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

John M. Abowd and Martha H. Stinson†

September 2007

Abstract

We quantify sources of variation in annual job earnings data collected by the Survey of Income and Program Participation (SIPP) to determine how much of the variation is the result of measurement error. Jobs reported in the SIPP are linked to jobs reported in a new administrative database, the Detailed Earnings Records (DER) drawn from the Social Security Administration's Master Earnings File, a universe file of all earnings reported on W-2 tax forms. As a result of the match, each job potentially has two earnings observations per year: survey and administrative. Unlike previous validation studies, both of these earnings measures are viewed as noisy measures of some underlying true amount of annual earnings. While the existence of survey error resulting from respondent mistakes or misinterpretation is widely accepted, the idea that administrative data is also error-prone is new. Possible sources of employer reporting error, employee under-reporting of compensation such as tips, and general differences between how earnings may be reported on tax forms and in surveys, necessitates the discarding of the assumption that administrative data is a "true" measure of the quantity that the survey was designed to collect. In addition, errors in matching SIPP and DER jobs, a necessary task in any use of administrative data, also contribute to measurement error in both earnings variables. We begin by comparing SIPP and DER earnings for different demographic and education groups of SIPP respondents. We also calculate different measures of changes in earnings for welfare respondents moving on and off public assistance and changes in earnings for individuals switching jobs. We estimate a standard earnings equation model using SIPP and DER earnings and compare the resulting coefficients. Finally exploiting the presence of individuals with multiple jobs and shared employers over time, we estimate an econometric model that includes random person and firm effects, a common error component shared by SIPP and DER earnings, and two independent error components that represent the variation unique to each earnings measure. We compare the variance components from this model and consider how the DER and SIPP differ across unobservable components.

1 Introduction

This paper makes use of a linked survey-administrative database to compare two different measures of earnings and consider the causes of the differences between them. We link job-level earnings reports from the Survey of Income and Program Participation (SIPP) to W-2 records extracted from the Social Security Administration's Master Earnings File. This extract is called the Detailed Earnings Records (DER) file. We use SSN and name of the employer to match SIPP reported jobs to the DER. These matched records provide a unique opportunity to assess differences in employee and employer reports and to consider the impact of these differences on correlations of earnings and other variables of interest in the SIPP survey.

The majority of past studies which compare two data sources, choose one of the sources to be the truth[1]. In contrast, we begin wtih an agnostic view about which data source is true, believing instead that there are legitimate differences between survey and administrative reports. Our goal is to first state the differences between the two data sources and then consider the reasons for these differences. We believe there are at least three reasons why administrative and SIPP survey reports on earnings might differ. First, there may be mis-matches between the records from the two sources. Our matching methodology uses a probablistic record link which produces some false matches. Also while identifiers for individuals remain constant over time in administrative databases, identifiers for firms do not and this complicates the matching process. Second, the

*We would like to thank Joanne Pascale, Tracy Mattingly, Gary Benedetto, Julia Lane, George Jakubson, Donald Kenkel, Simon Woodcock, Kevin McKinney, Kristin Sandusky, Lars Vilhuber, and Marc Roemer for helpful comments and support. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau, Cornell University, or any of the project sponsors. This work was 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. All of the data used in this paper are confidential data. The U.S. Census Bureau supports external researchers' use of some of these data items; please visit www.sipp.census.gov/sipp/ and click on "Access SIPP Synthetic Data."

†John M. Abowd is Professor of Industrial and Labor Relations at Cornell University, Distinguished Senior Research Fellow at the U.S. Census Bureau, Research Associate NBER, Research Affiliate CREST/INSEE, and Research Fellow IZA. Martha Stinson is an economist at the U.S. Census Bureau. contacts: john_abowd@cornell.edu, martha.stinson@census.gov

[1] For example, Duncan and Hill (1985) and Bound et al. (1994) compare employee and employer earnings reports from a large manufacturing company and treat the employer reports as truth. Bound and Krueger (1991) compare CPS earnings reports to SSA records and treat the SSA data as truth.

administrative records we utilize do not contain some types of earnings such as health insurance premiums that are usually reported on a pay stub and would probably be reported by a survey respondent. The reverse is also true: there are types of earnings which would appear on a W-2 form and would not appear on a pay stub, increasing the likelihood that these earnings would not be reported in a survey. Finally there are differences in reporting that can be labeled mistakes or measurement error. We believe these types of differences arise in both data sources. Employers make mistakes on W-2 forms just as employees make mistakes when they report earnings to a survey collector. The Social Security Administration does make corrections to the Master Earnings File when revised W-2 forms are submitted so over time these mistakes become less prevelant. One might hypothesize that errors in an administrative database are in general less prevelant than in a survey but they still undoubtedly exist.

Our strategy for investigating the differences between the SIPP and DER is to first focus on average differences for demographic sub-groups of interest. If differences between the two data sources vary across sub-groups, this may be evidence of mistakes made more frequently by some types of survey respondents. On the other hand it could also be that some sub-groups have more complicated types of earnings which will give rise to more definitional types of discprepancies between the two data sources. After these initial comparisons, we next focus on differences between the sources that are due to unobservable factors. Within each cell defined by a set of stratifying variables, there are average SIPP and DER earnings and then there is variation due to unobservable person, firm, and time period characteristics. We consider how much of this variation is common between the SIPP and the DER and how much might be unique to one data source or the other.

Our paper is organized as follows. We begin by describing the two data sources and our process for matching them. We discuss problems with this process and describe our final set of person-employer matches. We then explain the details of our analytic comparisons and our model for estimating the variance due to unobservable factors. We follow with results and discussion of potential causes of the differences that we observe. We end by trying to interpret our results in a measurement error framework, where we define measurement error as failure on the part of the survey to collect the intended quantity. We provide a range of estimates for the reliability of the SIPP data, making different assumptions about the sources of differences between SIPP and DER data.

2 Data Description

The fundamental unit of observation in this paper is a job, defined as a match between an individual and a firm. Data on jobs come from two sources: five Survey of Income and Program Participation (SIPP) Panels conducted during the 1990's and the Detailed Earnings Records (DER) extracted from the Social Security Administration Master Earnings File for the respondents in each of the five panels. In both sources, data on earnings were reported on a sequential, calendar-year 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 we addressed in each data source. After job records were created, individuals in each data set were linked by Social Security Number. Finally, jobs for each individual from the two data sources were matched to each other. Each step of this process is described below.

2.1 Creating a SIPP Jobs Data Set

All the SIPP Panels conducted in the 1990s interviewed respondents every 4 months over the course of 2 1/2 to 4 years. Interviewees were divided into 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[2]. Respondents were asked to report detailed information for up to 2 jobs they held during this time period. The employer name, industry, occupation, union status, usual weekly hours, and monthly earnings of each job were recorded, as well as any applicable start and end dates. Each job was also assigned a unique identification number, or job ID, with the intent that this identifier be time-invariant and allow the linking

[2] The SIPP panels generally began by interviewing the second rotation group in February of the panel reference year and collecting information about the previous 4 months. Interviews with the third, fourth, and first rotation groups followed and the reference period for collecting information moved forward each time so as to correspond to the preceding four months. For example, in the 1990 panel, the first interview cycle went as follows:

Second rotation group: interviewed in Feb.1990 and answered questions about Oct.1989-Jan.1990.
Third rotation group: interviewed in Mar. 1990 and answered questions about Nov.1989-Feb.1990
Fourth rotation group: interviewed in Apr.1990 and answered questions about Dec.1989-Mar.1990
First rotation group: interviewed in May 1990 and answered questions about Jan.1990-Apr.1990

This cycle continued until Sept. 1992 when the 8th interview with the first rotation group was completed, giving 8 complete waves of data for each rotation group. Thus the 1990 panel contains information on Oct.1989-Aug.1992. The following list gives the number of waves for each SIPP panel and the span of months referenced by the interview questions.

1990 panel:	8 waves, Oct.1989-Aug.1992
1991 panel:	8 waves, Oct.1990-Aug.1993
1992 panel:	10 waves, Oct.1991-Apr.1995
1993 panel:	9 waves, Oct.1992-Dec.1995
1996 panel:	12 waves, Dec.1995-Feb.2000

The cycle of interviews is slightly different for the 1996 panel. The first rotation group was interviewed in April 1996 followed by the second, third, and fourth rotation groups in May, June, and July. Thus the span of referenced months begins in December instead of October and ends in February instead of April, August, or December.

of job information across survey waves. For the first four panels (1990-1993), a paper and pencil survey instrument was used. During each interview, the Census field representative (FR) was required to record an employer name and assign a job ID for each job reported by the respondent, even if the job was a continuation of a job reported in a previous wave. While the FR was supposed to assign the same job ID to a continuing job and a new job ID to a newly begun job, there was no quality check to ensure that this procedure was followed. Beginning in 1996, a major survey redesign was implemented and information was collected using a Computer Assisted Personal Interview (CAPI) system. As a result, as long as the individual did not miss an interview, during the second and subsequent interviews, the CAPI instrument automatically assigned the same employer name and job ID each time further information about a continuing job was collected. When the respondent reported that a new job had started, the CAPI instrument assigned the next available job ID.

Job records were indexed by the longitudinal SIPP person ID, the wave (interview) number, and the job ID. We combined these records to create one observation per job that contained some time invariant job information, such as industry, and some time-varying information such as annual earnings. Table 1 shows the total number of respondents in each SIPP panel, the number that report holding at least one job over the course of the SIPP panel, the total number of person-wave-job records, and the total number of jobs reported, using the assigned SIPP job ID to count jobs. A careful examination of the person-wave-job records revealed serious problems with the SIPP job ID coding process. Because the definition of a "job" was so crucial to our comparison of job earnings from the SIPP and the DER data, we investigated the nature and causes of the job ID coding problems and developed an editing procedure that would resolve some of the inconsistencies we found. This section describes some of the problems we found and gives a summary of how we repaired the job id variables. Details about our edits are contained in Appendix A.

In the 1996 SIPP panel, the largest problem with jobs arose when the 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 generic 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. Consequently, 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 discrepancy arose. The case illustrated in Table 2 was the most common cause of the early start date problem. However it was not the only cause. The problem affected 21.6% of all jobs (29,520) and about 40% of the time there appears to have been a missing wave problem, while the rest of the time, the cause could not be determined. Whatever the reason, it was clear that the survey job IDs sometimes failed to link job records.

The problems encountered in the early SIPP panels (1990-1993) were considerably more complicated. There were two major types of problems - improper re-use of job IDs and improper assigning of new job IDs. Tables 3A and 3B give generic examples of these problems. In Table 3A, the SIPP respondent held the same job throughout the first four waves of the survey. However, in wave 3, the job ID was incorrectly changed, causing it to appear as if there had been a job transition. This error is identifiable because the name of the employer stays the same across the waves. Table 3B shows the second type of problem. In this case, the person changed jobs between waves 3 and 4 but the job ID was not changed. Thus, it appears that the person remained at the same job through all four waves and, consequently, a job transition was missed. Again, the true work history is apparent only through scrutinizing the employer names.

We developed an editing procedure that used employer name and person-level total job counts from the DER data to identify and correct SIPP job ID coding errors. This procedure involved several phases. First, employer names from job records were compared using statistical name matching software and those names deemed to be the same were grouped together by assigning new job IDs. Second, using these new job IDs, the total number of jobs held over the course of the survey was counted for each individual. A similar count was performed in the administrative DER data. A comparison was made between these two counts to identify cases where the name matching software had failed to correct or had introduced new job history errors. In the 1990-1993 panels, the group of problem observations was large enough to warrant further editing efforts. A second pass with the name matching software was performed and then a series of clerical edits was undertaken. In the 1996 panel, the job count comparison showed that the name matching step had corrected the most obvious problems and further editing was deemed unlikely to provide enough improvements to be worth the resource cost involved. More detail on the name matching procedure and the clerical edits can be found in Appendix A. A detailed description of the full process for the 1990-1993 panels can be found in Stinson (2003)[3]. Tables 4A and 4B provide a summary of the process for all the SIPP panels, showing how many unique jobs resulted from each step of the editing procedure and how many records were affected at each step. Row 5 of Table 4A and row 2 of Table 4B show the final number of jobs defined by the revised set of job IDs and row 6 and row 3, respectively, show the number of jobs belonging to people who still have discrepancies between job counts in the DER and the SIPP. We are confident that the majority of these cases are the result of reporting differences between the survey and the administrative data and not failure to link job records in the SIPP.

Once we had defined a set of jobs for each SIPP panel, we created annual earnings measures by summing monthly earnings reports. It is important to understand the concept of earnings as used during the SIPP interview. During the 1990-1993

[3] Recognizing that a similar edit could not be performed by researchers lacking access to the DER data, the Census Bureau has publicly released the revised SIPP job ids for the 1990-1993 SIPP panels on the SIPP website, http://www.sipp.census.gov/sipp/access.html.

SIPP panels, respondents were asked about earnings from a specific employer in the following way: "The next question is about the pay ... received from this job during the 4-month period. We need the most accurate figures you can provide. Please remember that certain months contain 5 paydays for workers paid weekly and 3 paydays for workers paid every 2 weeks. Be sure to include any tips, bonuses, overtime pay, or commissions. What was the total amount of pay that ... received BEFORE deductions on this job in ...?" The field representative reads the name of each month and separately records earnings for that month. A special caveat is added for members of the Armed Forces, "Be sure to include cash housing allowances and any other special types of pay." The intent of the survey question was to collect gross earnings and if the person responded that he or she did not know the earnings amounts, the field representative asked if the person could provide the information during a follow-up phone call.

