THE WORKER-ESTABLISHMENT CHARACTERISTICS DATABASE

By

Kenneth R. Troske*
Center for Economic Studies
U.S. Census Bureau
Email: ktroske@census.gov
(301) 457-1839

CES 95-10 June 1995

Editor, Discussion Papers, Economic Planning and Coordination, Center for Economic Studies, Room 1587, FB 3, Bureau of the Census, Washington, DC 20233-6101, (301-457-1882) or INTERNET address snguyen@info.census.gov.

Abstract

A data set combining information on the characteristics of both workers and their employers has long been a grail for labor economists. The reason for this interest is that while a number of theoretical models in labor economics stress the importance of employer-employee matching in determining labor market outcomes, almost all empirical work relies on either worker surveys with little information about employers or establishment surveys with little information about workers. The Worker-Establishment Characteristic Database (WECD) represents just such an employer-employee matched database. Containing 199,557 manufacturing workers matched to 16,144 manufacturing establishments, the WECD is the largest worker-firm matched data set available for the U.S. This paper describes how this data set was constructed and assess the usefulness of these data for economic research. In addition, I discuss some of the issues that can be addressed using employer-employee matched data and plans for creating future versions of the WECD.

Keywords: employer, employee, matched, database

*Prepared for the CRIW Labor Statistics Measurement Issues Conference, December 15-16 1994, Washington D.C. I would like to thank Will Carrington, Stacey Cole, Tim Dunne, Brian Greenberg,

Erica Groshen, Robert McGuckin, Nash Monsour, Brian Richards, Richard Sigman, SuZanne Troske and seminar participants at the U.S. Census Bureau for helpful comments. All remaining errors are the responsibility of the author. The opinions expressed herein are solely those of the author and do not reflect the opinions of the U.S. Bureau of the Census.

I. **Introduction**

A data set combining information on the characteristics of both workers and their employers has long been a grail for labor economists. In his article in the *Handbook of Labor Economics* Sherwin Rosen (1986 pg. 688) writes:

> On the empirical side of these questions the greatest potential for future progress rests in developing more suitable sources of data on the nature of selection and matching between workers and firms. Virtually no matched worker-firm records are available for empirical research, but obviously are crucial for the precise measurement of job and personal attributes required for empirical calculations.[1]

The motivation behind the Rosen quote is that existing data sources have proven inadequate for understanding the matching of workers and employers in the labor market. Currently, almost all empirical work in labor economics relies on either worker surveys with little information about the characteristics of a worker's employer, or establishment surveys with little information about the characteristics of workers in the establishment. Obviously, a more complete understanding of how workers and employers sort in the labor market is required before we will begin to understand a number of current empirical puzzles in labor economics such as the rising wage inequality or the establishment-size wage premium. As the Rosen quote makes clear, further understanding of this employer-employee matching process will only come about through the use of employer-employee matched data.

---

[1] In another article in the *Handbook of Labor Economics* Robert Willis (1986) writes:

> Future progress in this area will hinge critically on the development of data which links information on the individual characteristics of workers and their household with data on the firms who employ them.

Employer-employee matched data should also prove useful in a number of other fields in economics. For example, economists interested in estimating production functions at either the aggregate or plant level have long been concerned about the possible biases resulting from treating labor as a unidimentional input in production (Griliches, 1969, 1970). Estimating production functions with employer-employee matched data allows researchers to avoid this problem by enabling them to treat labor as a multidimensional input in the production function.

The Worker-Establishment Characteristics Database (WECD) represents just such an employer-employee matched data set. Containing 199,557 manufacturing workers matched to 16,144 manufacturing establishments, the WECD is the largest worker-firm matched data set available for the U.S. The primary purpose of this paper is to describe the data set and to assess its quality. In addition, I explore some of the issues that can be investigated using worker-firm matched data, and present preliminary plans for creating larger, more representative versions of the WECD.

