.03	20.10	-18.07
Norm_age	57.38	57.33	0.04	5.29	17.77	-12.49
Norm_srv	7.89	7.91	-0.02	3.23	10.59	-7.36
Nr_pay	0.31	0.30	0.01	1.21	1.34	-0.13
Nr_int	0.31	0.32	-0.01	1.21	1.41	-0.20
Ebaseyr1	2.97	2.79	0.18	1.70	2.40	-0.71
Ebaseyr2	21.24	20.76	0.48	11.67	35.52	-23.85
Poe_dol1	243.58	234.11	9.47	146.37	1877.95	-1731.58
Poe_dol2	0.00	0.00	0.00	0.00	0.00	0.00
Poe_pct1	10.19	10.24	-0.04	5.64	7.03	-1.39
Poe_pct2	0.76	0.67	0.09	0.43	0.85	-0.42
Poe_pct3	0.00	0.18	-0.18	0.00	0.43	-0.43
Poe_pct4	0.00	0.02	-0.02	0.00	0.14	-0.14
Poe_pct5	0.00	0.04	-0.04	0.00	0.21	-0.21
Poe_yrs1	5.40	5.22	0.18	2.91	11.30	-8.39
Poe_yrs2	0.50	0.43	0.06	0.50	2.28	-1.78

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