The 1996 survey instrument asked, "Each time he/she was paid by [Name of Employer] in [MonthX], how much did he/she receive BEFORE deductions?" The field representative then followed up with questions about whether there were any other payments such as tips, bonuses, overtime pay, or commissions. The FR was trained to probe several times to make sure all the payments from an employer in a given month were accurately reported. There were also consistency checks built into the CAPI instrument that were meant to spot earnings amounts that seemed unreasonable and provide the FR with the opportunity to make corrections. Respondents were also asked to refer to earnings records if possible so as to give accurate responses. Thus in the best case, these earnings reports most likely reflected the gross pay from monthly pay stubs.

2.2 Creating a DER Jobs Data Set

The second source of data, Detailed Earnings Records (DER) from SSA, contained earnings histories for each SIPP respondent in the 1990, 1991, 1992, 1993, and 1996 panels with a validated SSN (for a definition and discussion of validation see section 3.3: "Matching SIPP and DER 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 during the time period covered by the survey questions were used[4]. Employers on this administrative data were identified by an IRS-assigned Employer Identification Number (EIN). Table 5 gives the total number of jobs that appear in the DER for SIPP respondents in all five panels and the total number of unique EINs, followed by the time period covered by the survey and the total number of jobs and unique EINs for this time period.

The earnings data contained in these DER files have as their source the W-2 records filed by employers on behalf of each employee. The primary earnings variable comes from Box 1 of the W-2 Form: wages, tips, and other compensation. This earnings variable is uncapped and represents all earnings that were taxable under federal income tax. There are at least two parts of earnings that would be reported on an employee's pay stub in "gross earnings" that are not included in Box 1: pre-tax health insurance plan premiums paid by the employee and pre-tax elective contributions made to deferred compensation arrangements such as 401(k) retirement plans. In the later case, these contributions are reported elsewhere on the W-2 form (for example Box 13 in 1999) and the DER file contains reports of these deferred earnings which can be added to Box 1 earnings to approximate gross earnings. While pre-tax health insurance plan premiums are reported on the W-2 Form, they are not contained in the DER extract created for research use. This omission represents one important way in which administrative records may differ from survey records that is not the result of error in the survey data collection process. DER earnings will be lower than SIPP earnings if the respondent reported gross earnings during the survey that included health insurance plan premiums.

There are other possible differences between Box 1 on the W-2 Form and gross earnings reported in the survey, most of which involve some kind of employee benefit that the employee is unlikely to consider wages and may also be unlikely to be reported as such on a pay stub, but which the employer is nonetheless required to report as taxable income. These include educational assistance above a certain monetary level, business expense reimbursement above the amount treated as substantiated by the IRS, payments made by the employer to cover the employee's share of Social Security and Medicare taxes, certain types of fringe benefits such as the use of a company car, golden parachute payments, group-term life insurance over $50,000 paid for by the employer, potentially some portion of employer contributions to Medical Savings Accounts, non-qualified moving expenses, and, in some circumstances, sick pay from an insurance company or other third party payer. In all these cases, DER earnings are likely to be higher than SIPP earnings.

A final potential problem with DER employer reports is that EINs do not necessarily remain constant over time. Unlike Social Security Numbers which serve as good longitudinal identifiers for individuals, EINs can change for many reasons that do not necessarily involve a person moving to a new employer. Company reorganizations that consist of mergers, acquisitions, or spin-offs of some parts of the company may result in a worker having two W-2 forms for a tax year, each with a different EIN, without having actually changed jobs. In cases such as these, the DER earnings will be lower than the SIPP earnings because a portion of the earnings for the year are missing. As part of the linking process between DER and SIPP earnings, we attempt to identify these kinds of successor-predecessor problems and merge the two DER jobs determined to be related to a single SIPP job (see Appendix C for details). However, at this early stage of research involving the administrative data, there is no way to know how well our method works.

[4] The Detailed Earnings Records did contain reports of self-employment earnings. 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.

The following list summarizes the potential definitional differences between SIPP and DER earnings.

Health insurance premiums not included in the DER	:	DER < SIPP
Employee benefits included in the DER	:	DER > SIPP
EIN changes due to change in firm organization or ownership	:	DER < SIPP

The EIN linked employers to the Business Register, the master list of all businesses maintained by the Census Bureau that serves as the sampling frame for firm-level surveys. Using this link, we merged information from the Business Register about the industry and name of the employer to each relevant job report in the DER data. Details about this merge can be found in Appendix B. The employer name is the key linking element between the SIPP and DER job data.

2.3 Matching SIPP and DER Jobs

After the creation of the SIPP and DER jobs data sets, the next step was to take people who had job reports in both files and try to match each SIPP job record to a DER job record. Table 6 shows the total number of people and jobs that were potential matches following the job record creation process. Except for the 1996 SIPP jobs, the total number of jobs that were potential matches is the same as row 5 in Table 4A for the SIPP jobs and row 4 of Table 5 for DER jobs. In 1996, one final problem necessitated the dropping of a few additional 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 3,908 job records of this type in the 1996 SIPP data, representing possibly triple that many actual jobs. These jobs were essentially impossible to match to the DER because they do not represent earnings from a single employer. Hence, they were dropped, giving a new total of 121,450 jobs.

We began the job matching process by first linking at the person level. The unique identifier for a person on the DER was the SSN while the SIPP 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, sex, race, and date of birth from the SIPP were compared to their administrative counterparts on the Numident. 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. Hence in order for an individual to have any earnings reports in the DER, he or she, by necessity, must have a validated SSN.

The third column of Table 6 shows the number of people who matched between the SIPP and the DER. In all panels, some people were lost from both the SIPP and DER job data sets as a result of this match. On the SIPP side, there were two reasons why a person might not match. First, he or she might not have a validated SSN. The third column of Table 7A shows the number of people affected by this problem. The second possibility was that the person had a validated SSN and reported working in the SIPP, but did not have any earnings reports in the DER. This would be caused by the jobs being relatively informal (baby-sitting, yard work, household help) and hence not generating W-2 forms, or by over-reporting of jobs. On the DER side, the only reason for a person not to match was because the person did not report any jobs in the SIPP survey. As seen in the third column of Table 7B, it was far more likely for a person to have jobs in the DER and not the SIPP than the reverse. It would appear that overall, the SIPP undercounts employed people.

As shown in Table 6, even for those people 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. In every SIPP panel, the survey asked employment questions of at least some respondents in the last few months of the year preceding the official beginning year. For example in the 1990 panel, the first interview reference period included between one and three months of 1989 for the second, third, and fourth rotation groups. Also the last interviews in the 1990 panel were conducted in September 1992, leaving the last quarter of 1992 uncovered for every rotation group. The 1991-1993 panels followed similar patterns. In the 1996 panel, the first interview reference period included December 1995 for the first rotation group and the last reference period included one or two months of 2000 for the third and fourth rotation groups. In order to attempt to match as many SIPP and DER jobs as possible, all DER jobs from the year preceding the main survey beginning year and jobs from the last survey year were included for the rotation groups that were surveyed at some point during these years. However, some of these DER jobs could clearly have ended before the survey began or started after the survey ended, thus artificially inflating the DER job counts. In the early SIPP panels (1990-1993), jobs that ended before the main survey year or began only in the survey end-year accounted for 27%-32% of all DER jobs. In the 1996 panel, these jobs only accounted for 11% of all DER jobs, largely because the timing of the survey conformed more closely to the calendar year. Another factor which artificially depressed the SIPP job counts is the fact that the survey only collected information covering a maximum of 2 jobs.

After we matched at the person-level by SSN, a job-to-job match was performed, again using the statistical name matching software Integrity. 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 certain matches using other fields. The primary basis for matching was self-reported name of the employer from the SIPP and administrative name of the employer from the Census Business Register. Earnings were not used in the match in order to prevent bias in the subsequent comparison of earnings. Appendix C gives the details of this match including which matching variables were used in each pass and how duplicate matches were handled. Table 8 describes the results of this job-level match. The first row of Table 8 gives the number of SIPP jobs that were successfully matched to a counterpart job in the DER. Of the jobs that matched, we then restricted ourselves to comparing earnings only in the full survey years. If a matched job did not have SIPP and DER earnings report in at least one full survey year, we dropped the job. The second row of Table 8 shows how many jobs were dropped because for this reason. This restriction ensured that we would be comparing jobs that had provided income during roughly the same time period. However, the timing of the job in the SIPP and the DER was not required to be identical. 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 DER job could have reports for all three years. In order for a job to be used in our comparisons, we required that there be at least one DER earnings report and one SIPP earnings report for that job but we did not require these reports to be in the same year. This resulted in slightly different sample sizes between the SIPP and the DER 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. The decision not to require exact matching in the earnings years was based on the fact that earnings essentially reported as zero in one source and positive in another source was a type of measurement error that we did not wish to exclude. The third row shows the final total number of jobs per panel that were used in the analysis. At this point jobs from all panels were combined to give a total of 197,337 jobs, 133,849 people, and 110,454 unique employers.

Tables 9 and 10 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. In the SIPP data, the correlations between adjacent years range from .54-.74. In the DER, they are higher, ranging from .79 to .81. The variance of earnings is also higher in the DER than in the SIPP. Table 11 gives the correlations between each year of DER and SIPP data. The correlations between SIPP and DER earnings in the same year is quite high: .74-.86. The correlations between adjacent years of SIPP and DER data are not as high as between adjacent years of DER data but the correlations are quite similar to adjacent years of SIPP data. They range from .58-.71. In general, correlations in the early 1990s are lower than correlations in the later years of the decade, as might be expected given the improvements of the 1996 panel.

3 Comparing Data Sources: observables and unobservables

Once the matching process was complete, we were able to compare SIPP and DER earnings at the job level. We began our comparisons with simple tables of means, stratified by SIPP demographic and economic variables. These tables show average differences between SIPP and DER earnings and allow us to consider which groups have the most pronounced differences. We look at all individual-job matches and then look separately at individuals who were on welfare and individuals who changed jobs. We also report regressions coefficients using both measures of earnings as dependent variables and compare the results.

After exploring differences in means, we estimate a model with both fixed and random effects (i.e. mixed effects model) that accounts for the observable differences discussed above and quantifies remaining differences in unobservable characteristics. Our modeling follows the spirit of Abowd and Card (1989). They examined the variance-covariance matrix of first-differenced earnings and tested the fit of various structural models, all of which included a measurement error component. Here our 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 effects but the parsing of variance among structural components is similar.

We estimate the following SIPP earnings equation:

$$\begin{aligned}\ln\left(SIPPEARN_{ist}\right) \;=\;& \beta_{oSIPP} + \beta_{1SIPP}Race.Gender + \beta_{2SIPP}Race.Gender.Educ + \beta_{3SIPP}Race.Gender.Exp_{it} + \\ & \beta_{4SIPP}Time_{it} + \beta_{5SIPP}\left[P_{1990}, P_{1991}, P_{1992}, P_{1993}\right] + \theta_i + \theta_{iSIPPDEV} + \psi_j + \psi_{jSIPPDEV} + \eta_{ist} + \omega_{ist}\end{aligned} \tag{1}$$

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

$$\begin{aligned}\ln(DEREARN_{ist}) \;=\;& \beta_{oDER} + \beta_{1DER}Race.Gender + \beta_{2DER}Race.Gender.Educ + \beta_{3DER}Race.Gender.Exp_{it} + \\ & \beta_{4DER}Time_{it} + \beta_{5DER}\left[P_{1990}, P_{1991}, P_{1992}, P_{1993}\right] + \theta_i + \theta_{iDERDEV} + \psi_j + \psi_{jDERDEV} + \eta_{ist} + \upsilon_{ist}\end{aligned} \tag{2}$$

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:

$P_{1990}, P_{1991}, P_{1992}, P_{1993}$ = vector of 4 indicator variables specifying the SIPP panel of the individual;
the 1996 panel is the excluded group

$Race.Gender$ = full interaction of male and white produces four categories:
white male, non-white male, white female, non-white female

$Educ$ = four levels of education fully interacted with race and gender:
high school degree, some college, college degree, graduate degree
separately for each demographic group

Exp_{it} = general labor market experience, fully interacted with race and gender:
actual exp. calculated using employment history collected in the SIPP;
experience enters as a piecewise linear spline with nodes at
2 years, 5 years, 10 years, and 25 years;
separate effects for each demographic group

$Time_{it}$ = calendar time, base year is 1990

Person heterogeneity = $(\theta) \sim N(0, G_1)$

Source-specific person heterogeneity = $(\theta_{SIPPDEV}, \theta_{DERDEV}) \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, G_{1DEV})$

Firm heterogeneity = $\psi \sim N(0, G_2)$

Source-specific firm heterogeneity = $(\psi_{SIPPDEV}, \psi_{DERDEV}) \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, G_{2DEV})$

Common error component = $\eta \sim N(0, G_3)$

Measurement error, SIPP and DER = $(\omega, \upsilon) \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, R)$
G_1, G_2, G_3, R are defined below.

This model accounts for average differences in the SIPP and DER using fixed effects for the intercept, race and gender interactions, education, experience, and a time trend. The effects, θ and ψ, are random person and firm effects, respectively, and capture unobservable effects of individual and firm heterogeneity. We also interact the person and firm effects with the data source indicator in order to tell whether there is source specific variation at the person and firm levels. Thus $\theta_{SIPPDEV}$, θ_{DERDEV}, $\psi_{SIPPDEV}$, and ψ_{DERDEV} represent deviations from the main person and firm effects. The effect η is a shared random error component that 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 both earnings measures, presumably by influencing "true" underlying earnings. The final terms in the model, ω and υ, are residuals that capture any remaining variation. Strictly interpreted, these terms capture variation across time within a job that is unique to each data source.