The WECD is created from two data sources. The first is the Sample Detail File (SDF) which contains all individual responses to the 1990 Decennial Census one-in-six long form. The second is the 1990 Standard Statistical Establishment List (SSEL), which is a complete list of all establishments operating in the U.S. in 1990. The WECD is constructed by using detailed location and industry information available in both data sets to assign an establishment identifier to a subset of manufacturing worker records in the SDF. This identifier in turn enables the worker data to be matched to establishment data available in the Longitudinal Research Database (LRD).[2] Each linked record

[2] The WECD is limited to manufacturing workers and plants for two reasons. First, because preliminary analysis suggested that it would not be possible to match nonmanufacturing employers and employees given the limited

provides both cross-sectional demographic information for workers such as age, sex, race, marital status, and earnings, along with longitudinal information for workers' employers such as the total value of output, cost of materials, investment, and total employment.

I assess the quality of the data in three steps. First, I examine the accuracy of the employer-employee match. Second, I ask whether these data are representative of the underlying population of manufacturing workers and establishments. Third, I examine whether these data replicate results from previous researcher using alternative data sources.

Results from this analysis are somewhat mixed. On the positive side, several facts suggests that most WECD workers are matched to the correct establishments. First, the matching of worker and establishment data produces two estimates of average earnings for each establishment. The average difference between these two estimates is less than five percent and they are positively and significantly correlated. Second, establishments in the WECD have on average 16% of their workforce matched, which is the expected match rate given the sampling frame of the SDF. Another positive finding is that parameter estimates from regressions of wages on worker or plant characteristics are almost identical to results alternative data sets.

On the negative side, only 6% of manufacturing workers in the SDF and 5% of manufacturing plants in the SSEL appear in the WECD, and this match rate varies by industry, plant location, and plant size. In addition, the WECD is not a representative sample of either workers or plants. The WECD contains a larger proportion of white, male, married, production workers than the

---

place-of-work information. Second, because the LRD only contains data for manufacturing plants. The availability of plant data depends on the year. In Census years (all years ending in a 2 or 7) data are available for all plants in existence. However, in all other years data are only available for plants included in the Annual Survey of Manufactures.

SDF, and it also contains a larger proportion of large, old, urban establishments, and establishments located in the Northeastern and Midwestern regions of the country than the SSEL. However, using weights based on the probability that a plant appears in the WECD, one can produce estimates of worker and plant characteristics that are very similar to estimates of these characteristics found using the SDF and SSEL data.

Given that the WECD does not contain a representative sample of workers and employers, and that we only have indirect evidence on whether workers are being matched to the correct establishments, one needs to use these data with caution. As is the case with any new data source, the usefulness of these data can only be established by using these data in empirical research and comparing the results found with these data to those obtained using alternative data sources. Nevertheless, the results from this analysis suggest that the WECD is appropriate for testing hypotheses about relationships between variables derived from theoretical models -- relationships that should hold for any sample of plants or workers, not just a representative sample of these groups.[3] Of course, it must be recognized that results based on these data only apply to a select group of workers and plants and may not generalize to the entire population. However, even with these limitations, these data offer a unique opportunity to examine a number of previously intractable issues.

Apart from the concerns about the representativeness of these data, the primary limitation of the WECD is that they only contains information for manufacturing workers and employers. To try and address both this problem and to make the data more

[3] For example, the competitive model of wage determination says that a worker's wage should equal the worker's marginal product. This should be true for all workers--not just a representative sample of workers. Therefore, we should be able to test this hypothesis using any available sample of workers. However, to conclude that this theory holds for all workers in the labor market we would need to test this hypothesis on a random sample of workers.

representative, future versions of the WECD will be created from data with much more detailed place-of-work information. The Census Bureau is currently planning on replacing the long form from the Decennial Census with a large, monthly, household survey, the Continuous Measurement (CM) survey. Because the CM data will have much more detailed workplace name and address information, it should be possible to can create larger, more representative versions of the WECD that contain workers and employers from all sectors of the economy.

