LONGITUDINAL EMPLOYER - HOUSEHOLD DYNAMICS

TECHNICAL PAPER NO. TP-2002-24

Estimating Measurement Error in SIPP Annual Job Earnings: A Comparison of Census Survey and SSA Administrative Data

Date	:	September 2002
Prepared by	:	Martha Stinson
Contact	:	U.S. Census Bureau, LEHD Program FB 2138-3 4700 Silver Hill Rd. Suitland, MD 20233 USA

 This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging, and the Alfred P. Sloan Foundation.

Estimating Measurement Error in SIPP Annual Job Earnings: A Comparison of Census Survey and SSA Administrative Data*

Martha H. Stinson†

September 12, 2002

Abstract

The third chapter investigates measurement error in SIPP annual job earnings data linked to SSA administrative earnings data. The multiple earnings measures provided by the survey and administrative data enable the identification of components of true variation and variation due to measurement error. We find that 18% of the variation in SIPP annual job earnings can be attributed to measurement error. We also find that in both the SIPP and the DER, measurement error is persistent over time. A lower level of auto-correlation in the SIPP measurement error than in the economic error component leads to a lower reliability ratio of .62 for first-differenced earnings.

1 Introduction and Background

Economists and statisticans have long recognized that survey data are prone to measurement error. Responses to questions about earnings, education levels, and job characteristics are not measured exactly but instead contain some truth and some error. The classical measurement error model as described by Fuller (1987) defines a dependent variable Y_t that is a linear function of a covariate x_t.

*I would like to thank John Abowd, Simon Woodcock, Kevin McKinney, Kristin Sandusky, Lars Vilhuber, and Marc Roemer for helpful comments and support. This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging, and the Alfred P. Sloan Foundation. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the U.S. Census Bureau, Cornell University, or the National Science Foundation. Confidential data from the LEHD Program were used in this paper. The U.S. Census Bureau is preparing to support external researchers' use of these data under a protocol to be released in the near future.

†Martha Stinson is an economist at the U.S. Census Bureau. Contact information: martha.stinson@census.gov; LEHD Program, Mailstop 8400, U.S. Census Bureau, Washington, D.C. 20233

$$Y_t = \beta_0 + \beta_1 x_t + e_t$$

However x_t is not observed directly, and instead we see

$$X_t = x_t + u_t$$

where x_t is the true value of the covariate and u_t is the measurement error. By assuming that the measurement error, the true values, and the errors are independently distributed as

$$\begin{bmatrix} x_t \\ e_t \\ u_t \end{bmatrix} \sim N\left\{ \begin{bmatrix} \mu_x \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{xx} & 0 & 0 \\ 0 & \sigma_{ee} & 0 \\ 0 & 0 & \sigma_{\mathrm{u}\,u} \end{bmatrix} \right\}$$

the joint distribution of Y and X can be written as

$$\begin{aligned} E\{Y, X\} &= (\beta_0 + \beta_1 \mu_x, \mu_x) \\ Var(Y, X) &= \begin{bmatrix} \beta_1^2 \sigma_{xx} + \sigma_{ee} & \beta_1 \sigma_{xx} \\ \beta_1 \sigma_{xx} & \sigma_{xx} + \sigma_{\mathrm{u}\,u} \end{bmatrix} \end{aligned}$$

When Y is regressed on X, the expected value of the estimated regression coefficient $\widehat{\beta_1}$ is attenuated.

$$E\{\widehat{\beta_1}\} = \beta_1 \frac{\sigma_{xx}}{(\sigma_{xx} + \sigma_{\mathrm{u}\,u})}$$

The ratio

$$\kappa_{xx} = \frac{cov(u, X)}{Var(X)} = \frac{\sigma_{xx}}{(\sigma_{xx} + \sigma_{\mathrm{u}\,u})}$$

is often called the reliability ratio and it defines the ratio of $\widehat{\beta_1}$ to β_1. The proportional attenuation bias resulting from measurement error is defined as $\frac{\beta_1 - \widehat{\beta_1}}{\beta_1} = 1 - \kappa_{xx}$.

The bias resulting from measurement error can be exacerbated if one is using first differenced data. As Angrist and Krueger (1999) describe, the reliability ratio for a variable $\triangle X = (X_t - X_{t-1}) = (x_t - x_{t-1}) + (u_t - u_{t-1})$ is equal to

$$\kappa_{\triangle x \triangle x} = \frac{\sigma_{xx}}{\sigma_{xx} + \sigma_{\mathrm{u}\,u}(\frac{(1-\tau)}{(1-\rho)})}$$

where τ is the auto-correlation coefficient of the measurement error and ρ is the auto-correlation coefficient of the true value of earnings. If $\rho > \tau$ then $\frac{(1-\tau)}{(1-\rho)}$ is greater than one and the signal to noise ratio declines. Thus determining the extent to which measurement error persists over time is important in assessing the impact on the estimated coefficient.

If the variance and structure of the measurement error is known, then unbiased estimators of β_1 can be obtained. Hence those studying measurement error have focused on estimating κ_{xx} and testing whether the assumptions of classical measurement error were violated. Studies that obtain a second report for the mismeasured variable of interest in order to calculate σ_{uu} and σ_{xx} have been termed validation studies. The most common approach is to view this second report as "truth" and calculate the measurement errors directly as $u = X - x$. The properties of these errors can then be investigated and researchers have often concluded that the assumptions of classical measurement errors were violated and that the errors were correlated with the true values, i.e. $\sigma_{xu} \neq 0$. However they acknowledge that their models were driven by the assumption that they obtained a true measure of x. Without this assumption, there would be no way to determine the relationship between the errors and the true values. This assumption is also fundamentally untestable and is justified solely by the authors' knowledge of the quality of the secondary data source.

One of the first validation studies was done by Mellow and Sider (1983) using a special supplement to the January 1977 CPS that obtained name and address information of employers from the survey respondents[1]. Matched pairs with both employer and employee wage reports totaled 3,612. In this data set, employer-reported wages exceeded worker reports by 4.8% on average. In order to test the sensitivity of statistical models to the source of the variables used, the authors estimated four different wage regressions. In the first two wage regressions, respondent-reported variables for union status, industry and occupation were regressed on worker and employer reported wages, respectively. In the second two wage regressions, employer-reported union status, industry and occupation were regressed alternatively on worker and employer reported wages. Returns to education and experience were strikingly constant across these four equations. The nonwhite-white differential was smaller when using employer-reported wages while the female differential was higher. The union wage-premium was smaller when using employer-reports of union coverage. Occupation and industry differentials were very similar across the different specifications. The authors concluded that the wage regressions were generally not that sensitive to the source of information: worker versus employer.

During the 1980s, a validation study at a large anonymous manufacturing company was undertaken. Results from this study were reported in Duncan and Hill (1985) and Bound et al. (1994). Workers at the company were interviewed using a PSID survey instrument and then information for these workers was obtained from company records. Bound et al. provided a comprehensive report on both waves. The first wave of data was collected in the summer of 1983 and included 418 workers and the second wave was conducted in 1987 with 341 of the originally interviewed workers and an additional 151 new workers. The authors treated the company reports of annual earnings as measures of true earnings

[1] Mellow and Sider also evaluate a second matched data set: the Employment Opportunity Pilot Project (EOPP). However this data set contains only general firm data such as industry and union status matched to specific workers and hence it is not possible to compare earnings reports from both the employer and employee using this data set.

values and considered any differences between worker and employer reports to be errors on the part of the workers. According to the authors, "We do this because of our confidence in the accuracy and recording of the company records, in part because of the extraordinary cooperation of the company involved. This is crucial, because if there were significant errors in the company records, one would have no way of knowing how they were correlated with other variables." By their own acknowledgement, the results in this paper were completely driven by this assumption.