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 total number of time periods is 10. The maximum number of time periods a job may be observed depends upon the origin SIPP panel. In the 1990, 1991, and 1993 SIPP panels, there are 2 years of complete earnings data. In the 1992 panel there are 3 years and in the 1996 panel, 4 years. Thus a job may be observed anywhere from 1 to 4 years depending on the tenure of the job and the source panel.

Written in matrix notation, the model is

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

where Y is an $(N \times 10 \times 2) \times 1$ vector of stacked SIPP and DER earnings, X is an $(N \times 10 \times 2) \times k$ matrix of covariates treated as fixed effects, β is a $k \times 1$ vector of fixed effects coefficients, Z is an $(N \times 10 \times 2) \times (I + J + N \times 10)$ design matrix of the random effects, u is a $(I + J + N \times 10) \times 1$ vector of random effects and e is an $(N \times 10 \times 2) \times 1$ vector of residuals.

The fixed effects represent shifts in the conditional mean of the distribution of SIPP or DER earnings. For example, the β_{0SIPP} term is the mean of the entire SIPP earnings distribution and β_{0DER} is the mean of the DER earnings distribution. The vector β_{5SIPP}=[$\beta_{5SIPP1990}, \beta_{5SIPP1991}, \beta_{5SIPP1992}, \beta_{5SIPP1993}$] captures shifts in the mean of the panel-specific earnings distributions due to differences across SIPP panels. The equivalent vector β_{5DER} reflects shifts in the panel-specific DER earnings distributions due to the same cause.

The random effects capture variation in the data due to individual, firm, or time period heterogeneity that is left after controlling for observed characteristics. In other words, there is variation around the mean earnings due to unobservable characteristics of the person, employer, or time period for every category of individuals defined by the variables treated as fixed (X). The random effects quantify the amount of variance due to the different sources. The random effects vector, u, contains the stacked random effects, $\theta_1 ... \theta_I$, $\theta_{1SIPPDEV} ... \theta_{ISIPPDEV}$, $\theta_{1DERDEV} ... \theta_{IDERDEV}$, $\psi_1 ... \psi_J$, $\psi_{1SIPPDEV} ... \psi_{JSIPPDEV}$, $\psi_{1DERDEV} ... \psi_{JDERDEV}$, $\eta_{111990} ... \eta_{IN1999}$. The variance matrices for the random person, firm, and shared error component effects, respectively, can be written as

$$
\begin{aligned}
G_1 &= I_{I \times I} \otimes \sigma^2_\theta \\
G_{1DEV} &= I_{I \times I} \otimes \begin{bmatrix} \sigma^2_{\theta SIPPDEV} & 0 \\ 0 & \sigma^2_{\theta DERDEV} \end{bmatrix} \\
G_2 &= I_{J \times J} \otimes \sigma^2_\psi \\
G_{2DEV} &= I_{JxJ} \otimes \begin{bmatrix} \sigma^2_{\psi SIPPDEV} & 0 \\ 0 & \sigma^2_{\psi DERDEV} \end{bmatrix} \\
G_3 &= I_{NxN} \otimes \sigma^2_\eta \begin{bmatrix} 1 & \rho & \rho^2 & ... & \rho^9 \\ \rho & 1 & \rho & ... & ... \\ \rho^2 & \rho & ... & ... & \rho^2 \\ ... & ... & ... & 1 & \rho \\ \rho^9 & ... & \rho^2 & \rho & 1 \end{bmatrix}_{10 \times 10} \\
\text{where } \sigma^2_\eta &= \frac{\sigma^2_\varsigma}{(1-\rho^2)}
\end{aligned}
$$

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 nor across individuals. The person and firm deviation effects reflect that some additional, uncorrelated variation might exist in either the SIPP or the DER or both.

The error vector, e, contains the stacked terms, $\omega_{111990} ... \omega_{IN1999}, \upsilon_{111990} ... \upsilon_{IN1999}$. The SIPP and DER errors follow separate $AR(1)$ processes with the covariance between them constrained to be zero. These errors are identified by differences in the SIPP and DER earnings reports for each year. The variance matrix for the residuals can be written as

$$
R = I_{(N \times 1) \times (N \times 1)} \otimes \begin{bmatrix} \sigma^2_\omega \begin{bmatrix} 1 & \rho_{sipp} & \rho^2_{sipp} & ... & \rho^9_{sipp} \\ \rho_{sipp} & 1 & \rho_{sipp} & ... & ... \\ \rho^2_{sipp} & \rho_{sipp} & ... & ... & \rho^2_{sipp} \\ ... & ... & ... & 1 & \rho_{sipp} \\ \rho^9_{sipp} & ... & \rho^2_{sipp} & \rho_{sipp} & 1 \end{bmatrix}_{10 \times 10} & 0_{(N \times 10) \times (N \times 10)} \\ 0_{(N \times 10) \times (N \times 10)} & \sigma^2_\upsilon \begin{bmatrix} 1 & \rho_{der} & \rho^2_{der} & ... & \rho^9_{der} \\ \rho_{der} & 1 & \rho_{der} & ... & ... \\ \rho^2_{der} & \rho_{der} & ... & ... & \rho^2_{der} \\ ... & ... & ... & 1 & \rho_{der} \\ \rho^9_{der} & ... & \rho^2_{der} & \rho_{der} & 1 \end{bmatrix}_{10 \times 10} \end{bmatrix}
$$

where ρ_{sipp} and ρ_{der} are the autocorrelation terms of the SIPP (ω) and DER (υ) errors, respectively.

Estimates of β_{0SIPP} to β_{5SIPP}, β_{0DER} to β_{5DER}, the variance components (σ^2_θ, $\sigma^2_{\theta SIPPDEV}$, $\sigma^2_{\theta DERDEV}$, σ^2_ψ, $\sigma^2_{\psi SIPPDEV}$, $\sigma^2_{\psi DERI}$ ρ, σ^2_ω, ρ_{sipp}, σ^2_υ, ρ_{der}), and realizations of the random effects (θ, $\theta_{SIPPDEV}$, $\theta_{SIPPDEV}$, ψ, $\psi_{SIPPDEV}$, ψ_{DERDEV}, 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.

Our estimation of this model allows us to parse variation due to unobservables into common variation and source-specific variation. However we cannot talk about measurement error without making an additional assumption: namely what is true variation? One possible assumption is that all the common components are true variation and the deviations $\theta_{SIPPDEV}$, θ_{DERDEV}, $\psi_{SIPPDEV}$, ψ_{DERDEV}, ω, and υ are measurement error. Another assumption is that the common variation and the DER person and firm deviations are true and the SIPP deviations are measurement error. Using both these assumptions in turn, we calculate the reliability ratio, a summary measure commonly used in the literature that compares true variation to total variation. If only the common variation is considered true, then the formulas used are as follows:

$$\begin{aligned}\kappa_{SIPP} &= \frac{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_\psi}{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_{\theta SIPPDEV} + \sigma^2_\psi + \sigma^2_{\psi SIPPDEV} + \sigma^2_\omega} \\ \kappa_{DER1} &= \frac{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_\psi}{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_{\theta DERDEV} + \sigma^2_\psi + \sigma^2_{\psi DERDEV} + \sigma^2_\upsilon}\end{aligned} \tag{3}$$

If the DER person and firm deviations are true variation then the reliability ratios become:

$$\begin{aligned}\kappa_{SIPP} &= \frac{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_\psi}{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_{\theta SIPPDEV} + \sigma^2_\psi + \sigma^2_{\psi SIPPDEV} + \sigma^2_\omega} \\ \kappa_{DER2} &= \frac{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_{\theta DERDEV} + \sigma^2_\psi + \sigma^2_{\psi DERDEV}}{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_{\theta DERDEV} + \sigma^2_\psi + \sigma^2_{\psi DERDEV} + \sigma^2_\upsilon}\end{aligned} \tag{4}$$

A third possible assumption is that all variation in the DER is true in which case $\kappa_{DER} = 1$. Hence these assumptions provide a range of reliability ratios for the DER data. We will examine this range and consider what can be determined about the reliability of the survey data versus the administrative data. If the range is small, because κ_{DER1} is large, this will indicate a high level of similarity between the two data sources in unobservables.

4 Results

We begin by comparing average earnings for race-gender-education subgroups. We then look at earnings levels and changes for welfare recepients and job changers. We estimate a standard earnings equation and compare coefficients. After these four comparisons, we estimate the model described in section 3 and specified in equations 1 and 2 and calculate reliability ratios for SIPP and DER data.

4.1 Race-Gender-Education subgroups

In Table 12, we present average earnings for white males, white females, non-white males, and non-white females, divided by a five category education variable. The education categories are no high school degree, high school degree only, some college, college degree, and graduate degree. For every sub-group in this table, average earnings are higher in the DER (column 2) than in the SIPP (column 1). For most groups, the differences become more pronounced as the level of education increases. For example, white males with no high school degree report earning approximately $12,000 on average in the SIPP. However in the DER, this same group earns approximately $14,000 on average. This $2000 difference is about 16% of SIPP earnings (columns 3 and 4). In contrast, white males with a graduate degree earn about $44,000 on average according to the SIPP while the DER average is almost $59,000. This difference is 33% of SIPP earnings. Thus it appears the largest discrepencies between the two data sources appear for more not less educated individuals. This trend is also present for white females and non-white females, although to a lesser extent. For white females, the differences range from 10.6% to 12.7% of SIPP earnings and for non-white females the differences are between 14.5% and 18.1% of SIPP earnings. For non-white males, the education groups on either end of scale have the largest differences between SIPP and DER earnings, on average. Individuals without a high school degree have earnings 21.1% lower in the SIPP while those with a graduate degree have earnings 20.8% lower in the SIPP.

When comparing across race and gender groups, it appears that SIPP and DER differences are often higher for men than women and for blacks than whites. However this is not always true nor are all of the differences statistically significant. When comparing across white and non-white men, the only significant differences are in the college degree and graduate degree categories, where in fact, white men have higher differences between SIPP and DER earnings than non-white men (23.4 versus 21.9 (significant at the 10% level) and 33.3 versus 20.8 (significant at the 1% level)). When comparing white and non-white women, the only significant difference is for women without a high school degree (10.6 versus 14.5 (significant at the 1% level)). There are more significant differences when comparing men to women. White men have larger differences between self-reported and administrative earnings than white women for every education category and these differences are all significant[5]. Non-white males have higher differences than non-white females but only in the college degree category is the difference significant.

[5] In differences between white men and white women are all significant at the 1% level except for the no high school degree category where the difference is significant at the 10% level.

Standard deviations follow similar patterns to average earnings; they are universally higher in the DER and this difference varies by education. There seems to be more dispersion in earnings as education levels increase and this is more pronounced in the DER than in the SIPP. The standard deviations are large due to the presence of a number of high earners in the left tail of the distribution.

These results are somewhat surprising given that concern about under-reporting earnings has often focused on lower-income and less educated individuals. In fact it appears that the real systematic differences between the SIPP and the DER happen for the most educated. One possible reason for this might be that highly educated professionals receive a large part of their compensation in the form of end-of-the-year bonuses and the SIPP frequently misses these one-time payments. Another reason might be that highly compensated individuals report a measure of earnings in the SIPP that does not include deferred compensation or some other form of compensation that these individuals consider separate from wages. However these earnings are still reported in Box 1 or Box 13 of the W-2 and so show up in total compensation in the DER. The problem is most pronounced with the highly educated because they are most likely to have these different types of compensation. Another possible explanation is missing data imputation. When individuals miss months of the survey in the middle of the panel, earnings are imputed. Perhaps these imputed earnings are too low or at least contribute to the lower level of variance.

4.2 Welfare Recepients

In Table 13, we look at average earnings summed across jobs for individuals classified based on whether they ever received welfare in 1996 or 1997, whether they worked in 1996 or 1997, and whether they were single females or not. Individuals in this table were exclusively from the 1996 SIPP panel. The first four lines present average earnings for individuals on and off welfare who worked both years and were not single females. The first row shows average earnings for 309 individuals who were on welfare at some point in both 1996 and 1997. The next row shows individuals who were on welfare at some point in 1996 and never on welfare in 1997. The third row shows the reverse: individuals who were never on welfare in 1996 and on welfare at some point in 1997. The last row shows individuals who were never on welfare in either year. Again, earnings in the DER are on average higher than in the SIPP. However, changes in earnings are not as far apart as might have been expected and none of the differences are statistically significant. For individuals moving off welfare, earnings increased by 21.4% in the SIPP and 24.3% in the DER. For individuals moving onto welfare, earnings decreased between 2% and 6% for the SIPP and DER respectively. For individuals never on welfare in 1996 and 1997, earnings changes were very similar - 5.7% versus 5.2%. The results are similar for single females, reported in the second set of four rows. For those moving off welfare, wages increased by 39% in the SIPP and 35% in the DER. For those moving onto welfare, the earnings change was -19.8% in the SIPP and -19.1% in the DER. The earnings changes are larger in the SIPP than in the DER for single females, which is mostly the reverse compared to all others. However since none of these changes are statistically significant, one cannot be certain of a unique pattern for single females. The small sample sizes in these various sub-groups make it difficult to obtain much statistical precision.