The rest of the paper proceeds as follows. Section II discusses the data sets used to match workers to establishments, and outlines the matching process. Section III investigates the accuracy of the match. Section IV presents examples of how these data can be used in empirical work to increase our understanding of the wage determination process. Section V summarizes and section VI present preliminary plans for creating new versions of the WECD

II. **The Data and Matching Algorithm**

A. *The Data*

Matching workers to establishments is based on detailed location and industry information available for both groups. Information on the location and industry of a worker's employer comes from two questions asked on the one-in-six long form of the 1990 Decennial Census:[4, 5]

---

[4] For a more complete discussion of data available from the 1990 Decennial Census, along with a copy of the long form, see the "1990 Census of Population and Housing-Guide Part A. Text" U.S. Bureau of the Census (1992a).

[5] The form is referred to as the one-in-six long form because it is sent to one in six households on average. However, this rate varies by location. In places with less than 2500 people a form was sent to one-in-two households, while in tracts with more than 2500 housing units it was sent to one-in-eight households.

At what location did this person work LAST WEEK?[6]

and

What kind of business or industry was this?

The Census Bureau assigns geographic and industry codes to each person's record in the SDF based on an individual's response to these questions. Using these codes it is possible to assign each respondent to a unique industry-location cell. For this project I selected all respondents who indicated that they worked in manufacturing and worked in the previous week. This file contains approximately 3.18 million individual records.[7]

Each plant record in the 1990 SSEL includes a four-digit Standard Industrial Classification (SIC) code indicating the establishment's primary industry, and geographic codes showing its location.[8] This information enables each plant in the U.S. to be assigned to a unique industry-location cell. For this project all 342,524 manufacturing establishments are selected from the 1990 SSEL.[9]

B. Matching Process

---

[6] One problem with these questions is that they refer to the business where a person worked last week, which is not necessarily a person's primary place of employment. Another problem is that these questions are only relevant if an individual was employed in the previous week.

[7] The estimated manufacturing workforce based on the 1990 Census is 20.5 million, so the SDF sample of 3.18 million represents approximately 16% of the population of manufacturing workers. While over 4.5 million workers indicated they worked in manufacturing, only 3.18 million of these worked in the previous week.

[8] For a more complete description of the SSEL see "The Standard Statistical Establishment List Program" U.S Bureau of the Census (1979).

[9] The entire 1990 SSEL contains approximately 7.04 million nonagricultural establishments, of which, 424,519 manufacturing establishments. However, once I eliminate records for establishments that are closed, duplicate records, records for establishments with zero payroll or employment, and records for nonproduction unit establishments, I am left with 342,524 establishments.

Assigning a unique establishment identifier to worker records proceeds in four steps:

1) Standardize the geographic and industry definitions in the two data sources.
2) Eliminate all establishments that are not unique in an industry-location cell.
3) Assign a unique establishment identifier to the records of all workers located in the same industry-location cell as a unique establishment.
4) Eliminate all matches based on imputed data.

First, I will briefly describe the geographic coding system of the U.S. Bureau of the Census as of 1990.[10] The Census Bureau divides the entire country into a hierarchy of geographic areas and assigns codes to each area. The most aggregate areas are the four Census regions and the nine Census Divisions. For example, the first region is the Northeast region which consists of the New England and Middle Atlantic divisions. The New England division consists of the states of Maine, New Hampshire, Vermont, Massachusetts, Connecticut and Rhode Island. Each state within a division is assigned a unique geographic code as is each county within a state. Thus, each county in the U.S. has a unique state-county code combination. Counties are further divided into incorporated and unincorporated areas, and incorporated areas with a population of over 2500 are assigned a unique place code.[11] Finally, highly populated places are further subdivided, with each separate physical block in a place

[10] For a more complete description of geographic codes see "1990 Census of Population and Housing-Guide Part A. Text" U.S. Bureau of the Census (1992a).

[11] Portions of counties not in a qualifying place are assigned a place code of 9999.

assigned a unique block code.[12] Thus, for addresses located in central cities, the Census Bureau assigns a unique code for the block, place, county, state, division and region of the address.