The authors reported a noise to total variance ratio ($\frac{\sigma_{\mathrm{u}\,u}}{(\sigma_{xx}+\sigma_{\mathrm{u}\,u})}$ in the notation above) of .302 for annual earnings in 1986 and .151 for annual earnings in 1982. They argued that this ratio was misleading because the errors in earnings were correlated with the true levels of earnings. In this case the true variance ratio should be

$$\frac{cov(X,u)}{var(X)} = \frac{\sigma_{\mathrm{u}\,u} + cov(u,x)}{\sigma_{xx} + \sigma_{\mathrm{u}\,u} + 2cov(u,x)}$$

This ratio was calculated by regressing the errors on the employee-reported annual earnings and was .239 in 1986 and .076 in 1982. Thus the authors claimed that when earnings measures are used as independent variables in regression analyses, the bias resulting from meaurement error will be mitigated by correlation between errors and true values.

Generally measurement error in a dependent variable will not cause bias in the regression coefficients but will make them less precise by increasing the overall variance of Y. However the correlation between the true value and the error of a dependent variable will introduce bias even if the independent variables are measured without error. The authors described this result in the following way:

$$\begin{aligned} Y &= (1+\delta)y + v = x\beta + \varepsilon \\ \widehat{b} &= \frac{(1+\delta)cov(y,x)}{var(x)} \\ \frac{\widehat{b}}{\beta} &= (1+\delta) \end{aligned}$$

Thus the proportional attenuation bias in the coefficient is δ which was estimated as -.172 for 1986 and -.104 for 1982. Again the calculation of these results was completely dependent on the strategy used to identify the errors separately from the true value of earnings.

The authors concluded by estimating two earnings equations, one using employee reported measures of earnings and tenure and the other using company recorded measures of the same variables. Education and experience were also included in the regressions. Since only one measure of education and experience was available (employee interview responses), these variables were considered measured without error. Regression coefficients from the worker-reported equation were measured against the "true" coefficients from the company-reported equation. According to this standard, the interview data overstated the return to education by 40% and the understated the return to tenure by 20%.

Bound and Krueger (1991) conducted a similar validation study using linked CPS-Social Security Earnings Records. March 1978 CPS respondents were asked to report their Social Security Numbers and, using SSN, name, age, sex, and race, respondents were linked to SSA records. About 50% of respondents who were in both the 1977 and 1978 March CPS were successfully linked to SSA data. This study was complicated by the fact that SSA earnings reports were truncated at the maximum Social Security taxable earnings amount ($16,500 in 1977 and $15,300 in 1976). The authors made the same error-identifying assumptions as Bound et al. Administrative records were viewed as truth with the exception that the truth was sometimes truncated. Thus the authors first estimated the relationship between the SSA and CPS earnings using a Tobit maximum likelihood approach which accounted for the truncation. The results from this estimation were used to calculate the variance/covariance matrix between CPS earnings and true SSA earnings. This matrix in turn was used to compute a variance/covariance matrix between x_t and u_t. The authors reported large negative correlations between measurement error and true earnings for both 1976 and 1977 (-.46 and -.42 respectively). They reported reliability ratios which did and did not take account of these correlations as .844 and 1.016 respectively for 1976 and .819 and .974 for 1977. They also noted that the reporting errors appeared to be positively correlated over time but "with only 2 years of data it is impossible to distinguish an autoregressive process in the measurement error from a person fixed effect or from other time-series processes."

Bound, Brown, and Mathiowetz (2001) summarized earnings validation studies and stated that the ideal information for correcting measurement error would be to know the joint distribution of all the true and observed variables, i.e. $f(y, x, Y, X)$. However the authors recognized that information about this joint distribution has often come at the cost of assuming that validation data is truth. They write, "Those collecting validation data usually begin with the intention of obtaining "true" values against which the errors of survey reports can be assessed; more often than not we end up with the realization that the validation data are also imperfect. While much can still be learned from such data, particularly if one is confident the errors in the validation data are uncorrelated with those in the survey reports, this means replacing one assumption (e.g. errors are uncorrelated with true values) with another (e.g. errors in survey reports uncorrelated with errors in validation data)."

Bound, Brown, and Mathiowetz also expressed the hope that future validation studies would be able to obtain secondary data reports for multiple consecutive years. Past validation studies have been able to create panels of earnings measures for at most two consecutive years. Thus it has been difficult to calculate the correlation of errors over time, an important component to assessing the impact of measurement error on panel data. Due to the high cost of validating panel data, the authors forsee the future of validation studies as being critically enhanced by opportunities to "merge administrative data to existing panel data."

This research will follow in the tradition of validation studies but four major innovations will be introduced. First, a new linked survey-administrative database will be used. Second, the administrative data will not be viewed as the measure of true earnings and a methodology will be developed to quantify measurement error without this assumption. Third, earnings records will be linked at the job level, allowing information about the identity and characteristics of the employer to be used. Fourth, earnings records from four consecutive years will be compared, thus providing valuable new insight into the time series properties of measurement error.

2 Model

The goal of this paper is to estimate levels of measurement error in SIPP survey data using an alternative source of earnings data: SSA administrative earnings records. Unlike past studies, the administrative records will not be treated as a measure of true earnings. Instead, both sources of earnings data will be treated as noisy measures of some underlying true value of earnings. Measurement error will be estimated by decomposing both measures of earnings into shared effects and separate effects. The shared effects will include the observable effects of general labor force experience and time as well as the unobservable effects of individual and firm heterogeneity. In addition there will be a shared error component which can be thought of as a nested individual/job/time period random effect. This effect is estimable due to the presence of two earnings observations for each year of the panel. It represents "economic" noise, or fluctuations in annual earnings due to unobservable economic factors which influence true earnings as opposed to reported SIPP or SSA earnings. The separate effects are then attributed to measurement error, as these effects are due to things which do not influence the underlying true value of earnings. Using the estimated variance components, the reliability ratios of both the SIPP and SSA earnings variables can then be calculated as the ratio of true to total variance.

This modeling follows the spirit of Abowd and Card (1989). Using several different long panel datasets, they first-difference earnings and hours and adjust for experience. They then examine the variance/covariance matrix of these differences and test the fit of various structural models, all of which include a measurement error component. This model will rely on random person and firm effects instead of first-differencing and has the advantage of a second source of data to identify the measurement error but the parsing of variance among structural components is the similar.

Given this statistical model, the SIPP earnings equation for a given individual i is:

$$\ln\left(SIPPEARN_{ist}\right) = \beta_{oSIPP} + \beta_1 Exp_{it} + \beta_2 Time_{it} + \theta_i + \psi_j + \eta_{ist} + \omega_{ist} \quad (1)$$

and the SSA earnings equation for the same individual is identical except for the last component:

$$\ln(SSAEARN_{ist}) = \beta_{oDER} + \beta_1 Exp_{it} + \beta_2 Time_{it} + \theta_i + \psi_j + \eta_{ist} + \upsilon_{ist} \quad (2)$$

where i subscripts the individual, j subscripts the firm, s subscripts the person-firm match or job, and t subscripts the year. The variables are defined as follows:

$$\begin{aligned}
Exp_{it} &= \text{general labor market experience (annual)} \\
Time_{it} &= \text{calendar time} \\
\text{Person heterogeneity} &= \theta \sim N(0, G_1) \\
\text{Firm heterogeneity} &= \psi \sim N(0, G_2) \\
\text{Common error component} &= \eta \sim N(0, G_3) \\
\text{Measurement error, SIPP and SSA} &= (\omega, \upsilon) \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, R) \\
& \quad G_1, G_2, G_3, R \text{ defined below}
\end{aligned}$$

The total number of jobs held by all individuals is N, the total number of individuals is I, the total number of firms employing individuals in the sample is J, the number of covariates included in X is k, and the maximum number of possible time periods for any job is 4.