The next group of results is for individuals who moved off welfare from 1996 to 1997 and started paid employment in 1997 after not working 1996. This group is compared to individuals who started working in 1997 but were not on welfare in either year. We again divide the group into those who are single females and all others. Here it is quite surprising that for single females and all others moving off welfare and getting a job, the point estimate of average SIPP reported earnings in 1997 is not very different from the DER point estimate for the same year. There is also no statistically significant difference between the two data sources, something that is less surprising given the small sample sizes. For individuals starting a job in 1997 who had never been on welfare, there is a statistically significant difference between SIPP and DER earnings for individuals who were not single females, with DER earnings being about 29% higher. DER point estimates are higher for single females not on welfare but the difference between the SIPP and the DER is not statistically significant.

Finally we look at individuals who move onto welfare in 1997 and have no reported earnings in 1997 after having positive earnings in 1996. We compare them to individuals who quit earning in 1997 but who were never on welfare. These subgroups have lower SIPP earnings on average than in the DER but for individuals who were not single females, the difference is only significant for those never on welfare. For single females starting welfare receipt, SIPP-reported earnings in 1996 are about 50% lower than in the DER compared to 30% lower for single females not going on welfare. However these differences are not significantly different from zero.

In general earnings levels for welfare receipients are lower in the SIPP than in the DER but the changes in earnings between years are similar. In some cases the differences between SIPP and DER earnings is actually lower for welfare receipients as in the case of individuals moving off welfare and beginning to work. Unfortunately the sample sizes are small and this result will need to be verified in the other SIPP panels.

4.3 Job Changers

In Table 14 we report earnings changes for individual who switch jobs. We define a job switch as an individual who reports two consecutive jobs that do not overlap. There may be a gap between the ending date of job 1 and the starting date of job 2 but the starting date of job 2 cannot come before the ending date of job 1. We allow at most one job switch per individual. Some individuals have only one job and others have jobs that overlap and these individuals are not included in our table. The first row shows that on average the SIPP-reported change in earnings after switching jobs is small, about $40, and is not siginficantly different from zero. In contrast, in the DER, on average job-changers earn around $900 less in the second job. However these results differ a great deal by industry of the first and second jobs. The first group of 15 rows reports earning changes classified by respondent-reported industry of the first job. Four industries are similar to the overall average with small earnings gains reported in the SIPP and losses reported in the DER: Construction, Retail Trade, Personal Services, and Professional Services. Retail Trade is slightly different than the others in that the DER change is only -$6 while the SIPP change is over $600. Three other industries report positive changes in earnings from switching jobs - Agriculture, Business and Repair Services, and Entertainment and Recreation Services. The SIPP reports a higher positive change for the first 2 of these industries and is just slightly lower for the last one. These three industries seem to have the closest SIPP and DER reporting. The remaining seven industries have negative earnings changes between jobs in both the SIPP and the DER but the changes are substantially smaller in the SIPP than in the DER. For example individuals with a job in the Finance, Insurance, and Real Estate (FIRE) industry who switched jobs reported a decline in earnings of just over $400 in the SIPP but their W-2 records show a decline of almost $2000.

The next group of 14 lines shows earnings changes for individuals classified by the industry of their second job. Here five industries have positive SIPP changes and negative DER changes, three industries have positive SIPP and DER changes, and six industries have negative earnings changes in both data sources. When earnings changes are positive in both sources, the SIPP report is substantially higher than the DER report (Public Administration, FIRE, and Manufacturing Non-durable). When both sources report negative earnings changes, the DER is substantially more negative, with the exception of Personal Services where the change is essentially the same. It is interesting that some industries have different signs on earnings changes between the Job1 table and the Job2 table, for example FIRE. This is probably due to the fact that many individuals with FIRE as their original job industry, have a different industry at their second job and they earn less in this new industry. For individuals with FIRE as their second job industry, they may have switched out of a lower-paying industry. This effect is captured in both the SIPP and the DER although the magnitude of the change is different between the two data sources.

The last two lines show earnings changes for individuals who do not switch industries and those who do switch industries. Overall, there do not appear to be significant differences among industry changers and stayers. Their SIPP-reported changes in earnings are very similar as well as their DER earnings changes. They follow the overall pattern of the data with slightly positive earnings gains in the SIPP and larger negative earnings losses in the DER.

Overall it appears that when earnings increase at the time of a job switch, the increase is stronger in the SIPP than in the DER and when earnings decrease, the decrease is stronger in the DER than in the SIPP. This leads to the SIPP reporting earnings gains on average while the DER reports earnings losses on average.

4.4 Earnings Regressions

In table 15, we present results from two earnings regressions, one using SIPP data for the dependent variable and the other using DER data. Our dependent variable is an annualized log wage rate. We calculated this wage rate by dividing total annual earnings by total annual hours worked. The information on hours comes from the SIPP survey and is used to create both the SIPP and DER wage rate. For each year, we chose a dominant employer based on earnings and then kept all annual earnings for that employer. In general this leads to only one observation per person per year but sometimes when employers overlapped, there were multiple observations per year. The explanatory variables are SIPP variables and are identical in both regressions. As fixed effects, we include an intercept, indicators for the source SIPP panel (1996 is the excluded case), a linear time trend, interactions of race and gender (white males are the excluded case), interactions of race, gender, and five levels of education (no high school degree is the excluded case), and interactions of race, gender, and a piecewise linear spline in experience. As random effects, we include a person effect and a person/labor force experience interaction and allow these effects to be correlated. These two terms allow for individual deviations from the overall intercept and from the overall labor force experience slope due to unobservable individual characteristics. In addition, we include a random firm effect and specify an AR(1) process for the error term.

The intercepts are remarkably similar for the SIPP and DER regressions but the panel indicators are substantially higher in the DER than in the SIPP. This means that there is a much larger difference in the overall average mean of SIPP and DER earnings in the panels conducted in the early 1990s compared to the panel that began in 1996. This differences is between 5%-10%. Given the fact that the SIPP was re-designed and switched to computer-assisted interviews for the 1996 panel, it is not surprising that the differences between the SIPP and DER are smaller for this panel. The linear time effect and race and gender interactions are also very similar. The coefficient on non-white males is small and positive in the DER and just barely negative in the SIPP but it is not significantly different from zero in either data source. The education coefficients compare differently depending on the demographic group. For white men, the education coefficients are slightly larger in the DER than in the SIPP by between1%-2.5%. The differences for college and graduate degrees are higher than the differences for high school degree and some college. For non-white males, the education coefficients are 1%-4% smaller in the DER than in the SIPP. White females are similar to white males, with earnings 1%-3% higher in the DER. Non-white females with a graduate degree have the biggest gap, with the DER coefficient being 7% higher than the SIPP coefficient.

To better visualize the differences between the labor force experience splines, we have graphed the experience profiles of the four demographic groups using both the SIPP and DER spline coefficients. The intercept for each profile is an individual in the 1996 SIPP panel with a high school degree. Figure 1 shows results for white men. Initially, the increase in earnings due to labor force experience accumulation is slower in the DER than in the SIPP. However, by around 14 years, the DER earnings have caught up and from then on, surpass the SIPP earnings. The gap continues to widen over the later years of an individual's career. Profiles for white women follow a similar pattern, as shown in Figure 2. However, the cross-over point for the DER and SIPP is not until approximately 25 years of experience. Non-white men have DER earnings well below SIPP earnings for much of the profile, reflecting a lower intercept for non-white men with a high school degree in the DER than in the SIPP and then lower growth for the first 2 years. After that the DER growth rate is higher than the SIPP and earnings cross around 25 years again. For non-white women, the DER growth rate is again lower for the first 2 years but after this, the DER catches up very quickly and surpasses the SIPP by year 6.

In general it appears that returns to experience in the first 2 years are higher in the SIPP than in the DER. For years 3-5, returns are higher in the DER. For years 6-10, the rate of return is quite close for all groups except non-white males where the DER is substantially higher. For years 11-25 and 26+, the DER is again uniformly higher. Hence it appears that there is a range over which the effect of labor force experience is very similar between the SIPP and the DER. However there are also portions of the profile where the effect is quite different.

Finally at the end of Table 15, we list values for the variance components of the mixed effects models. The main person effect, or person intercept as we label it in the table, is three times as large in the DER as in the SIPP. This result means that for this type of regression, person heterogeneity is much higher in the SIPP than in the DER. The person slope, or interaction of person and labor force experience, is 2.5 times larger in the DER than the SIPP. The firm effect is 3 times larger in the DER than in the SIPP. Finally, the variance of the error term is significantly higher in the DER but the AR(1) correlation coefficient is similar in both regressions. DER earnings overall have higher variance and this is reflected in the random effects.

4.5 Reliability Ratios

The models used to estimate reliability ratios are similar to the earnings regressions described in the previous section but with some important differences. First, the sample of people-job matches used in the estimation is different. All person-job matches that have SIPP and DER earnings in some years are used to estimate variance components used for calculating reliability ratios. We do not select a dominant employer. This is because we want to quantify the sources of variance for all the data, not estimate economic relationships. Second, we use earnings not wages because it is variation in earnings that is of interest. Finally, we jointly estimate equations 1 and 2 so that there are some common variance components and some components particular to either the SIPP or the DER. In Table 16, we show the estimated individual components and then combine them using the different reliability ratio formulae given in section 3. The main person and firm effects are .28 and .32 respectively. The interactions of person and firm and data source produce variance components that go to zero for the SIPP. Essentially there is no variation left in the SIPP at the person and firm levels after taking account of the variation that is common to both the SIPP and the DER. However there is additional variation in the DER and θ_{DERDEV}=.04 and ψ_{DERDEV}=.1. These magnitudes mean that about 25% of the variation due to unobservable firm characteristics in the DER is not found in the SIPP and about 10% of the variation due to unobservable person characteristics. The variance in the SIPP measurement error term is also lower than the DER measurement error term and the SIPP error is less correlated over time. The common time period component has a higher variance than either measurement error term. The magnitudes imply that of time period specific variation in the SIPP, 23% is unique to the SIPP and 33% is unique to the DER. We calculate $\kappa_{SIPP} = .86$ and $\kappa_{DER1} = .72$. These magnitudes indicate that approximately 70% of total DER variation due to unobservables is common to both the SIPP and the DER. In the SIPP 86% of total variation is common to both sources. The higher ratio in the SIPP reflects two things. First, overall variation in the SIPP is lower than in the DER. Second, σ_ω is smaller than σ_υ. If we believe that the SIPP is missing variation that is actually true variation, then a higher reliability ratio is not an indication of less measurement error. If we believe that $\sigma^2_{\theta DERDEV}$ and $\sigma^2_{\psi DERDEV}$ are really true variation then the appropriate reliability measure for the DER is $\kappa_{DER2} = .80$. Thus the range of ratios for the DER, depending on how much of the DER variation one is willing to term truth, is $.72 - 1$.

If one chooses to adopt the hypothesis that $\sigma^2_\eta + \sigma^2_\theta + \sigma^2_\psi$ represents true variation and all other variation is measurement error of some kind, then the overall level of SIPP variation is 14% too high. However if one chooses instead to adopt the hypothesis that $\sigma^2_\eta + \sigma^2_\theta + \sigma^2_{\theta SIPPDEV} + \sigma^2_\psi + \sigma^2_{\psi SIPPDEV}$ is true variation, then overall variation in the SIPP is too low. The ratio $\frac{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_\psi}{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_{\theta SIPPDEV} + \sigma^2_\psi + \sigma^2_{\psi SIPPDEV}}$ which is the ratio of common variation to true variation under the second hypothesis is .91 which indicates that about 9% of true variation is missing from the SIPP. If one chooses to believe that all variation in the DER is truth, then the ratio becomes $\frac{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_\psi}{\sigma^2_\eta + \sigma^2_\theta + \sigma^2_{\theta SIPPDEV} + \sigma^2_\psi + \sigma^2_{\psi SIPPDEV} + \sigma_\upsilon}$ which is equal to .72. Under this hypothesis, about 28% of true variation is missing from the SIPP. This later ratio is exactly equal to κ_{DER1} but the interpretation is different because of the different assumption made about the DER.

Which hypothesis about the DER to adopt is not answered by the data. However it is interesting to note that σ_υ and ρ_υ are both relatively large in magnitude. Thus this DER-specific time period effect accounts for more variation than the total DER person effect and, unlike classical measurement error, it does not immediately die out in the next time period. Because of these estimates, we hesitate to label υ as strictly measurement error and hypothesize that the true reliability ratio of the

DER is somewhere between $.8 - 1$ and the SIPP is missing between 9% and 28% of true variation.

5 Conclusion

In comparing the SIPP and the DER we have found two consistent results. DER earnings are on average higher than SIPP earnings and there is more variation due to unobservables in the DER than the SIPP. Of the definitional differences discussed earlier, it appears that lack of health insurance premiums and EIN changes are not dominant factors since these would give rise to lower DER earnings on average. We cannot say for certain how much of the differences we find might be due to employee benefits appearing on a W-2 form and not on a pay stub. However our opinion is that these differences are unlikely to be solely the result of the SIPP and the DER measuring different quantities. In particular, it seems likely to us that highly educated SIPP respondents with high incomes do under-report their earnings to some extent. Our data on job changers are particularly interesting in that the major differences between the SIPP and the DER are in earnings changes not levels.