The first step in matching workers to establishments is to standardize the geographic and industry codes across the two data sources. Originally, only place code information was available for establishments in the 1990 SSEL. I used the Census Bureau's 1990 Address Reference File (ARF) to assign block codes to 36% of the establishments in the 1990 SSEL.[13, 14]

Industry codes must also be standardized since establishments in the 1990 SSEL are classified into industries using the SIC system, while workers in the SDF are classified into industries using Census industry codes. To make the industry data for both workers and establishments compatible, the SIC codes in the 1990 SSEL are converted to Census industry codes using a concordance table.[15]

The second step in the matching process is to eliminate non-unique establishments. This occurs in two steps. First, I keep all establishments that are unique in an industry-block cell. However, because some plants have missing block codes, I only keep establishments that are unique in an industry-block cell

---

[12] In 1990 block codes were only available for addresses in Tape Address Register (TAR) areas. These roughly correspond to central cities or Metropolitan Statistical Areas (MSAs). By 1992 block codes were available for most addresses in the U.S., not just addresses in TARs.

[13] The ARF is a file of address ranges with the corresponding geographic codes. Given a street address one can use the ARF to assign the appropriate geographic codes.

[14] The main reason why establishments in the 1990 SSEL do not have block codes is that in 1990 block code information is only available for establishments located in TARs. Data from the 1990 SSEL shows that 40% of manufacturing establishments are located in an MSA. Thus, I am missing block codes for only 4% of the establishments.

[15] See "1990 Census of Population and Housing--Classified Index of Industries and Occupations" U.S. Bureau of the Census (1992b). SIC codes are converted to census codes because the census codes are more aggregate than SIC codes.

when all establishments in the industry-place cell have valid block codes, or when an establishment is unique in an industry-place cell.[16] Eliminating non-unique establishments reduces the number of establishments available for matching from 342,171 to 63,949.

Next, workers and establishments are assigned to industry-location cells and workers and establishments in the same cell are matched. This is a two step process. First, workers and establishments are assigned to industry-*block* cells and matched. Second, all remaining workers and establishments are assigned to industry-*place* cells and matched.

Finally, to minimize the probability of incorrectly matching workers to establishments, I drop all worker-establishment matches based on imputed industry or geographic data.[17] In addition, I drop all matches where the total number of workers matched to a given establishment is greater than the establishment's reported employment.[18]

---

[16] Multiple establishments that are owned by the same firm that are in the same block or place cell are kept.

[17] For example, if I match a worker to an establishment using block code information, and the worker's block code is imputed, I throw out the match. However, if I match a worker to an establishment using place code information, and the place code is not imputed, I keep the match, regardless of whether the block code is imputed. I chose to eliminate imputed data after I matched workers and establishments to increase the number of successful matches. This way I keep matches based on place codes even when the block codes have been imputed. In the SDF 1,790,851 worker records have imputed block codes, 218,558 have imputed place codes and 157,185 have imputed industry codes. Imputation of these items is done by cold decking. In this process, when information for an individual is missing the computer draws another individual at random from a distribution of individuals with similar characteristics. Then information from the selected record replaces the missing information in the original record. Obviously, using imputed data would increase the number of incorrect matches.

[18] Dropping matches based on imputed geographic or industry codes eliminates 218,507 matches. Dropping matches where the number of workers matched to an establishment is greater than the establishment's reported employment eliminates 17,826 matches. There are a number of possible reasons why I matched more workers to an establishment than the establishment's reported employment. First, a worker's industry or geographic code could be misassigned. Second, an establishment's employment may have changed between March 12th, the date employment is recorded in the SSEL, and April 1st, the

The resulting data set contains 199,557 worker records matched to 16,144 different plants.[19] Table A.1 in the appendix provides a list of variables available for workers in the WECD and for establishments in the LRD.

III. **Evaluating the WECD**

A. *Examining the Accuracy of the Match*

One advantage to using the matching algorithm described above is that coding errors should be the primary reason for incorrectly assigning workers to establishments.[20] The matching algorithm only matches workers to establishments that are unique in an industry-location cell. Therefore, given that workers and establishments have the correct geographic and industry codes, all workers in an industry-location cell which contains an establishment *must* work in that establishment. Furthermore, all workers in the same industry-location cell who filled out the long form in the Census are matched to the same plant. This

---

date of the Census. Third, reported employment in the SSEL does not include the owner of an establishment, while the owner could be in the SDF. Matching the owner to the establishment may make it appear that more workers are matched to an establishment than the establishment's reported employment. These latter two reasons are more likely to be a problem with small establishments.