Written in matrix notation, the model is

$$Y = X\beta + Zu + e$$

where Y is an $(N \times 4 \times 2) \times 1$ vector of stacked SIPP and SSA earnings, X is an $(N \times 4 \times 2) \times k$ matrix of covariates treated as fixed effects, β is a $k \times 1$ vector of fixed effects coefficients, Z is an $(N \times 4 \times 2)x(I + J + N \times 4)$ design matrix of the random effects, u is a $(I + J + N \times 4) \times 1$ vector of random effects and e is an $(N \times 4 \times 2) \times 1$ vector of residuals. The random effects vector, u, contains the stacked random effects, $\theta_1 ... \theta_I$, $\psi_1 ... \psi_J$, $\eta_{111996} ... \eta_{IN1999}$. The design matrix of the random effects, Z, contains one column for each individual, one column for each firm, and one column for each individual-firm-time period match. The error vector, e, contains the stacked error terms, $\omega_{111996}, \upsilon_{111996}, ... \omega_{IN1999}, \upsilon_{IN1999}$. The variance matrices for the person, firm, and shared error component random effects, respectively, can be written as

$$\begin{aligned}
G_1 &= I_{IxI} \otimes \sigma_\theta^2 \\
G_2 &= I_{JxJ} \otimes \sigma_\psi^2 \\
G_3 &= I_{NxN} \otimes \sigma_\eta^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \\
\text{where } \sigma_\eta^2 &= \frac{\sigma_\varsigma^2}{(1-\rho^2)}
\end{aligned}$$

while the variance matrix for the measurement errors can be written as

$$R = I_{(Nx4)x(Nx4)} \otimes \begin{bmatrix} \sigma_\omega^2 & 0 \\ 0 & \sigma_\upsilon^2 \end{bmatrix}$$

The shared error component is modeled as an AR(1) process where errors are correlated within the same job for a given individual but not across jobs and nor across individuals. This effect is identified by the fact that there are two observations for each time period. The measurement errors are modeled as being independently distributed with the covariance between the SIPP and SSA errors constrained to be 0. These errors are identified by differences in the SIPP and SSA earnings reports for each year.

The intercepts and the coefficients on experience and time are treated as fixed effects while the unobserved person, firm, and job/time period effects are treated as random. Estimates of β_{0SIPP}, β_{0DER}, β_1, β_2, the variance components (σ_θ^2, σ_ψ^2, σ_η^2, σ_υ^2, σ_ω^2), and realizations of the random effects (θ, ψ, η) and the residuals (υ, ω) can be obtained by solving the mixed model equations.

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\beta} \\ \widehat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix}$$

The estimation is done by restricted maximum likelihood (REML) using an average information (AI) algorithm, developed and programmed by Gilmore, Thompson, and Cullis (1995). This method closely follows the Fisher scoring algorithm proposed by Patterson and Thompson (1971). Parameters are chosen to maximize the log likelihood function by satisfying a set of first order conditions, or score equations. Solutions to the score equations are calculated iteratively. The user furnishes a set of starting values for the variance components and the algorithm calculates the log likelihood and produces initial estimates of the fixed effects ($\beta's$) and the realized random effects. The information matrix is calculated using an averaging method that simplifies the process for large data sets with multiple random effects. The information matrix is then used to update the variance component estimates. The process is repeated until the estimates converge.

3 Data

The fundamental unit of observation in this paper is a job, defined as a match between an individual and a firm. Data on jobs comes from two sources: the 1996 Survey of Income and Program Participation (SIPP) Panel and the Detailed Earnings Records (DER) extracted from the Social Security Administration Master Earnings File for the 1996 SIPP Panel respondents. In both sources, data on earnings were reported on a sequential, calendar basis and job records had to be created by combining earnings records over time that belonged to the same job. Hence appropriately grouping earnings records and defining jobs was the first fundamental difficulty that was addressed in each data source. After job records were created, individuals in each data set were linked by Social Security Number (SSN). Finally, jobs for each individual from the two data sources were matched to each other. Each step of this process is described below.

3.1 Creating a SIPP Jobs Data Set

The 1996 SIPP Panel interviewed respondents every 4 months beginning in December 1995 for the first rotation group and finishing in February 2000 for the last rotation group. There were 4 rotation groups and one group was interviewed each month. At the time of the interview, retrospective information about the previous 4 months was collected. Respondents were asked to report information for up to 2 jobs they held during this time period. The industry, occupation, union status, usual weekly hours, and monthly earnings of each job were recorded, as well as any applicable start and end dates. Information was collected using a CAPI (Computer Assisted Personal Interview) system and some responses were carried forward and were available to the interviewer at the time of the next interview. Employment information was one example of such a response. Each time an individual reported a new job, it was assigned a unique identifier, EENO, with the intent that this identifier be time-invariant and allow the linking of job information across survey waves.

The 1996 SIPP Panel contains 498,553 person, wave, job records. These records are indexed by the longitudinal SIPP person id (INTID), the wave number (WAVE), and the job id (EENO). These records were combined to create one observation per job which contained a time invariant industry code (INDJOB), a set of dummy variables that indicated which years the job was held (YEAR1995 - YEAR2000), and annual earnings for the survey period (ANNEARN1995-ANNEARN2000). There were 136,550 unique job observations after this process and 63,600 unique individuals. This step led to the observation of two problems. First, for about 10% of all jobs, the reported start date differed across waves. Table 1 gives an example of this type of problem.

In this example, the SIPP respondent held job 1 for waves 1 and 2, from January 1, 1996 to August 31,1996. The respondent then quit job 1 and began job 2 in wave 3 on September 1, 1996. This second startdate was accurately

Table 1: SIPP Startdate Problems: Type I

EENO	WAVE	Startdate	Enddate	INDJOB
1	1	Jan. 1, 1996		591
1	2	Jan. 1, 1996	Aug. 31, 1996	591
2	3	Sep. 1, 1996		612
2	4	Jan. 1, 1996		591

Table 2: SIPP Startdate Problems: Type II

EENO	WAVE	Startdate
1	1	Feb. 1, 1996
1	2	Feb. 1, 1996
1	3	Feb. 1, 1996
1	4	Feb. 1, 1996
	5	
2	6	Jan. 1, 1996

recorded along with a new industry code. However in wave 4, the second startdate and industry were erroneously replaced with the first job start date and industry. This mistake occurred due to an error in the data processing framework for some respondents. It was easily fixed by ignoring the start dates given in subsequent waves and always using the first start date given for a job. When made aware of this problem, the SIPP processing branch at the Census Bureau corrected their system and posted a warning notice to SIPP users on the SIPP homepage at the Census Bureau website[2]. The longitudinal 1996 SIPP Panel, which was still being processed at the time this error was discovered, has been corrected.

The second problem with job start dates was much more complicated and difficult to correct. This problem arose when jobs had start dates prior to the beginning of the first wave in which they were reported and prior to the beginning of the previously held job. Table 2 gives one example of the cause of this problem.

In this case, the individual was interviewed in waves 1 through 4 and reported a job which began February 1, 1996. However, the individual missed the fifth interview. When the next interview was conducted in wave 6, a new job was reported but the start date was prior to the beginning of wave 6 and prior to the beginning of job 1. The CAPI system was not designed to allow job ids to be carried forward through missed interviews and so when this person temporarily dropped out of the panel, she was automatically given a new job id at the time of the next interview, regardless of whether the job had actually begun in wave 6 or not. However, there were no restrictions placed on the start date she reported and hence this discrepency arose.

The case illustrated in Table 2 was the most common cause of the early

[2]See http://www.sipp.census.gov/sipp/core/1996/usernotes/1996_Cross_Section.htm

start date problem. However it was not the only cause. The problem affected 16.7%(10,635) of all SIPP respondents who held 22.7% of all jobs (31,085) and about 40% of the time there appears to have been a missing wave problem while the rest of the time, the cause is still unknown. Whatever the cause, it seems clear that in some cases the EENO job ids used in the survey data do not always correctly link jobs over time.