Our results which examine differences due to unobservables lead us to believe that there is too little variation in SIPP earnings. Without further research, we cannot give a definite reason why this might be the case. However, we hypothesize that difficulty in capturing with-in year fluctuations in pay and imputation procedures that use earnings from one wave to impute earnings values for missing waves contribute to the lower SIPP variation. The SIPP collects earnings at the monthly level so to some extent there will always be difficulties in comparing to an administrative database that is annual. In light of these results, it might be useful to consider ways in which the SIPP could better capture an annual earnings measure that included bonuses and infrequent extra earnings. Also, an imputation procedure which used the DER to help model SIPP earnings could reduce the differences between the two sources, by allowing draws for SIPP earnings values to be taken from a distribution with greater variance.

In spite of the differences we find, we feel that there are reasons to be confident in the use of SIPP data. Of the variation that is found in the SIPP, 86% of it can also be found in the DER. Earnings regressions using SIPP and DER data produce similar coefficients. Total earnings changes across years for individuals moving on and off welfare are similar in both the SIPP and the DER. More work is needed to continue to study effectively measuring earnings using the SIPP survey instrument. Ideally, the administrative data will continue to inform this study and contribute to SIPP improvements.

References

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

[2] Abowd, John M. and Lars Vilhuber. 2002. "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers." *Technical Paper TP-2002-17*, LEHD, U.S. Census Bureau.

[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] 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.

[6] Fellegi, I.P. and A.B.Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64:1183-1210.

[7] 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.

[8] Newcombe, H.B., J.M. Kennedy, S.J. Axford, and A.P. James. 1959. "Automatic Linkage of Vital Records." *Science* 130:954-959.

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

[10] Stinson, Martha H. 2003. "Technical Description of SIPP Job Identification Number Editing, 1990-1993 SIPP Panels." *LEHD Technical Paper*, U.S. Census Bureau.

6 Appendix A Editing of SIPP JOB ID Variable

As described in Section 3.1, the SIPP job id variable had difficulty in longitudinally tracking jobs (see tables 2, 3A, 3B for examples of the problems). In the early SIPP panels, the problems were mostly the result of field representatives being required to collect information about an on-going job over and over again. Inconsistencies crept in over time as the name of the employer was collected and written down separately at the time of each interview. Wave-specific names differed both across and within the original SIPP job IDs. Different spellings, use of abbreviations in later waves, and slightly different wording were the most common differences within job IDs. In contrast, the 1996 panel only recorded employer names when new jobs were begun and hence employer names differed only across job IDs and not within. In the 1990-1993 panels, the goal was to create an entirely new set of job IDs that was not derived from the old job IDs because these were deemed too unreliable. Hence it was necessary to compare all person-job-wave records for a given individual and group those with the same name. In 1996, however, the goal was simply to check and see if some jobs for an individual should be linked because an individual may have missed a survey wave and been incorrectly assigned a new job id when he or she was next interviewed. Hence person-job-wave observations with the same job ID assigned were accepted as belonging to the same job and job-level records were created. These job-level records were then compared and those with names deemed to be the same were grouped together.

Because of the spelling and wording differences across observations, we used probabilistic matching methods as developed by Newcombe, Kennedy, Axford, and James (1959) and Fellegi and Sunter (1969) and implemented in a commercial software program called Integrity. These methods have been used extensively at the U.S. Census Bureau to solve problems of miss-coded identifiers (for an example of an application of probabilistic person name matching to fix SSN miscodes, see Abowd and Vilhuber (2002)). This method involves grouping records into "blocks" of possible matches and then computing matching weights for pairs of records within the "block". Pairs with matching weights above a certain threshold, or cutoff point, are deemed to be matches and those with weights below another threshold are deemed to be non-matches. Those pairs with matching weights in between the two thresholds are termed uncertain and clerical review is suggested.

A matching weight for a pair of records is a composite score that is created by comparing the records across a variety of fields, assigning a weight to each field based on a determination of whether the field agrees or disagrees, and then summing the weights from all the fields involved in the comparison. Each field used in the matching is assigned an m and u probability. The m probability is the probability that the same field on two separate records agrees given that the two records were indeed a match. When this probability is set to less than one, it is assumed that there are some errors in the fields and that even if two records are a match, there is still some probability that the field is miscoded on one of the records and the two fields will disagree. The u probability is the probability that the same field on two separate records agrees given that the records are not a match. This is the probability that a field agrees at random. Given the m and u probabilities, agreement and disagreement weights for each field are calculated using the following formulas:

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

The decision of whether a field agrees or disagrees, and hence whether it receives the agreement or disagreement weight, can be implemented in a variety of different ways. One can be quite strict and insist on absolute identity in order to declare agreement or one can allow some level of discrepancy between fields without declaring disagreement. This flexibility is especially useful for name matching because it allows the user to take account of potential misspelling of words.

In our application, we blocked on the SIPP person identifier and hence only job records for the same person were compared. To create the fields for comparison we parsed the reported name into several pieces. Common words such as "Inc," "Company," or "Firm" were saved in one set of fields while geography words such as state names were saved in another set. The remaining words from the name were thought most likely to be unique to a particular employer and were saved in a third set of fields. We performed several sets of comparisons, or passes, using different fields in each pass. The choice of m and u probabilities and cutoff levels was determined both by knowledge about the fields and by experimentation. For the fields that contained unique name words, a high m probability and low u probability were chosen. Since these words were deemed to be the part of the employer name that was unique to that firm, matching values were essential to matching records, thus requiring the high m probability. At the same time, these words were 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 the unique word names received very high agreement weights and also very high disagreement weights. The fields that contained common words and geography words, 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. Cutoff values were chosen by examining certain and uncertain matches and determining the range of their weights. Appendix Table A1 gives the exact blocking and matching fields used along with their m and u probabilities.

For the early SIPP panels, probabilistic name matching alone proved inadequate for creating a consistent set of job IDs. While the name matching procedure both separated jobs records originally assigned the same job ID and connected job records originally assigned different job IDs, the former was the most common outcome. This can be seen in Table 4A by the fact that the number of total jobs rose substantially after Integrity processing. This result was due to the fact that the most common problems in the survey were the re-use of job IDs, as described in Table 3B, and the high degree of irregularity in the spelling of job names and the common use of abbreviations in later waves. Universities and government agencies with common acronyms were especially problematic. For example, the Integrity software could not recognize the names "University of X" and "UofX" or "Department of Y" and "DOY" as being the same. Hence in these panels, a considerable amount of clerical review was undertaken in order to separate cases where Integrity correctly and incorrectly split job records. In cases where there were discrepancies between the total job count in the SIPP and the total job count in the DER, job records were output and reviewed by two separate individuals. When one of the reviewers discovered two jobs in a respondent's job history that appeared to be the same, she manually changed the job ID to reflect this determination. The second reviewer re-checked all these changes as a quality assurance measure. After this extensive manual review, a few final edits were performed to locate any final obvious cases where Integrity had erroneously failed to link job records. The work history of any person who had one job that consisted of at least four linked job records and a second job that consisted of only one job record was examined to see whether the single job record in fact belonged to the job with at least four linked records. Corrections to job IDs were made to link job records that were determined to belong to the same job.

7 Appendix B

The merge between the DER and the Business Register 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, we also searched for information about the EIN in the second part of the Business Register, the Multi-unit file. Here we 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, we kept one record for each unique three-digit 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.

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 at the time of the beginning of the survey, the start year was coded to be the first survey year 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. Appendix Table B1 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, the match rates are extremely high, 98% for every panel except 1996. The low match rate for the 1996 panel relative to the other four panels can be explained by the fact that the latest year for which the Business Register was available at the time this research was conducted was 1999. Thus any job in the DER for a SIPP respondent from the 1996 panel that began in the year 2000 could not be matched to the Business Register. For the purposes of this study, this lack of data 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 accurately constructed for SIPP jobs. As described in Section 3.3., jobs beginning in the year 2000 were dropped from both the SIPP and the DER before comparing earnings. The second interesting thing to note is that although only 27%-32% of all EINs were multi-unit companies, these EINs accounted for 39%-44% of all jobs. SIPP respondents disproportionately work for large companies. Third, a small percentage of EINs and jobs 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.

8 Appendix C

The job-level match between the SIPP and DER data compared employer name, calendar year indicators, and industry in order to link records from each source. On one side of the match were all the SIPP jobs deemed to be reports of employment at a single employer. 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 Multi-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. This was done in an attempt to maximize the number of job matches obtained by using all possible name information associated with multi-unit companies. For example a person might report working for company X in the SIPP and have a job report in the DER with EIN A that is a multi-unit. The main company name of EIN A may be Y but one of the subsidiary establishments may be called X. By attaching names X and Y to EIN A in the DER, we increase the likelihood that this job will match to the SIPP job reported at company X.

Appendix Table C1 gives the blocking and matching fields for each pass along with the accompanying m and u probabilities for the matching fields. Several variables were also used in multiple passes, with the requirements for matching gradually relaxed. For example, in the third pass, three-digit Single-unit 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 with the first possible year being the year that data was first collected in the survey and the last possible year being the last year data was collected in the survey. Likewise in pass seven, only one-digit Single-unit 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.

Appendix Table C2 shows the results of the matching. Of the SIPP jobs, between 77%-79% were successfully matched to a DER job. Of these matches, 86%-88% were deemed high probability matches that surpassed the clerical editing threshold, while the remaining matches were between the clerical threshold and the no-match cutoff point or were duplicate matches. The majority of the matching took place in the first pass (between 75%-83% of all matches). The next most successful passes were 3 (5%-9%) and 7 (5%-6%).

Appendix Tables C3 and C4 highlight two problems that resulted from the matching. First, two different SIPP jobs could match to the same DER Job. An example of this type of case is illustrated in Table C3. 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. Careful inspection of duplicate cases led to the adoption of the following rule: if the two SIPP jobs had been matched to the one DER job in either the first or second pass and there were 2 or fewer residual DER jobs left that had not matched to any SIPP job, then the second SIPP job was declared a true duplicate. It was combined with the first SIPP job to become one single SIPP job matched to the one DER job. Otherwise if the two SIPP jobs had matched to the one DER job in pass 3 or later or they had matched in pass 1 or 2 and there were 3 or more residual DER jobs, the duplicate was declared false and only the master record match was kept. The duplicate SIPP job was changed to be a residual, non-matching SIPP job. The total number of duplicates that were determined to be “true” and hence were subsequently combined is shown in row 3 of Table 9.

The second problem was the reverse duplication issue: two different DER jobs sometimes matched to the same SIPP job. Table C4 gives an example. 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. This is the successor-predeccessor problem described earlier. 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 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.

We made a first attempt to tell whether two DER jobs that matched one SIPP job were indeed the “same” job by using some additional information from the Census Business Register. Previously we had augmented our list of EINs from the DER to include parent company information and name and industry information from one establishment of every unique three-digit industry group within the parent company. We then added annual geography information (a geocode created from the exact address) to each EIN at both the parent company level and the establishment level. We compared this geography information for each year of the survey across the two EINs and if the geocode was ever the same for the parent company or the establishment, we declared the two jobs to be duplicates. The intent of this geocode comparison exercise was to find cases where an EIN changed but the physical location did not change and hence it was likely that the SIPP respondent still considered himself to be at the same job. Since we did not keep every establishment within an industry group, we clearly did not compare every possible geocode. Hence our determination of how many DER jobs were duplicates and should be combined is probably an undercount.

We also added parent company identifiers to the EINs so we could tell if two EINs had some kind of ownership relationship. Two DER jobs that matched to one SIPP job but had the same parent company identifier were also declared to be a match. In this case it seemed possible that the SIPP respondent had kept the “same” job but had moved within the company or had simply experienced a company re-organization where the EIN tax reporting structure had changed. Row 4 of Table C2 shows how many DER job duplicates were determined to be legitimate.