[19] While the matching algorithm results in 16,144 unique establishment-level identifiers being attached to the 199,557 worker records, detailed information is not available for all of these plants in all years. This is because detailed information on plant inputs and outputs comes from the LRD which consists of the plant-level records contained in the various Census of Manufactures and Annual Survey of Manufactures. Therefore, the number of plants for which detailed data is available depends on the year (in particular, whether a survey or a census was conducted in a year). For example, matching the worker file to 1989 LRD data (a survey year) results in a match of 152,987 worker records to 5,423 establishments. In contrast, matching the worker data to 1987 LRD data (a census year) results in 195,943 worker records matched to 15,557 establishments.

[20] One large source for coding error is assigning an industry code to a worker's description of the primary industry of their employer. Another possible source of error is mismatching workers who work in new establishments that are not yet included in the SSEL to older establishments in the SSEL in the same industry-location cell.

means that the WECD will contain a random sample of workers in the plant.[21]

In spite of these assurances, some tests of the match are desirable. To begin, Table 1 presents statistics examining the quality of the match. One test of whether workers and establishments are correctly matched is to compare similar information from the worker and establishment data. This is done in rows (1)-(4) in Table 1. Row (1) presents the cross-plant mean of worker earnings using data from the SSEL. Per-worker earnings in a plant are estimated by dividing the 1990 annual payroll for the establishment by the plant's March 12, 1990 employment. The numbers in row (1) are an average of this per-worker earnings estimate across all plants in the data. I will refer to this number as SSEL worker earnings. Row (2) presents the cross-plant mean of worker earnings based on worker data. Each worker in the SDF reports their total earnings in the previous year. Per-worker earnings in a plant are estimated by taking the average earnings for all workers matched to the plant. The numbers in row (2) are then the average of this per-worker earnings estimate across all plants in the data. I will refer to this number as SDF worker earnings. Row (3) presents the cross-plant mean log difference in these two estimates of worker earnings, while row (4) presents the cross-plant correlation of these two estimates of worker earnings. Row (5) presents the cross-plant mean of total employment in the plants (based on SSEL data), while row (6) presents the average proportion of workers matched to the plant. Column (1) in Table 1 presents numbers for all plants and workers in the WECD, column (2) presents numbers for workers and plants that contain workers who are between 18 and 65 years old and who usually worked between 30 and 65 hours a

[21] This assumes that there is no systematic bias in response rates to the long-form. Seen Bates, Fay and Moore (1991) and Kulka, et.al (1991) for a discussion of response rates to the 1990 Decennial Census.

week in 1989, while column (3) presents numbers for plants with more than 10% of the workforce matched to the plant.

The numbers in Table 1 suggest that workers are matched to the correct establishments. The numbers in rows (1) and (2) show that the estimates of worker earnings from the SSEL and SDF data are very similar. The numbers in row (3) show that for all plants and workers in the data the average plant-level difference in the two estimates is less than 5%.[22] Further, when we consider the samples in columns (2) and (3) this difference falls to less than 1% and is statistically insignificant. The numbers in row (4) show that the SSEL and SDF worker earnings are positively and significantly correlated.[23] Finally, row (6) shows that on average 16% of a plant's workforce is matched to the plant. This is the exact rate one would expect given the one-in-six sampling frame of the SDF.

Table 2 breaks out the numbers in Table 1, first by the size of the plant (Panel A), and second by the nine Census Divisions (Panel B). The numbers in Table 2 are for workers who are

---

[22] There are a number of reasons why these two estimates might differ. First, the estimates of earnings per-worker based on plant data is an estimate of the earnings paid to a worker by the plant, while the estimates based on worker data is the total earnings paid to a worker by all employers. If some workers in a plant hold multiple jobs then the estimate based on worker data will be larger. Second, worker earnings reflect total earnings of a worker in 1989, while the estimate based on plant data is the total amount paid in salary and wages by the plant to all workers in 1990 divided by the number of workers in the plant on March 12, 1990. If a worker is on vacation on March 12th then the worker will not appear in the employment figure, however any paid vacation they receive will appear in the wages paid by the plant. This will tend to make SSEL worker earnings larger than SDF worker earnings. Also, if employment in the plants is seasonal, and March 12th is a period of low (high) employment then SSEL earnings will appear higher (lower) than SDF earnings. Finally, part-time workers may be missed in the plant earnings estimate, but have a major impact on the estimate of earnings based on worker data.