In order to produce a more accurate set of jobs for each individual, with the appropriate start and end dates, I turned to another source of identifying job information: the name of the employer as reported by the SIPP respondent. Out of 136,550 jobs, 108,297 had non-missing name information. Using stasticial name matching software called Vality, an unduplication procedure was performed that searched for matching job names within all the jobs for a given individual. The first step of this procedure was to standardize the names of the employers. The reported name was parsed into four new fields: compname, type, qualifier, and geo. Common words such as "Inc," "Company," or "Firm" were saved in the qualifier or type fields and geography words such as state names were saved in the geo field. The remaining words were saved in compname and this field was used as the primary set of unique words describing an employer. The second step was to perform statistical name matching using the four standardized fields. Each field used in the matching was given an m and u probability. The m probability was the probability that the same field on two separate records agreed given that two records were indeed a match. When this probability was set to less than one, it was assumed that there were some errors in the fields and that even if two records were a match, there was still a small chance that the field on the first record was miscoded and would be different than the same field on the second record. The u probability was the probability that the same field on two separate records agreed given that the records were not a match. This was the probability that a field agreed at random. Next a blocking field was chosen. Records with the same value in this field were assigned to the same block and only records in the same block were actually compared. In this application, the blocking value was the SIPP person identifier and hence only jobs held by the same person were compared.

Using the m and u probabilities, weights for each potential pair within a block were calculated. If the field was determined to agree then an agreement weight was assigned (positive value) and if the field was determined to disagree, a disagreement weight was assigned (negative value). The following formulas were used to calculate the two weights:

$$\begin{aligned}\text{agreement weight} &= \log_2(\frac{m}{u}) \\ \text{disagreement weight} &= -(\log_2(\frac{1-m}{1-u}))\end{aligned}$$

Once weights were assigned to each field, a composite weight was calculated for the pair by summing all the individual weights. Using the composite weights, record pairs were determined to be matches or not based on the cutoff values

Table 3: Statistical Name Matching Probabilities

field	m-prob	u-prob
names	.9	.3
qualifier	.9	.3
type	.9	.3
geo	.7	.5

specified by the user. The cutoff values gave the minimum weight required for a match and the minimum value required to be declared a potential match requiring clerical review. Any weights below the clerical review minimum were determined to be non-matches.

The choice of m and u probabilities and cutoff levels was determined both by knowledge about the fields and by experimentation. For the compname field, a high m probability and low u probability were chosen. Since compname was deemed to be the part of the employer name that was unique to that firm, matching values of compname were essential to matching records, thus requiring the high m probability. At the same time, compname was unlikely to agree at random and hence produce false matches, so a low u probability was chosen. The result of these choices was that matching values of compname received very high agreement weights and also very high disagreement weights. The type, qualifier and geo fields, on the other hand, had higher u probabilities. Agreement in one of these fields produced a lower agreement weight because matches were more likely to happen at random while disagreement produced a more negative disagreement weight because non-matches meant the companies were unlikely to be the same. Table 3 gives the m and u probabilities for each field used in the matching.

Cutoff values were chosen by examining certain and uncertain matches and determining the range of their weights. The values chosen for matches and clerical matches in this application were 2 and .3 respectively.

After the matching was completed, a new job id variable, EENO_2, was created and attached to each person, wave, job level record. Like the original job id variable, EENO_2 linked job observations over waves of the SIPP panel but it incorporated the new information about matching records provided by the name matching software. Using this new job id, records were again combined to create a job level file, now with 125, 282 unique jobs. The procedure of combining job records with different EENO numbers but similar names did not solve all the start date problems but it did make some improvements. Only 11.5% (7,292) of people holding 20.7% of jobs (25,996) now had at least one job with a start date problem.

One final problem arose in the processing of SIPP jobs. Respondents were only allowed to report at most two jobs per interview. In cases where people had a series of short or part-time jobs, interviewers recorded a single job which was labeled as "various employers" or "work arrangement." There were 3,926 job records of this type in the SIPP data, representing possibly triple that many actual jobs. These jobs were thought to be impossible to match to the DER because they did not represent earnings from a single employer. Hence they were dropped, giving a new total of 121,356 jobs.

3.2 Creating an SSA Jobs Data Set

The second source of data, Detailed Earnings Records (DER) from SSA, contained earnings histories for each SIPP respondent in the 1996 Panel with a validated SSN (for a definition and discussion of validation see section 3.3: "Matching SIPP and SSA Jobs"). These histories included reports of annual earnings, by employer, from 1978-2000. For the purposes of this earnings comparison study, however, only non-self-employment jobs held from 1995-2000 were used since this covered the time period of the survey questions[3]. Employers on this administrative data were identified by an IRS-assigned Employer Identification Number (EIN). There were 607,873 jobs held by SIPP respondents from 1978-1999, representing 315,471 unique EINs. Of these, 192,720 non-self-employment jobs representing 105,095 unique EINs were held from 1995-2000 and hence were potential matches to the jobs reported in the SIPP 1996 Panel.

The EIN linked employers to the Business Register, the master list of all businesses maintained by the Census Bureau to use in sampling firms for surveys. Using this link, I merged information from the Business Register about the industry and name of the employer to each relevant job report in the SSA data. This merge was somewhat complex because the Business Register had two parts. The first part was called the Single-unit file and contained records for all EINs that were either single-unit companies or sub-masters. Single-unit companies were firms with only one establishment that had a single EIN. Sub-masters were companies with multiple establishments that shared an EIN, i.e. multi-unit companies. For single-unit companies, the names and industries found on the Single-unit Business Register file were likely to correspond to the names and industries of employers reported in the SIPP. However for sub-masters, the name and industry were potentially quite different because these represented some aggregate concept - name of parent company or major industry out of a group of industries represented within a multi-unit company. Hence for sub-masters, I also searched for information about the EIN in the second part of the Business Register, the Multi-unit file. Here I obtained multiple records for each EIN representing the names and industries of all the different establishments associated with a sub-master record. For these multi-unit companies, I kept one record for each unique industry. Establishments within the same industry tended to have extremely similar names and hence this choice resulted in both a manageable number of observations to match to SIPP jobs while still providing additional information that might assist in the match.

[3]The Detailed Earnings Records did contain reports of self-employment earnings. These were coded with an EIN of 999999999. The SIPP also collected information about self-employment, but responses to these questions were treated separately from responses to the questions about jobs with employers. Self-employment reports from either source were not included in this study.

Table 4: Matching DER Data to the Business Register

DER Total		Match to		
		Business Register	Single-Unit File	Multi-Unit File
EINs	105,095	95,122 (90.5%)	94,438 (89.8%)	28,923 (27.5%)
Jobs	192,720	172,832 (89.7%)	171,585 (89.0%)	82,546 (42.8%)

Table 5: DER to Business Register Match Rates by Year

	Year Job first reported in Detailed Earnings Records						
	1995*	1996	1997	1998	1999	2000	Total
Total jobs	51,115	30,905	31,471	31,603	32,078	15,548	192,720
Non-matched Jobs	1,848	598	752	589	553	15,548	22,982
$\frac{\text{Non-matched Jobs in year}}{\text{Total Jobs in year}}$	3.6	1.9	2.4	1.8	1.7	100	10.3
$\frac{\text{Non-matched Jobs in year}}{\text{Total Non-matched Jobs}}$	9.3	3.0	3.8	3.0	2.8	78.2	100

*Jobs in this column either began in 1995 or were already in progress by 1995. Jobs in all other columns began in the year listed.

The Business Register is maintained on a yearly basis. Initially an EIN from a job was sought in the Business Register year that corresponds to the first year the job was reported in the DER. If a job was already in progress in 1995, the start year was coded as 1995 since this was the first year the job was at risk to match to the SIPP. If the job was not found in the Business Register year corresponding to the start year, it was sought in the following two Business Register years. The Business Register for 2000 was not yet available so jobs beginning in the DER in 2000 were not able to be matched. For the purposes of this study, this did not present a serious problem because so little SIPP data was collected in 2000 that annual earnings from jobs beginning in 2000 could not be compared. Table 4 presents a summary of the match rates between the DER and the Business Register.