Table 1: Original SIPP Job Summary

SIPP Panel	1990	1991	1992	1993	1996
Total SIPP respondents	69,432	44,373	62,412	62,721	116,636
Respondents who ever report a job	37,291	23,520	33,920	32,972	63,600
Person-job-wave observations	216,851	136,693	228,214	208,748	498,553
Jobs defined by original SIPP jobid	57,800	35,515	55,453	52,591	136,550

Table 2: SIPP Job ID Problems, 1996 Panel

Wave	Start date	Jobid
1	Feb. 1, 1996	1
2	Feb. 1, 1996	1
3	Feb. 1, 1996	1
4	Feb. 1, 1996	1
5		
6	Jan. 1, 1996	2

Table 3A: SIPP Job ID Problems, 1990-1993 Panels

Failure to link job across waves		
Wave	Firm Name	SIPP Jobid
1	AAAA	1
2	AAAA	1
3	AAAA	2
4	AAAA	1

Table 3B: SIPP Job ID Problems, 1990-1993 Panels

Failure to separate jobs across waves		
Wave	Firm Name	SIPP Jobid
1	AAAA	1
2	AAAA	1
3	AAAA	1
4	BBBB	1

Table 4A: Summary of the Job ID Editing Process, 1990-1993 SIPP Panels

	SIPP Panel	1990	1991	1992	1993
1	Number of Jobs, post name matching pass 1	78,225	46,316	74,078	68,803
2	Jobs belonging to people with conflict with DER	45,725	24,149	38,752	35,352
3	Number of Jobs, post name matching pass 2	69,138	41,814	66,602	62,251
4	Jobs belonging to people with conflict with DER	10,011	5,106	8,131	8,330
5	Number of Jobs, post clerical edits	66,991	40,818	65,278	61,094
6	Jobs belonging to people with conflict with DER	7,089	3,800	6,448	6,670

Table 4B: Summary of Job ID editing process, 1996 SIPP Panel

	SIPP Panel	1996
1	Jobs with startdate probs	29,520
2	Number of Jobs, post name matching	125,358
3	Jobs belonging to people with conflict with DER	15,331
4	Jobs with startdate probs	22,353

Table 5: Jobs from the DER

	SIPP Panel	1990	1991	1992	1993	1996
1	Total DER jobs	432,105	263,063	364,261	347,485	607,873
2	Total EINs	235,910	155,013	205,951	197,881	315,471
3	Years covered by survey	1989-1992	1990-1993	1991-1995	1992-1995	1995-2000
4	DER jobs in survey time	96,086	58,020	99,524	81,320	192,720
5	EINs	60,131	38,628	62,406	51,880	105,095

Table 6: Match Rates for People with Jobs in the SIPP and DER

				Both	
SIPP Panel		SIPP	DER	SIPP	DER
1990	People with jobs	37,291	35,032	30,993	
	Total Jobs held	66,991	96,086	55,087	88,324
1991	People with jobs	23,520	21,729	19,056	
	Total Jobs held	40,818	58,020	32,447	52,797
1992	People with jobs	33,920	31,557	27,394	
	Total Jobs held	65,278	99,524	51,650	90,360
1993	People with jobs	32,972	29,831	26,267	
	Total Jobs held	61,094	81,320	47,723	74,317
1996	People with jobs	63,116	55,894	48,542	
	Total Jobs held	121,450	192,720	97,149	173,623

Table 7A: Reasons SIPP Workers Do Not Match DER

SIPP Panel	Total SIPP People	SIPP People without Valid SSNs	People Who Have Only SIPP Jobs	People in SIPP and DER
1990	37,291	4,856	1,442	30,993
1991	23,520	3,629	835	19,056
1992	33,920	5,477	1,049	27,394
1993	32,972	5,535	1,170	26,267
1996	63,116	12,425	2,149	48,542

Table 7B: Reasons DER Workers Do Not Match SIPP

SIPP Panel	Total DER people	DER People without Valid SSNs	People Who Have Only DER Jobs	People in SIPP and DER
1990	35,032	0	4,039	30,993
1991	21,729	0	2,673	19,056
1992	31,557	0	4,163	27,394
1993	29,831	0	3,564	26,267
1996	55,894	0	7,352	48,542

Table 8: Final Sample of Matched Jobs

SIPP Panel	1990	1991	1992	1993	1996	Total
Number of Matched Jobs after Combining Duplicates	41,885	25,258	39,729	36,469	75,110	218,451
Jobs w/out SIPP and DER Earnings in Sample Years	5,716	3,497	2,706	6,904	2,291	21,114
New Matched Job Total	36,169	21,761	37,023	29,565	72,819	197,337

Table 9: Covariance/Correlation Matrix for Ln(SIPP Job Annual Earnings)

	1990	1991	1992	1993	1994	1996	1997	1998	1999
1990	2.0293	0.61							
1991	1.0170	2.0643	0.54						
1992		0.8618	1.8204	0.57	0.51				
1993			0.9260	1.8653	0.74				
1994			0.7042	1.0943	1.9877				
1996						2.0875	0.72	0.66	0.63
1997						1.1456	2.0732	0.72	0.66
1998						0.8343	1.0889	2.0162	0.72
1999						0.7060	0.8084	1.0457	1.8932

Notes: Covariances on and below the diagonal; correlations above the diagonal.

Table 10: Covariance/Correlation Matrix for Ln(DER Job Annual Earnings)

	1990	1991	1992	1993	1994	1996	1997	1998	1999
1990	1.9604	0.81							
1991	1.1795	2.0640	0.80						
1992		1.2313	2.0615	0.80	0.74				
1993			1.2486	2.1542	0.79				
1994			1.0129	1.2330	2.1987				
1996						2.2320	0.80	0.75	0.71
1997						1.3094	2.2750	0.80	0.75
1998						1.0805	1.3083	2.3040	0.80
1999						0.9456	1.0736	1.3284	2.2816

Notes: Covariances on and below the diagonal; correlations above the diagonal.

Table 11: Correlation Matrix of SIPP/DER Job Annual Earnings

	Ln(SIPP Job Annual Earnings)								
Ln(DER Job Annual Earnings)	1990	1991	1992	1993	1994	1996	1997	1998	1999
1990	0.74	0.58							
1991	0.59	0.75	0.58						
1992		0.63	0.77	0.61	0.54				
1993			0.71	0.85	0.69				
1994			0.67	0.71	0.86				
1996						0.85	0.70	0.65	0.61
1997						0.68	0.84	0.68	0.63
1998						0.63	0.68	0.84	0.67
1999						0.60	0.63	0.68	0.83

Table 12: Annual Earnings at a job by demographic and education groups

			Average Annual Earnings				Standard Deviation	
		N	SIPP	DER	DIFF	%of SIPP	SIPP	DER
Demo. Group	Education level		(1)	(2)	(3)	(4)	(5)	(6)
white male	no high school	20023	12,203.6	14,217.4	2,013.7	16.5%	12,736.2	15,815.8
	high school	46175	19,952.6	23,468.3	3,515.7	17.6%	17,253.8	21,134.5
	some college	46870	21,351.2	24,887.4	3,536.2	16.6%	22,863.3	31,646.0
	college degree	21658	36,979.3	45,640.2	8,660.9	23.4%	39,136.2	72,316.8
	graduate degree	17993	43,997.9	58,639.8	14,641.8	33.3%	41,487.1	229,519.0
white female	no high school	14815	6,956.4	7,694.5	738.1	10.6%	7,610.6	8,674.6
	high school	45084	12,530.4	13,993.3	1,462.9	11.7%	12,381.0	13,078.5
	some college	50330	13,346.6	14,851.8	1,505.2	11.3%	13,958.1	15,687.4
	college degree	21004	21,395.3	23,980.2	2,584.9	12.1%	20,044.9	23,918.1
	graduate degree	15804	27,348.1	30,833.5	3,485.4	12.7%	27,035.4	41,567.3
non-white male	no high school	3362	10,791.9	13,064.1	2,272.2	21.1%	10,315.6	13,100.0
	high school	6605	15,738.4	18,652.3	2,913.9	18.5%	13,978.3	16,006.2
	some college	6300	17,080.0	19,876.0	2,796.0	16.4%	17,117.0	19,095.0
	college degree	2319	28,925.8	35,250.8	6,324.9	21.9%	26,318.5	61,884.6
	graduate degree	2031	36,679.7	44,314.2	7,634.5	20.8%	31,895.1	54,168.0
non-white female	no high school	3685	7,800.1	8,932.0	1,131.9	14.5%	7,658.3	9,097.2
	high school	7767	11,451.9	13,300.9	1,849.0	16.1%	10,107.4	12,960.8
	some college	8960	13,246.5	15,179.4	1,932.9	14.6%	13,058.7	14,344.3
	college degree	3216	22,125.9	25,625.7	3,499.7	15.8%	19,986.1	21,228.5
	graduate degree	2033	28,447.6	33,606.4	5,158.8	18.1%	22,653.9	26,357.2

an observation is an annual earnings report for a person-job match
individuals may be in this table multiple times if they have multiple jobs
earnings are reported in real 1999 dollar terms

Table 13: Welfare Receipients from 1996 panel - Annual Earnings in 1996 and 1997

				SIPP				DER		
Welfare		N	Earn 1996	Earn 1997	**Change**	**%Change**	Earn 1996	Earn 1997	**Change**	**%Change**
1996	1997	all individuals except single females								
yes	yes	309	9,792.6	10,917.3	**1,124.7**	**11.5%**	12,002.0	14,180.4	**2,178.4**	**18.2%**
yes	no	315	13,129.4	15,932.5	**2,803.1**	**21.4%**	13,681.0	17,002.7	**3,321.7**	**24.3%**
no	yes	170	15,069.1	14,700.2	**-368.9**	**-2.4%**	16,956.7	15,922.8	**-1,034.0**	**-6.1%**
no	no	21413	26,769.8	28,298.0	**1,528.2**	**5.7%**	32,095.6	33,751.3	**1,655.6**	**5.2%**
		single females								
yes	yes	481	6,078.1	7,581.0	**1,502.9**	**24.7%**	6,628.3	7,586.6	**958.3**	**14.5%**
yes	no	203	8,475.5	11,815.9	**3,340.3**	**39.4%**	9,315.9	12,604.7	**3,288.8**	**35.3%**
no	yes	126	9,525.8	7,641.6	**-1,884.3**	**-19.8%**	10,229.2	8,277.2	**-1,952.0**	**-19.1%**
no	no	5724	18,003.3	19,438.9	**1,435.6**	**8.0%**	20,357.8	21,557.0	**1,199.2**	**5.9%**

Welfare		N	Earn 1996	Earn 1997	**Change**	**%Change**	Earn 1996	Earn 1997	**Change**	**%Change**
1996	1997	all individuals except single females								
yes	no	43	none	7,424.0			none	7,733.1		
no	no	1385	none	8,921.3			none	11,535.5		
		single females								
yes	no	30	none	6,011.4			none	6,008.4		
no	no	623	none	4,701.8			none	5,237.8		

Welfare		N	Earn 1996	Earn 1997	**Change**	**%Change**	Earn 1996	Earn 1997	**Change**	**%Change**
1996	1997	all individuals except single females								
no	yes	33	6,493.1	none			8,134.7	none		
no	no	1448	10,085.5	none			13,991.0	none		
		single females								
no	yes	34	3,839.8	none			5,681.3	none		
no	no	459	6,812.1	none			8,853.4	none		

an observation is the sum of earnings from all jobs for an individual
individuals from 1996 panel only
earnings are reported in real 1999 dollar terms
welfare indicator is "yes" if individual was ever on welfare during the year
welfare indicator is "no" if individual was never on welfare during the year

Table 14: Job changers - earnings comparisons at old and new jobs

			SIPP			DER	
	N	Earn Job1	Earn Job2	**Change**	Earn Job1	Earn Job2	**Change**
OVERALL	26241	8265.063	8306.522	**41.45894**	9685.1804	8768.884	**-916.2965**
Respondent reported Industry for JOB1							
Agriculture	481	4,687.2	5,399.1	**711.9**	4,736.1	5,242.6	**506.4**
Mining	121	14,812.0	14,464.0	**-348.1**	19,362.7	14,922.9	**-4,439.8**
Construction	1487	10,147.2	10,158.6	**11.4**	10,353.6	9,982.8	**-370.8**
Manufacturing Nondurable	1457	10,383.2	9,625.3	**-758.0**	12,982.5	10,373.5	**-2,609.0**
Manufacturing Durable	2036	12,724.0	11,024.2	**-1,699.9**	16,349.8	12,117.4	**-4,232.3**
Transp., Comm., Public Ut.	1104	11,744.5	11,035.5	**-709.0**	14,142.9	11,777.3	**-2,365.6**
Wholesale Trade	936	12,144.8	11,535.7	**-609.1**	14,277.0	12,079.6	**-2,197.4**
Retail Trade	8154	4,446.5	5,063.0	**616.5**	5,225.9	5,219.2	**-6.6**
Finance, Insur., Real Estate	1270	13,682.7	13,248.9	**-433.8**	16,719.9	14,811.3	**-1,908.6**
Business & Repair Services	2212	8,064.1	9,022.0	**957.9**	8,701.1	9,337.2	**636.2**
Personal Services	886	5,337.7	5,381.0	**43.3**	5,490.0	5,243.0	**-246.9**
Entertain. & Recreation Ser.	666	5,137.4	6,078.3	**941.0**	5,580.0	6,585.9	**1,005.9**
Professional Services	4798	9,336.4	9,498.9	**162.5**	10,711.8	10,050.7	**-661.2**
Public Administration	555	11,996.7	10,542.2	**-1,454.4**	14,671.6	12,224.1	**-2,447.5**
Respondent reported Industry for JOB2							
Agriculture	422	5,146.1	5,312.4	**166.3**	5,339.0	5,053.0	**-286.0**
Mining	115	16,481.9	16,398.9	**-83.0**	18,909.6	17,564.1	**-1,345.5**
Construction	1624	9,673.9	10,127.0	**453.1**	10,193.3	9,947.5	**-245.8**
Manufacturing Nondurable	1349	9,334.8	9,996.7	**661.9**	11,088.6	11,265.4	**176.8**
Manufacturing Durable	2039	11,455.9	12,120.0	**664.1**	14,138.8	13,313.5	**-825.2**
Transp., Comm., Public Ut.	1279	10,069.4	10,575.2	**505.7**	11,912.6	11,302.8	**-609.8**
Wholesale Trade	953	10,854.2	12,139.6	**1,285.4**	13,193.9	13,058.4	**-135.5**
Retail Trade	7307	4,921.3	4,456.6	**-464.8**	5,895.1	4,584.9	**-1,310.2**
Finance, Insur., Real Estate	1354	12,220.5	13,794.5	**1,574.0**	14,667.6	15,300.2	**632.6**
Business & Repair Services	2646	9,051.7	8,495.1	**-556.6**	10,627.5	8,432.2	**-2,195.3**
Personal Services	818	5,657.2	5,154.6	**-502.5**	6,002.3	5,537.1	**-465.2**
Entertain. & Recreation Ser.	620	6,034.1	5,963.6	**-70.6**	6,442.3	6,124.7	**-317.6**
Professional Services	5114	9,217.4	8,936.0	**-281.4**	10,659.0	9,516.0	**-1,143.1**
Public Administration	582	9,865.6	11,335.4	**1,469.7**	11,692.4	11,818.2	**125.7**
Did respondent switch Industries between JOB1 and JOB2?							
no switch	11494	9,348.5	9,416.3	**67.8**	11,041.5	10,166.4	**-875.0**
switch	14656	7,371.2	7,430.2	**59.0**	8,555.1	7,670.6	**-884.5**