There are several interesting things to notice in this table. First, although only 27.5% of all EINs represented multi-unit companies, these EINs accounted for 42.8% of all jobs. Second, there was also a small percentage of EINs and jobs that were found in the Multi-Unit file but not in the Single-Unit file. The cause of this is unknown at this time and will need further research. These aggregate match rates disguise the fact that the majority of non-matching EINs represented jobs that began in 2000. Table 5 breaks down the match rate by year the job was first reported in the Detailed Earnings Records.

3.3 Matching SIPP and SSA Jobs

Table 6 shows the total number of people and jobs that were potential matches following the job record creation process described in the previous two sections.

The next step in the linking process involved linking at the person level. The unique identifier for a person on the DER was the SSN while the SIPP

Table 6: Counts of People and Jobs in DER and SIPP

	DER	SIPP
people	55,894	63,116
jobs	192,720	121,356

Table 7: Matching Individuals in SIPP and DER

	people	DER jobs	SIPP jobs
match	48,542	173,623	97,081

contained a longitudinal person identifier specific to the survey. A crosswalk file matched SSNs and SIPP person ids. This crosswalk was developed using self-reported SSNs and a validation procedure. Each SIPP respondent was asked to provide an SSN. After this information was collected, SSA searched for each SSN in an administrative data base called the Numident, a universe file containing demographic information collected when every SSN was issued. Self-reported name, gender, race, and date of birth were compared to their administrative counterparts. If a respondent's name and demographics were deemed close enough to the name and demographics associated with the SSN in the administrative data base, then the SSN was declared valid. For respondents who answered "do not know" to the SSN question, an attempt was made to find the missing SSN by locating the person in the Numident based on their reported name and demographic characteristics. When a respondent refused to provide an SSN, no attempt was made to link this person to any administrative data and the SSN was left missing. Validated SSNs were included in the crosswalk file and served as the basis for extracting Detailed Earnings Records from the SSA Master Earnings File. The 1996 SIPP Panel crosswalk contained 92,033 validated SSNs, 8,657 invalid SSNs, and 15,363 refused SSNs.

The DER data set for this SIPP panel contained only 68,652 unique SSNs because not all people with validated SSNs held jobs between 1978 and 2000. Of those who were found in the DER, 55,894 held jobs between 1995 and 2000. Using the crosswalk, individuals in the SIPP and DER data sets were merged. Table 7 shows the results of this merge.

Table 8 describes the people from both the SIPP and DER who do not match.

There were people who reported jobs in the SIPP but for whom no jobs were found in the administrative records. Most of this difficulty was caused by lack

Table 8: Individuals in SIPP and DER with Validated SSNs by Matching Status

	DER	SIPP	Both
people with validated SSNs	55,894	50,691	48,542
people with jobs in DER not SIPP	7,352		
people with jobs in SIPP not DER		2,149	

of validated SSNs. Of the 63,116 respondents who reported jobs, only 50,691 had validated SSNs. This left only 2,149 respondents who had a validated SSN and reported a job in the SIPP but who were not found in the administrative records. These jobs are likely to be uncovered employment or erroreously reported self-employment, i.e. babysitting, yard work, etc[4]. Potentially more troubling was the existence of 7,352 SIPP respondents with validated SSNs who had administrative job records between 1995 and 2000 but did not report any jobs in the SIPP[5]. While some jobs were missing on both sides, clearly the administrative data picked up more employed people than the survey.

As shown in Table 7, even for those SIPP respondents who had employment reports in both the SIPP and the DER, the number of jobs reported was much higher in the administrative data compared to the survey data. At least one factor that influenced the job count on each side was the timing of the survey. For the first rotation group, the survey began collecting data in 1995 and for the last two rotation groups, the survey continued to collect data until January and February of 2000. For rotation groups who had data collected in either 1995 or 2000, I included all jobs found in the SSA data which had earnings reports for these years. However some jobs in the administrative data would have ended before the survey began in 1995 or started after the survey ended in 2000 and thus there were "extra" jobs for some people.

After the match by SSN had been performed, a job-to-job match was performed again using the statistical name matching software Vality. On one side of the match were all the SIPP jobs deemed to be reports of employment at a single employer, a total of 97,081 records. On the other side of the match were all the records associated with the DER jobs deemed to have taken place during the at risk time frame. Each DER record contained the name and industry of the EIN as found on the Single-unit part of the Business Register. When the EIN was also found on the Multi-unit part of the Business Register, the record contained a second name and industry representing information about a particular establishment of this EIN. When an EIN was associated with multiple establishments with different industries on the Mulit-unit file, multiple records were created for this DER job. Each record contained the same Single-unit name and industry information but different Multi-unit name and industry information. There were a total of 196,845 DER records representing the 173,623 jobs described in the previous section.

The matching was performed in several steps, called passes. The goal was to first link jobs that were almost certain matches based on the fields deemed to be the most reliable matching indicators and then to link jobs that were less

[4]Of the 2,149 working SIPP respondents who do not have a regular job in the DER, 610 have a self-employment record in the DER. However since self-employment cannot be matched to the Business Register, there is not enough information to match the SIPP job to the DER self-employment job.

[5]There were 217 people with validated SSNs who had at least one earnings report in the DER but only one job reported in the SIPP as "various employers" or "work arrangement." Since SIPP jobs of this type were previously excluded, these people were dropped from the analysis. This slightly exaggerates the person-level non-match rate between the DER and SIPP.

certain matches using other fields. Table 9 gives the blocking and matching fields for each pass and along with the accompanying m and u probabilities for the matching fields. Earnings were not used in the match in order to prevent bias in the subsequent comparison of earnings. Some special steps were taken to deal with missing values for names. In passes where the name information was used in matching, the complete name of the firm was included as a matching field. However, this field was not given m and u probabilities. Instead it was used to prevent jobs with missing names from matching. If either the SIPP or DER complete names were missing, the weights were set such that the records automatically did not match in this pass. If the names were not missing then this field automatically received a weight of zero and did not contribute to matching decision.

Several variables were also used in multiple passes, with the requirements for matching gradually relaxed. For example, in the third pass, three-digit SU industry was used as a blocking variable and the four year indicators were used as matching variables. Pass five was quite similar except that instead of requiring records to match on all four year indicators, only start year was required to match. Start year was a field that indicated the first year that a record was found for this job. The first possible year was 1995 and the last possible year was 2000. Likewise in pass seven, only one-digit industry was used as a blocking variable. This process enabled the detection of high-probability matches in early passes and then the addition of lower-probability matches in later passes.

Table 10 shows the results of the matching. Of the SIPP jobs, 75,127 (77.4%) were matched to a corresponding SSA job. Of these matches, 66,387 (68.4%) were deemed high probability matches that surpassed the clerical editing threshhold, while 8,740 (9.0%) were clerical matches. The majority of the matching took place in the first pass (76.3% of Master Pairs and 39.3% of Clerical Pairs). The next most successful passes were 3 (11.1% MPs and 12.6% CPs) and 5 (.3% MPs and 19.6% CPs).

The third row in Table 10 highlights two problems that result from the matching. First, two different SIPP jobs could match to the same DER Job as illustrated in Table 11.

There were several possible causes of this problem. First, it was possible that the two SIPP jobs were indeed the same and the SIPP job creation phase erroneously failed to link them. In this case the duplicate record was a "true" duplicate and both jobs were correctly matched to one DER job. However another possibility was that the matching software mistakenly matched a second SIPP job to the same DER job due to lack of differentiating information for the SIPP jobs. This was particularly likely in the later passes where matches were based on year and industry indicators alone. In this case, the duplicate was false and only one of the two matches was correct. Initial editing of the data indicated that about 50% of the duplicate SIPP jobs matched in passes 1 and 2 were "true" duplicates (495 out of 991 jobs) and that some improvements in the SIPP job creation phase could be made. However given that the number of jobs affected was relatively small, I chose to drop duplicate pairs (DA) and

Table 9: Description of Matching Algorithm by Pass

	Blocking Variables	Matching Variables	m, u prob.
Pass 1	Person ID	Fields from SU Name:	
		Array: first 4 words	.95, .1
		Array: first 2 qual. words	.9, .3
		Array: first 2 type words	.9, .3
		Geo word	.7, .5
		year1996 - year 1999	.75, .3
		Complete SU name	
Pass 2	Person ID	Fields from MU Name:	
		Array: first 4 words	.95, .1
		Array: first 2 qual. words	.9, .3
		Array: first 2 type words	.9, .3
		Geo word	.7, .5
		year1996 - year 1999	.75, .3
		Complete MU name	
Pass 3	Person ID	year1996 - year1999	.9, .3
	3-digit SU Ind.		
Pass 4	Person ID	year1996 - year1999	.9, .3
	3-digit MU Ind.		
Pass 5	Person ID	start year	.9, .3
	3-digit SU Ind.		
Pass 6	Person ID	year1996 - year1999	.9, .3
	1-digit SU Ind.		
Pass 7	Person ID	year1996 - year1999	.9, .1
		3-digit SU Ind.	.9, .1

Table 10: Match rates of SIPP and DER Jobs

	SIPP Jobs	Percent	Der Jobs	Percent
Master Pair (MP)	66,387	68.4	66,387	38.1
Clerical Pair (CP)	8,740	9.0	8,740	5.0
Duplicate (DA, DB)	1,856	1.9	7,418	4.3
Residual (RA, RB)	20,098	20.7	91,078	52.6
Total	97,081	100	173,623	100

Table 11: Duplicate SIPP Jobs Match One DER Job

	DER	SIPP
type	EIN	Jobnum
MP	A	1
DA	A	2

Table 12: Duplicate DER Jobs match One SIPP Job

	DER	SIPP
type	EIN	Jobnum
MP	A	1
DB	B	1

Table 13: Determining Duplicate DER Jobs to be the Same Job

			DER Years				SIPP Years			
SIPP Job	EIN	Type	1996	1997	1998	1999	1996	1997	1998	1999
Jobnum1	EINA	MP	1	1	0	0	1	1	1	1
Jobnum1	EINB	DB	0	0	1	1	1	1	1	1

kept only master and clerical pairs (MP and CP), pending further investigation into SIPP job linking. There was only one exception to this rule. In a very few cases, the matching software declared both SIPP jobs to be Master or Duplicate Pairs. In this case, I dropped both SIPP jobs associated with the EIN (122 total records) since I was unable to determine which was the better match.

The second problem shown in Row 3 of Table 10 is the reverse duplication issue. Two different DER jobs sometimes matched to the same SIPP job as shown in Table 12.

This type of duplication was more common and it was more difficult to know the causes. The first possibility was that a company changed its EIN due to a change in ownership structure or some other reason. If this EIN change did not represent a change in the employment of the workers, the SIPP respondent might have reported holding the same job while the administrative data for the individual contained a new earnings report from the second EIN. Another possibility was that SIPP respondents reported "lump" jobs, meaning that one SIPP job was really a combination of several jobs. Since administrative records pertained to the source of the earnings, it was possible that some individuals considered themselves as holding only one job but were in reality paid from several different source EINs. It was also possible that individuals consciously grouped jobs in order to ease the burden of responding to the survey. These issues warrant further research.

For the purposes of this study, I combined duplicate DER jobs that had a strong indication of representing the same company. I judged this by first summing the year indicator variables attached to DER jobs labeled as master and duplicate pairs by Vality. If the pattern of the summed year variables matched the exact pattern of SIPP year variables, I then concluded that the DER jobs combined to form one job that was equivalent to the SIPP job and I summed earnings from the two DER jobs for each year. Table 13 gives an example of a hypothetical case where the year pattern of the DER jobs sums to equal the year pattern of the SIPP job. This exercise combined duplicate DER records for 1,275 SIPP jobs.

Table 14: Summary Statistics for Matched Jobs by Year

	1996	1997	1998	1999	Total
N SIPP	42,476	41,145	37,646	34,932	156,199
N DER	44,003	44,311	42,451	39,617	170,382
log (SIPP real ann. earn.)	9.09	9.14	9.21	9.27	9.17
	(2.09)	(2.07)	(2.02)	(1.90)	(2.03)
log (SSA real ann. earn.)	9.11	9.14	9.19	9.25	9.17
	(2.49)	(2.51)	(2.54)	(2.48)	(2.51)
years of experience	18.10	18.34	18.59	19.12	18.53
	(152.18)	(152.18)	(152.76)	(154.37)	(152.98)

Table 15: Covariance/Variance Matrix of SIPP Job Annual Earnings

Log SIPP Job Annual Earnings	1996	1997	1998	1999
1996	2.09	.72	.66	.64
1997	1.15	2.07	.72	.66
1998	.84	1.09	2.01	.71
1999	.71	.81	1.04	1.90

After dropping duplicates and residuals as described above, as well as job matches where either the SIPP or DER job did not have any earnings between 1996 and 1999, the resulting data set to be used in the analysis contained 74,059 jobs at 47,601 unique employers held by 44,388 unique individuals.[6] Table 14 gives summary statistics for earnings and experience for all jobs from 1996 to 1999. Numbers in parentheses are variances of the preceding row. As is clear from the table, there were some jobs which matched but did not have the same number of years of reported earnings. For example a SIPP job could have earnings reports for 1996 and 1997 but not 1998 while the SSA job could have reports for all three years. This resulted is slightly different sample sizes between the SIPP and the SSA data for each year. Missing values were modeled in the maximization routine as conditionally missing at random and hence the panel was not required to be balanced.

Tables 15 and16 describe the variance/covariance structure of the SIPP and DER earnings over time. The covariances are listed below the diagonal and correlations are listed above.

Table 17 gives the correlations between each year of DER and SIPP data.

[6] Of the 75,127 jobs which were either master or clerical pairs, 122 were dropped because they were associated with a case of one EIN matching two different SIPP Job IDs and both matches being declared master or clerical. In addition, 816 jobs were dropped because there were no DER earnings between 1996 and 1999 and 130 were dropped because there were no SIPP earnings during this time period.

Table 16: Covariance/Variance Matrix of DER Job Annual Earnings

Log DER Job Annual Earnings	1996	1997	1998	1999
1996	2.49	.77	.73	.69
1997	1.47	2.51	.77	.72
1998	1.20	1.43	2.54	.78
1999	1.03	1.15	1.45	2.48

Table 17: Correlation Matrix of SIPP and DER Job Annual Earnings

Log DER Annual Earnings	Log SIPP Annual Earnings			
	1996	1997	1998	1999
1996	.85	.68	.64	.61
1997	.70	.84	.68	.63
1998	.66	.68	.84	.68
1999	.62	.63	.67	.83

4 Results

Table 18 presents results from estimating equations 1 and 2 with the previously described variance structure. The first half of the table reports estimates of the variance components and the intercepts and linear time trend fixed effects. General labor force experience was modeled as a linear spline with nodes at 2, 5, 10, and 25 years and interacted with gender and race to give separate experience profiles for four different groups: white males, non-white males, white females, and non-white females. The experience measure used was calculated from survey responses to questions about the year entering the labor force and time taken off work. These coefficients are presented in Table 19 and experience profiles for each group are graphed in Figure 1.