an observation is an individual who reports two consecutive, non-overlapping jobs during the SIPP survey

only one job-switch per person is included in the table

earnings are last annual earnings for job1 and first annual earnings for job2

earnings are reported in real 1999 dollar terms

Table 15: Earnings Regressions Comparisons

N		1,641,180	1,617,320		
		SIPP	DER	SIPP	DER
FIXED EFFECTS		Coefficients		Standard Errors	
non-white female	lf exp years 1-2	0.067	0.009	0.0170	0.0287
	lf exp years 3-5	0.044	0.068	0.0072	0.0119
	lf exp years 6-10	0.028	0.030	0.0037	0.0059
	lf exp years 11-25	0.011	0.014	0.0011	0.0018
	lf exp years 25+	-0.004	-0.001	0.0011	0.0018
white female	lf exp years 1-2	0.077	0.053	0.0071	0.0119
	lf exp years 3-5	0.049	0.053	0.0031	0.0051
	lf exp years 6-10	0.036	0.034	0.0016	0.0026
	lf exp years 11-25	0.011	0.015	0.0005	0.0008
	lf exp years 25+	-0.008	-0.005	0.0005	0.0007
non-white male	lf exp years 1-2	0.100	0.036	0.0191	0.0323
	lf exp years 3-5	0.055	0.076	0.0082	0.0134
	lf exp years 6-10	0.038	0.045	0.0042	0.0067
	lf exp years 11-25	0.012	0.017	0.0013	0.0020
	lf exp years 25+	-0.006	-0.004	0.0011	0.0017
white male	lf exp years 1-2	0.118	0.063	0.0074	0.0125
	lf exp years 3-5	0.062	0.086	0.0033	0.0054
	lf exp years 6-10	0.053	0.054	0.0017	0.0027
	lf exp years 11-25	0.019	0.022	0.0005	0.0008
	lf exp years 25+	-0.010	-0.008	0.0004	0.0006
non-white female	high school	0.124	0.158	0.0229	0.0145
	some college	0.276	0.293	0.0225	0.0143
	college degree	0.623	0.645	0.0287	0.0183
	graduate degree	0.790	0.863	0.0320	0.0205
white female	high school	0.164	0.188	0.0111	0.0070
	some college	0.288	0.298	0.0110	0.0069
	college degree	0.596	0.617	0.0131	0.0083
	graduate degree	0.748	0.767	0.0139	0.0088
non-white male	high school	0.175	0.137	0.0244	0.0154
	some college	0.259	0.241	0.0248	0.0156
	college degree	0.568	0.551	0.0319	0.0203
	graduate degree	0.760	0.728	0.0327	0.0208
white male	high school	0.163	0.176	0.0101	0.0064
	some college	0.247	0.248	0.0101	0.0064
	college degree	0.561	0.586	0.0120	0.0076
	graduate degree	0.677	0.700	0.0125	0.0079
non-white female		-0.005	-0.008	0.0518	0.0303
white female		-0.029	-0.030	0.0287	0.0169
non-white male		-0.001	0.016	0.0569	0.0333
linear time effect		0.021	0.033	0.0014	0.0008
panel1993		0.083	0.139	0.0082	0.0051
panel1992		0.100	0.173	0.0083	0.0051
panel1991		0.130	0.233	0.0106	0.0065
panel1990		0.151	0.253	0.0110	0.0067
intercept		1.266	1.250	0.0235	0.0138
RANDOM EFFECTS					
person intercept		0.180	0.540		
person slope		0.076	0.180		
correlation		-0.099	-0.270		
firm effect		0.069	0.210		
variance of residual		0.160	0.400		
AR1 correlation coefficient of residual		0.500	0.450		

dependent variable is
annualized log real wage

use all observations for
each dominant employer

dominant employer is
highest paying employer
in at least one year

Table 16: Comparison of unobservable heterogeneity: random effects

person effect	
main	0.2814
SIPPdev	0.0000
DERdev	0.0367
firm effect	
main	0.3163
SIPPdev	0.0000
DERdev	0.1013
common time period	
variance	0.7486
AR1 correlation	0.5737
residual	
SIPP variance	0.2171
DER variance	0.3753
SIPP AR1 correlation	0.2705
DER AR1 correlation	0.6365
reliability ratio	
SIPP	0.8612
DER1	0.7240
DER2	0.7982