The variation explained by the person and firm effects is approximately equal. The variance of the SIPP and DER measurement error terms is similar although the DER variance is larger. This result is probably related to the fact that the overall variance of the DER earnings is larger than the variance of the

Table 18: Estimation Results: Specification 1

Variance Components		Intercepts and Time	
σ^2_θ	.3694		
σ^2_ψ	.3704	β_{0SIPP}	6.6604
σ^2_η	.7505	β_{0DER}	6.6981
ρ	.6978	time trend β_2	-.00085
σ^2_ω	.3249		
σ^2_υ	.3859		

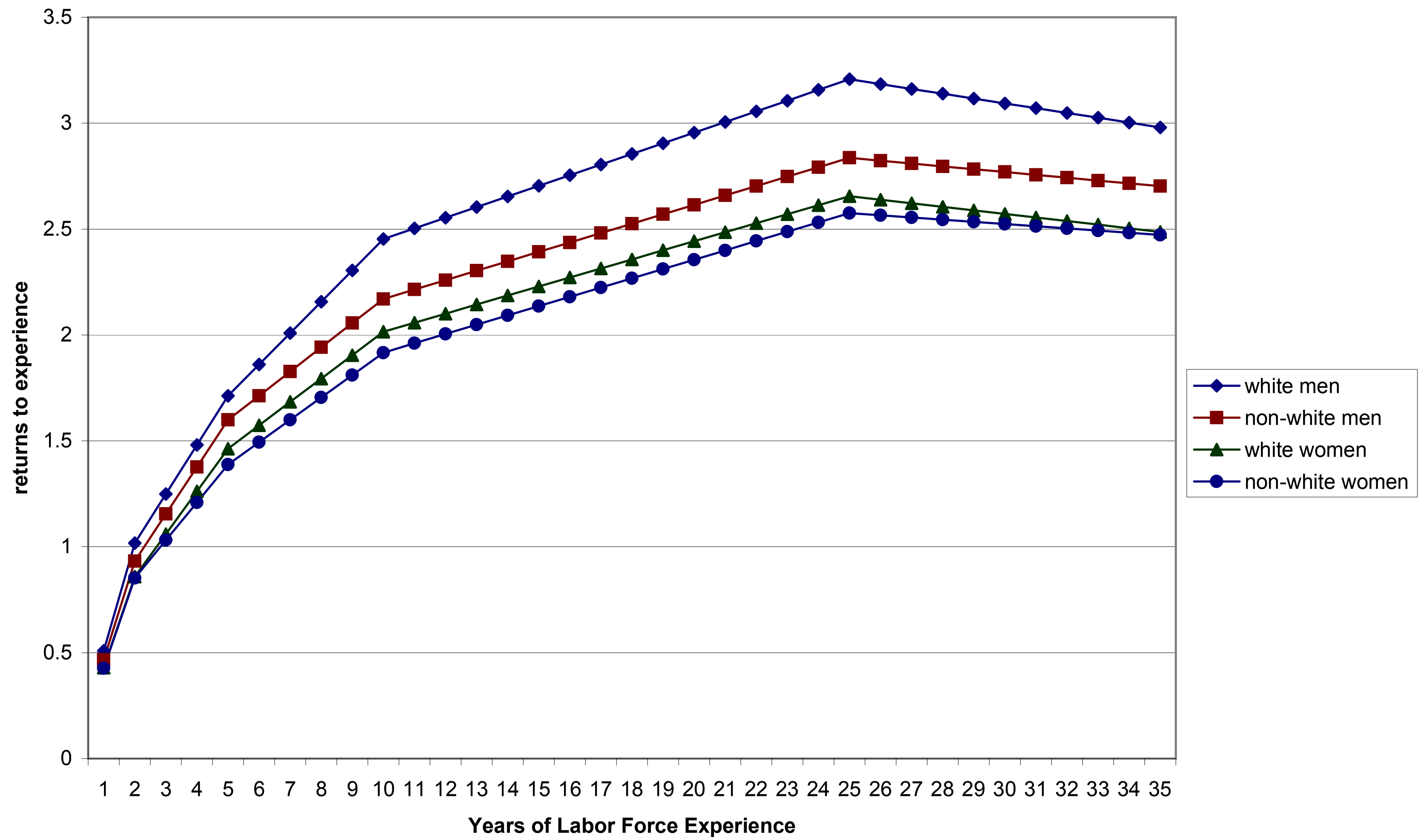

Figure 3.1: Labor Force Experience Profiles

Table 19: Estimation Results: Specification 1, Experience Effects

Years of Experience				
	white males	non-white males	white females	non-white females
0-2 years	.5087	.4663	.4292	.4260
2-5 years	.2316	.2221	.2015	.1789
5-10 years	.1481	.1143	.1105	.1056
10-25 years	.0503	.0444	.0427	.0439
25+ years	-.0227	-.0134	-.0167	-.0103

Table 20: SIPP Residuals Variance/Covariance Matrix: Specification 1

	ω_{1996}	ω_{1997}	ω_{1998}	ω_{1999}
ω_{1996}	.313	.39	.43	..47
ω_{1997}	.125	.338	.38	.43
ω_{1998}	.139	.128	.327	.42
ω_{1999}	.149	.143	.135	.322

SIPP earnings. The experience splines have the expected concave slope with white males having the steepest slope.

The traditional reliability ratio can be calculated using the ratio of the sum of the "economic" variance components (η, θ, ψ) to the sum of the economic and measurement error variance components. Thus for the SIPP and SSA earnings measures, respectively, the reliability ratios are calculated as

$$\begin{aligned} \kappa_{SIPP} &= \frac{\sigma_\eta^2 + \sigma_\theta^2 + \sigma_\psi^2}{\sigma_\eta^2 + \sigma_\theta^2 + \sigma_\psi^2 + \sigma_\omega^2} = .8210 \\ \kappa_{SSA} &= \frac{\sigma_\eta^2 + \sigma_\theta^2 + \sigma_\psi^2}{\sigma_\eta^2 + \sigma_\theta^2 + \sigma_\psi^2 + \sigma_\upsilon^2} = .7943 \end{aligned}$$

These ratios are similar to those obtained by Bound et al. in the PSID validation study (.7 in 1986 and .85 in 1982) and Bound and Krueger in the CPS validation study (.84 for 1976 and .82 for 1977). However our ratios were obtained in a much different manner. Our model estimates both the true and the measurement error variation, using the repeated earnings measures in a given time period to identify the model.

Tables 20 and 21 report the covariance/variance matrices of the SIPP and SSA measurement errors, ω and υ, for the four years in the panel, 1996 to 1999. Table 22 reports the covariance/variance matrix of the common measurement error, η. As before, covariances are listed below the diagonal and correlations are listed above.

Measurement error seems to persist over time in the SIPP and SSA data and the common error component also seems to be correlated across years, although this correlation declines over time.

Table 21: SSA Residuals Variance/Covariance Matrix: Specification 1

	υ_{1996}	υ_{1997}	υ_{1998}	υ_{1999}
υ_{1996}	.376	.41	.43	.44
υ_{1997}	.156	.387	.40	.43
υ_{1998}	.163	.154	.386	.43
υ_{1999}	.170	.167	.168	.397

Table 22: Common Error Component Var./Covar. Matrix: Specification 1

	η_{1996}	η_{1997}	η_{1998}	η_{1999}
η_{1996}	.710	.79	.68	.62
η_{1997}	.573	.741	.79	.69
η_{1998}	.496	.586	.745	.80
η_{1999}	.448	.508	.598	.738

It is of interest to compare the results above to results from an econometric specification with no person and firm random effects. These results are reported in Tables 23-27.

$$\kappa_{SIPP} = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\omega^2} = .8689$$

$$\kappa_{SSA} = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\upsilon^2} = .8463$$

The variance of the common error component is significantly larger than in Table 18 as is the auto-correlation coefficient, while the measurement error variances are lower. These results combine to give higher reliability ratios. The variance/covariance matrices of the residuals, again with correlations in the upper triangle, are reported in Tables 25-27.

The correlation patterns are significantly stronger for the common residual when the person and firm random effects are left out, as would be expected. The cross-year correlations of the measurement errors are also slightly stronger, but this difference is less pronounced.

A final specification of interest is one where the variance matrix of the errors is defined to allow correlation over time in the measurement error residuals.

Table 23: Estimation Results: Specification 2

Variance Components		Intercepts and Time	
σ_η^2	1.5423	β_{0SIPP}	6.610
ρ	.8598	β_{0DER}	6.648
σ_ω^2	.2327	time trend β_2	-.00706
σ_υ^2	.2801		

Table 24: Estimation Results: Specification 2, Experience Effects

Years of Experience				
	white males	non-white males	white females	non-white females
0-2 years	.4849	.4438	.3932	.3917
2-5 years	.2342	.2384	.2028	.1882
5-10 years	.1537	.1181	.1173	.1030
10-25 years	.0561	.0501	.0472	.0518
25+ years	-.0209	-.0097	-.0164	-.0085

Table 25: SIPP Residuals Variance/Covariance Matrix: Specification 2

	ω_{1996}	ω_{1997}	ω_{1998}	ω_{1999}
ω_{1996}	.311	.37	.43	.48
ω_{1997}	.120	.338	.37	.43
ω_{1998}	.137	.122	.327	.40
ω_{1999}	.152	.142	.130	.319

Table 26: SSA Residuals Variance/Covariance Matrix: Specification 2

	υ_{1996}	υ_{1997}	υ_{1998}	υ_{1999}
υ_{1996}	.380	.40	.43	.45
υ_{1997}	.154	.391	.38	.43
υ_{1998}	.164	.150	.389	.42
υ_{1999}	.177	.168	.165	.400

Table 27: Common Error Component Var./Covar. Matrix: Specification 2

	η_{1996}	η_{1997}	η_{1998}	η_{1999}
η_{1996}	1.463	.87	.82	.80
η_{1997}	1.284	1.492	.87	.83
η_{1998}	1.222	1.305	1.503	.89
η_{1999}	1.181	1.240	1.327	1.485

Table 28: Estimation Results: Specification 3

Variance Components							
Person, Firm effects		Common Error		SIPP Error		SSA Error	
σ^2_θ	.3600	σ^2_η	.7785	σ^2_ω	.2248	σ^2_υ	.5519
σ^2_ψ	.3547	ρ	.5913	ρ_ω	.1362	ρ_υ	.7326

Essentially the R matrix previously described is split into two pieces, R_ω and R_υ, that separately describe the variance of the SIPP and DER residuals.

$$
\begin{aligned}
R_\omega &= I_{NxN} \otimes \sigma^2_\omega \begin{bmatrix} 1 & \rho_\omega & \rho^2_\omega & \rho^3_\omega \\ \rho_\omega & 1 & \rho_\omega & \rho^2_\omega \\ \rho^2_\omega & \rho_\omega & 1 & \rho_\omega \\ \rho^3_\omega & \rho^2_\omega & \rho_\omega & 1 \end{bmatrix} \\
R_\upsilon &= I_{NxN} \otimes \sigma^2_\upsilon \begin{bmatrix} 1 & \rho_\upsilon & \rho^2_\upsilon & \rho^3_\upsilon \\ \rho_\upsilon & 1 & \rho_\upsilon & \rho^2_\upsilon \\ \rho^2_\upsilon & \rho_\upsilon & 1 & \rho_\upsilon \\ \rho^3_\upsilon & \rho^2_\upsilon & \rho_\upsilon & 1 \end{bmatrix}
\end{aligned}
$$

Variance components are presented in Table 28. The fixed effects were similar to the previous specifications. The estimated random person and firm effects are also similar to the estimates in 18. However the estimated auto-regressive structures on the two measurement error components are quite different from each other. The SSA error is much more highly auto-correlated than either the SIPP error or the common error component.

The reliability ratios for first-differenced variables are

$$
\begin{aligned}
\kappa_{\triangle SIPP} &= \frac{\sigma^2_\eta}{\sigma^2_\eta + \sigma^2_\omega(\frac{(1-\rho_\omega)}{(1-\rho)})} = .6210 \\
\kappa_{\triangle SSA} &= \frac{\sigma^2_\eta}{\sigma^2_\eta + \sigma^2_\upsilon(\frac{(1-\rho_\upsilon)}{(1-\rho)})} = .6831
\end{aligned}
$$

Notably, the reliability ratio for first-differenced SSA earnings is higher than for first-differenced SIPP earnings. This is due to the fact that the auto-correlation in the SSA errors is substantially higher than the auto-correlation in the other error components. High auto-correlation in the SSA errors means that some of the error gets differenced away because it is non-transitory. At the same time, some of the signal in η gets differenced away as well, but since the auto-correlation in η is not as high, less of the signal is lost in the first differencing. The opposite is true for the SIPP and hence the SSA reliability ratio is the larger of the two.

Finally, the variance/covariance matrices of the residuals are described in the Tables 29-31. Covariances are again in the lower triangle and correlations in the upper triangle.

Table 29: SIPP Residuals Variance/Covariance Matrix: Specification 3

	ω_{1996}	ω_{1997}	ω_{1998}	ω_{1999}
ω_{1996}	.218	.41	.47	.54
ω_{1997}	.091	.231	.41	.48
ω_{1998}	.105	.094	.225	.44
ω_{1999}	119	.108	.099	.223

Table 30: SSA Residuals Variance/Covariance Matrix: Specification 3

	υ_{1996}	υ_{1997}	υ_{1998}	υ_{1999}
υ_{1996}	.496	.69	.61	.57
υ_{1997}	.344	.499	.71	.64
υ_{1998}	.307	.361	.508	.74
υ_{1999}	.290	.326	.381	.524

5 Conclusion

The levels of measurement error calculated in this paper give some cause for optimism. First, for jobs found in both the SIPP and the SSA DER records, the earnings reports from the two sources are highly correlated. Second, measurement error accounted for only 18% of the variation in SIPP annual earnings and 21% of the variation in DER annual earnings. However errors in both types of data are correlated over time and in the case of the SIPP, this correlation makes the attenuation bias resulting from measurement error worse. Future research will focus on improving the job linking, specification checks, and investigating the role played by outliers. This study could also be expanded to incorporate multiple measures of labor market experience drawn from SIPP survey responses and counts of the number of years with administrative earnings records. Both measures of experience would be viewed as containing errors and the interaction of measurement error among multiple variables could be studied.

References

[1] Abowd, John M. and David Card. 1989. "On the Covariance Structure of Earnings and Hours Changes." *Econometrica* 57(2):411-446.

Table 31: Common Error Component Var./Covar. Matrix: Specification 3

	η_{1996}	η_{1997}	η_{1998}	η_{1999}
η_{1996}	.728	.63	.55	.51
η_{1997}	.474	.777	.62	.55
η_{1998}	.414	.483	.777	.65
η_{1999}	.379	.424	.498	.764

[2] Angrist, Joshua D. and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics Vol. 3A*, ed. Orley Ashenfelter and David Card, 1277-1357. Amsterdam: Elsevier.

[3] Bound, John and Alan B. Krueger. 1991. "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?." *Journal of Labor Economics* 9(1):1-24.

[4] Bound, John, Charles Brown, Greg J. Duncan, Willard L. Rodgers. 1994. "Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data." *Journal of Labor Economics* 12(3):345-368.

[5] Bound, John, Charles Brown, Nancy Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics Vol. 5*, ed. James J. Heckman and Edward Leamer, 3705-3843. Amsterdam: Elsevier.

[6] Duncan, Greg J. and Daniel H. Hill. 1985. "An Investigation of the Extent and Consequences of Measurement Error in Labor-economic Survey Data." *Journal of Labor Economics* 3(4):508-532.

[7] Fuller, Wayne. 1987. *Measurement Error Models*. New York: John Wiley and Sons, Inc.

[8] Gilmour, Arthur R., Robin Thompson, Brian R. Cullis, "Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models," *Biometrics* 51(4):1440-1450.

[9] Mellow, Wesley and Hal Sider. 1983. "Accuracy of Response in Labor Market Surveys: Evidence and Implications." *Journal of Labor Economics*, 1(4):331-344.

[10] Patterson, H.D. and R. Thompson. 1971. "Recovery of Interblock Information when Block Sizes are Unequal." *Biometrika*. 58(3):545-554.