Figure 1
Labor Force Experience Profiles, Job Level:
White Men

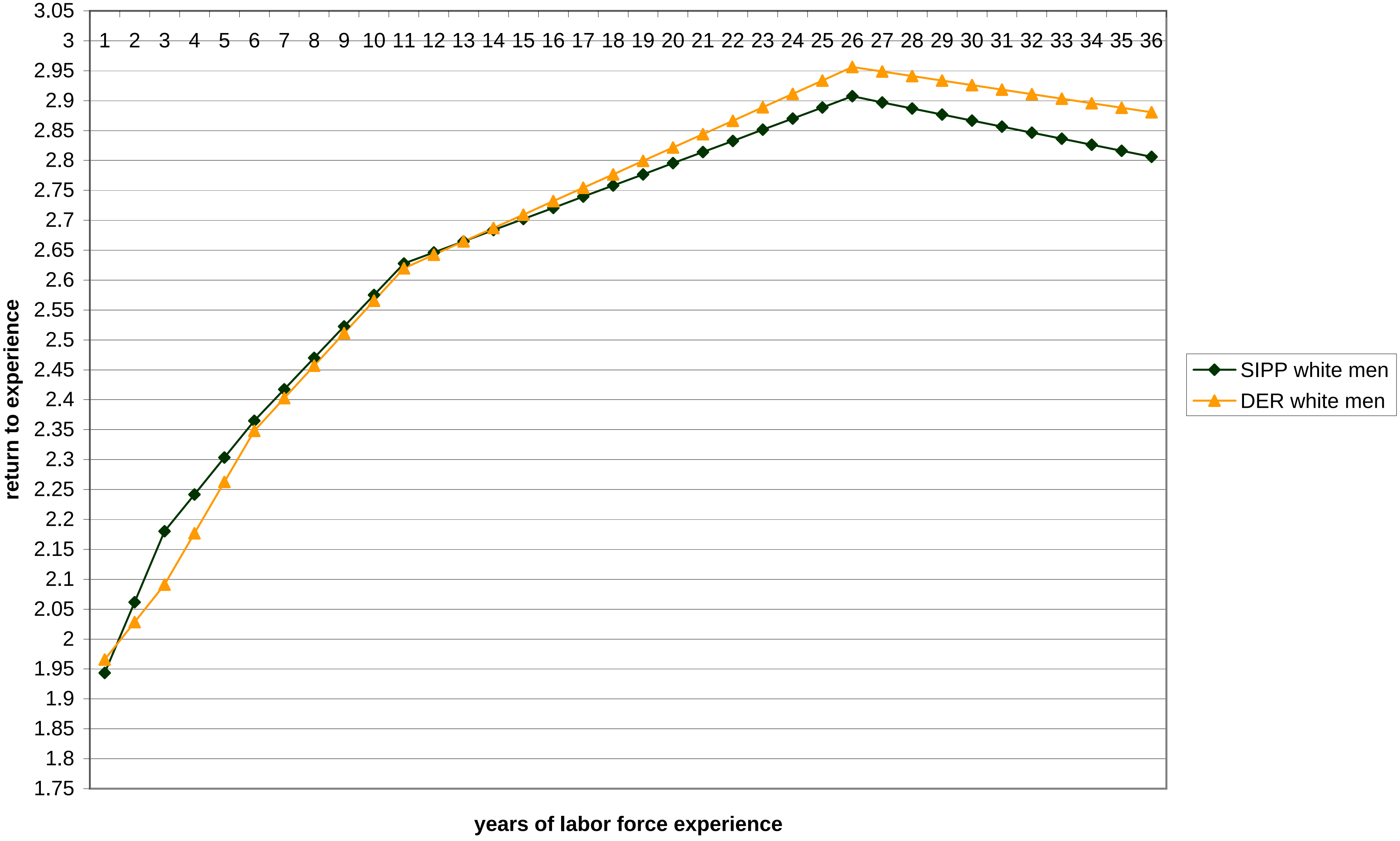


Figure 2
Labor Force Experience Profiles, Job Level: White Women

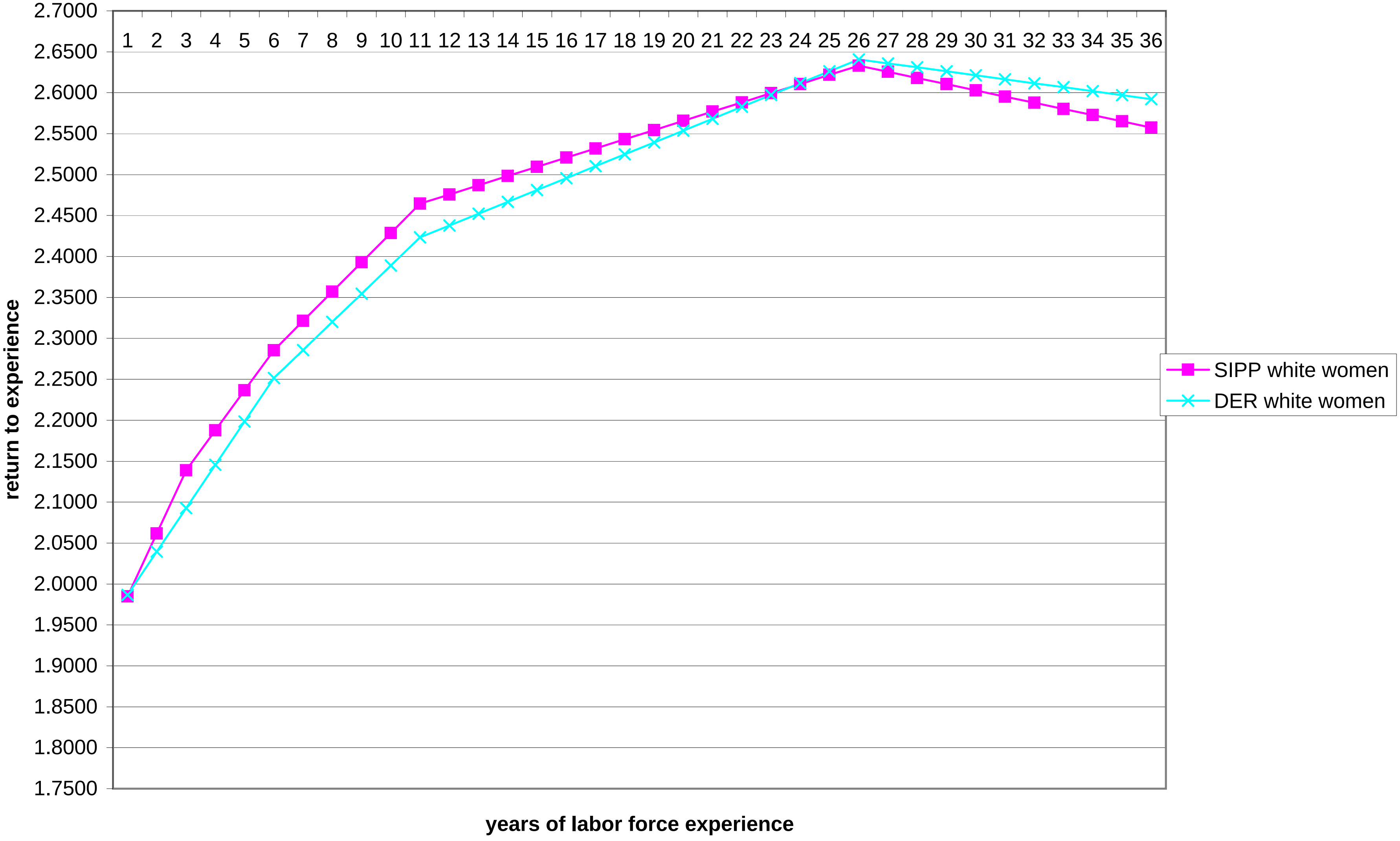


Figure 3
Labor Force Experience Profiles, Job Level:
Non-white Men

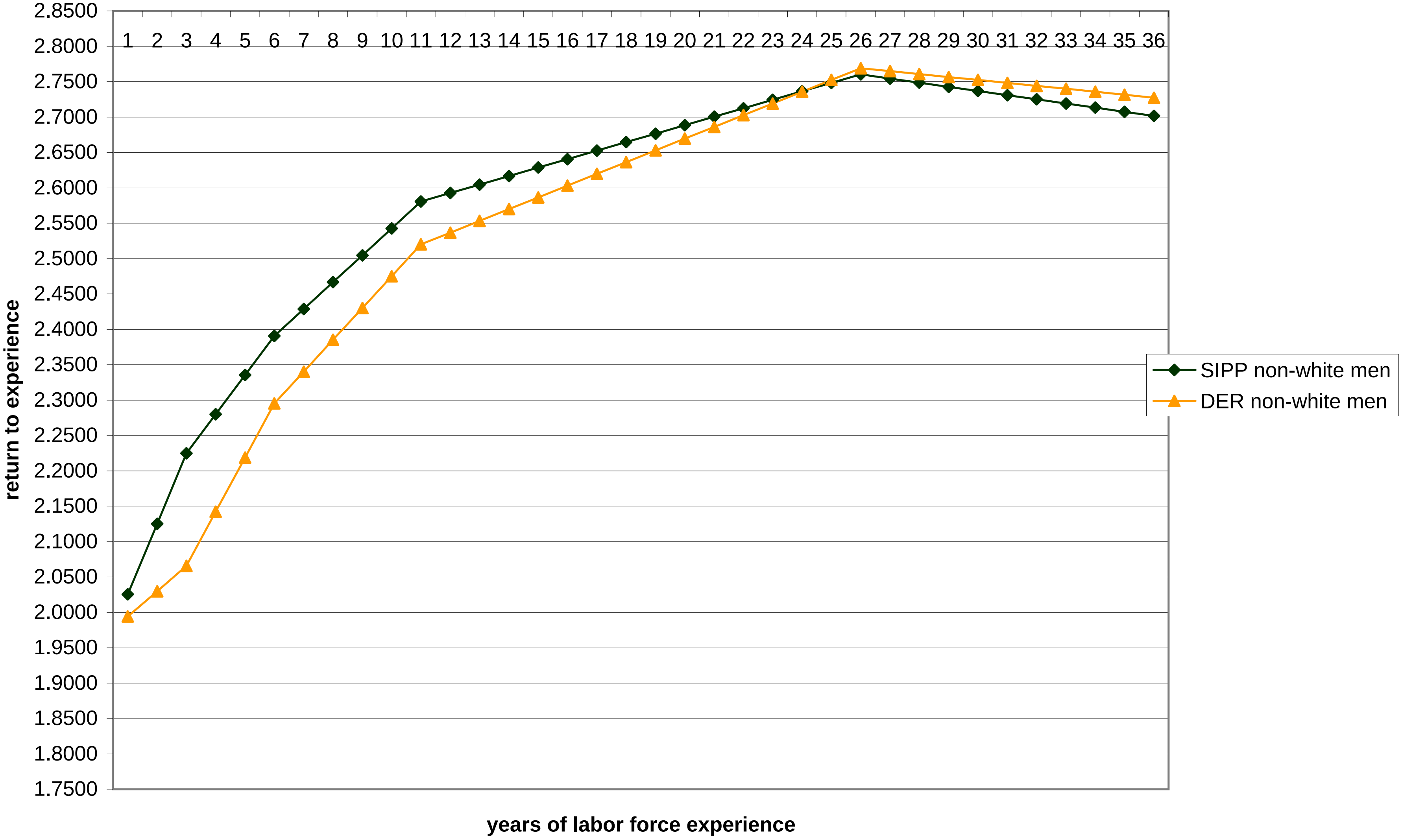
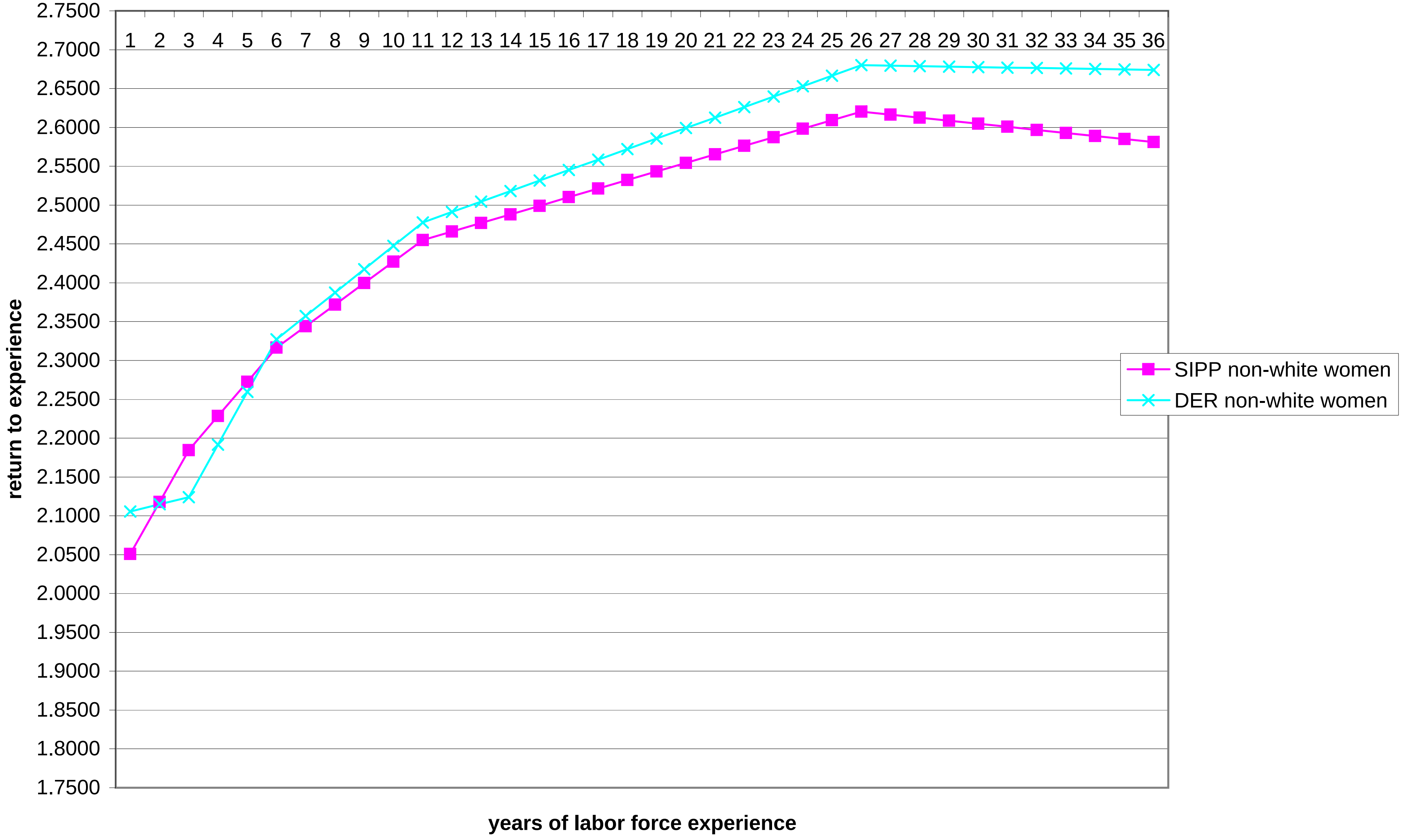


Figure 4
Labor Force Experience Profiles, Job Level:
Non-white Women



Appendix Table A1:
First Round of Job Name Matching, SIPP Panels 1990-1993
Description of Person-Job-Wave observation matching by pass

	Blocking Variables	Matching Variables	m, u prob.	cutoffs
Pass 1	Person ID	Full employer name*	.9, .02	2, .3
Pass 2	Person ID	Fields from employer name*:		2, .3
		word one	.9, .15	
		word two	.95, .6	
		word three	.95, .6	
		word four	.95, .6	
		qualifier word one	.95, .6	
		qualifier word two	.95, .6	
		type word one	.95, .6	
		type word two	.95, .6	
		SIPP original job id number	.6, .5	
Pass 3	Person ID	Fields from employer name*:		5, .05
		Array: first 4 words	.95, .6	
		word one	.9, .15	
		qualifier word one	.95, .6	
		qualifier word two	.95, .6	
		type word one	.95, .6	
		type word two	.95, .6	

*Jobs with missing names were excluded from this round of name matching

Appendix Table A1 (continued):
Second Round of Job Name Matching, SIPP Panels 1990-1993
Description of Person-Job-Wave observation matching by pass

	Blocking Variables	Matching Variables	m, u prob.	cutoffs
Pass 1	Person ID	Fields from employer name:		.1, .1
		full name**		
		array: first 4 words	.9, .1	
		Array: first 2 qualifier words	.9, .15	
		Array: first 2 type words	.9, .15	
		Geo word	.7, .5	
Pass 2	Person ID	Full employer name**	.95, .1	.1, .1

Job Name Matching, SIPP Panel 1996
Description of Person-Job record matching by pass***

	Blocking Variables	Matching Variables	m, u prob.	cutoffs
Pass 1	Person ID	Full employer name**		2, .15
Pass 2	Person ID	Fields from employer name:		2, .15
		full name**		
		word one	.9, .15	
		word two	.95, .6	
		word three	.95, .6	
		word four	.95, .6	
		qualifier word one	.95, .6	
		qualifier word two	.95, .6	
		type word one	.95, .6	
		type word two	.95, .6	
		geo word	.7, .5	
Pass 3	Person ID	Fields from employer name:		2, .15
		full name**		
		Array: first 4 words	.95, .6	
		word one	.9, .15	
		qualifier word one	.95, .6	
		qualifier word two	.95, .6	
		type word one	.95, .6	
		type word two	.95, .6	
		geo word	.7, .5	

**When no weights were assigned, complete employer name was included but given zero weight unless it was blank and then the full disagreement weight was assigned. This was used to prevent jobs with blank names from matching. If weights were assigned, full disagreement weight was also assigned if name was missing.
***Job records with observations in the same wave were disqualified from matching to each other because the same job could not be reported on twice in the same wave.

Appendix Table B1: DER Match to the Business Register

SIPP Panel		DER Total	Match to					
			Business Register		Single-Unit File		Multi-Unit File	
1990	EINs	60,131	58,991	98.10%	58,255	96.88%	16,990	28.25%
	Jobs	96,086	93,520	97.33%	92,379	96.14%	37,807	39.35%
1991	EINs	38,628	38,096	98.62%	37,686	97.56%	12,497	32.35%
	Jobs	58,020	56,725	97.77%	56,118	96.72%	24,526	42.27%
1992	EINs	62,406	61,391	98.37%	60,777	97.39%	19,361	31.02%
	Jobs	99,524	96,982	97.45%	96,029	96.49%	43,197	43.40%
1993	EINs	51,880	50,839	97.99%	50,376	97.10%	17,029	32.82%
	Jobs	81,320	78,933	97.06%	78,198	96.16%	35,977	44.24%
1996	EINs	105,095	95,122	90.51%	94,438	89.86%	28,923	27.52%
	Jobs	192,720	172,832	89.68%	171,585	89.03%	82,546	42.83%

Appendix Table C1:
Description of SIPP Job to DER Job Matching Algorithm by Pass

	Blocking Variables	Matching Variables	m, u prob.	cutoffs
Pass 1	Person ID	Fields from SU name:		2, .3
		Array: first 4 words	.95, .1	
		Array: first 2 qualifier words	.9, .3	
		Array: first 2 type words	.9, .3	
		Geo word	.7, .5	
		year indicators*	.75, .3	
		Complete SU name**		
Pass 2	Person ID	Fields from MU name:		2, .3
		Array: first 4 words	.95, .1	
		Array: first 2 qualifier words	.9, .3	
		Array: first 2 type words	.9, .3	
		Geo word	.7, .5	
		year indicators*	.75, .3	
		Complete MU name**		
Pass 3	Person ID 3-digit SU Industry	year indicators*	.9, .3	2, .3
Pass 4	Person ID 3-digit MU Industry	year indicators*	.9, .3	2, .3
Pass 5	Person ID 3-digit SU Industry	start year***	.9, .3	2, .3
Pass 6	Person ID 1-digit SU Industry	year indicators*	.9, .3	2, .3
Pass 7	Person ID	year indicators*	.9, .1	2, .3
		3-digit SU Industry	.9, .1	2, .3

*Year Indicators by Panel
1990: 1990, 1991, 1992
1991: 1991, 1992, 1993
1992: 1992, 1993, 1994, 1995
1993: 1993, 1994, 1995
1996: 1996, 1997, 1998, 1999

**Complete employer name was included but given zero weight unless it was blank and then the full disagreement weight was assigned. This was used to prevent jobs with blank names from matching in the first 2 passes.

***Start year was first year during survey time frame when job was observed in the SIPP or DER.

Appendix Table C2: SIPP Jobs matched to DER Jobs

		1990		1991		1992		1993		1996	
		SIPP		SIPP	DER	SIPP	DER	SIPP	DER	SIPP	DER
		Jobs	DER Jobs	Jobs	Jobs	Jobs	Jobs	Jobs	Jobs	Jobs	Jobs
1	Master Match	37,207	37,207	22,513	22,513	34,593	34,593	32,289	32,289	66,366	66,366
2	Clerical Match	4,678	4,678	2,745	2,745	5,136	5,136	4,180	4,180	8,476	8,476
3	Duplicate Match on SIPP side	529		294		490		436		971	
4	Duplicate Match on DER side		907		609		1,007		771		1,920
5	Total Matches	42,414	42,792	25,552	25,867	40,219	40,736	36,905	37,240	75,813	76,762
6	Match Rate	76.99%	48.45%	78.75%	48.99%	77.87%	45.08%	77.33%	50.11%	78.04%	44.21%
7	Percent of Matches that are Master	87.72%	86.95%	88.11%	87.03%	86.01%	84.92%	87.49%	86.71%	87.54%	86.46%
8	Residual Job (non-match)	12,673	45,532	6,895	26,930	11,431	49,624	10,818	37,077	21,336	96,861
9	Total Jobs	55,087	88,324	32,447	52,797	51,650	90,360	47,723	74,317	97,149	173,623

Appendix Table C3: Example of Duplicate Match on SIPP Side

	DER	SIPP
Type of Match	EIN	Jobnum
Master Match	A	1
Duplicate Match on SIPP side	A	2

Appendix Table C4: Example of Duplicate Match on DER Side

	DER	SIPP
Type of Match	EIN	Jobnum
Master Match	A	1
Duplicate Match on DER side	B	1