[23] The reader should note that, because the SDF earnings estimates are based on a sample of workers in a plant, even if all workers are matched to the correct establishment the estimate of D will in general be less than 1 because of sampling error. Thus, the fact that these correlations are significantly greater than zero is fairly strong evidence that workers are being matched to the correct establishments.

between 18 and 65 years old and who usually worked between 30 and 65 hours a week in the previous year.[24]

The numbers in panel A reveal no systematic relationship between the difference in SSEL and SDF worker earnings and plant size. The largest difference, 14%, is found for plants with 1-9 employees, while the smallest difference, 0.1%, is found for plants with 10-24 employees. However, there is a strong negative relationship between plant size and the proportion of workers matched to the establishment, and a strong positive relationship between plant size and the correlation of the two measures of worker earnings. Plants with 1-9 employees average 40% of their workforce matched to the plant. However, the correlation between SSEL and SDF worker earnings in these plants is only 0.20. In contrast, plants with over 1000 workers average 8% of their work forced matched to the plant, while the correlation between the two earnings measures is 0.78. The negative relationship between the proportion of workers matched and size is the result of an integer constraint. Plants must have at least one worker matched to the plant to appear in the data. For a plant with 5 employees this means that the minimum percent matched will be 20%. Obviously, as a plant gets larger, this minimum approaches zero. The reason that the correlation between the two measures of worker wages increases with plant size is that as the size of a population increases it requires a smaller percentage of the population to have a representative sample. Thus, in plants with more than 1000 employees we are able to get a relatively accurate estimate of worker wages with only 8% of the workforce. Overall, while it appears that smaller plants have a much larger

[24] I focus on these workers for three reasons. First, because these are workers with the strongest labor market attachments and therefore should have the most reliable earnings and hours worked data. Second,because the log difference across plants (row 3 - Table 1) is small and insignificant for these workers. Third, because of the problem with part-time workers mentioned in footnote 22.

proportion of their workforce matched, larger plants appear to have a much more representative sample of workers matched.

The numbers in Panel B show no systematic relationship between the difference in the two earnings measures and plant location. While the mean difference in the two earnings measures varies between -0.037 and 0.032, this difference is never significantly different from zero for plants in any Census Division. In addition, there is very little variation in either the proportion matched or in the correlation between the two earnings measures across plants in the various Census Divisions. The numbers in Panel B suggest that the matching process works equally well for plants in all areas of the country.

Table 3 breaks out the numbers presented in Table 1 by two-digit industry again for workers between 18 and 65 years old who usually worked between 30 and 65 hours a week in the previous year. Column (3) in Table 3 shows that the log difference in the measures of worker earnings varies from a high of 0.24 for Tobacco to a low of -0.13 for Chemicals. However, of the 20 two-digit industries, 12 have an absolute difference of less than 0.05 and in 13 industries the difference is not significantly different from zero at the 1% significance level. Further, in all 20 industries there is a positive correlation between these two measures of workers earnings, and in 18 of the 20 industries the correlation is significantly different from zero at the 0.1% significance level. Viewed as a whole the numbers is Tables 1-3 suggest that workers are being matched to the correct establishments.

B. *Examining the Representativeness of the WECD*

To begin examining whether the WECD data are representative of the underlying population of workers and plants, Table 4 compares the number and annual earnings of workers in the SDF with workers in the WECD, for all workers (the Total row) and by

two-digit industry. Columns (1) and (2) present the number of workers in the SDF and WECD, respectively, while column (3) presents the proportion of workers in the industry matched to an establishment (column 2 ÷ column 1). Columns (4) and (5) present the industry mean of worker earnings in the SDF and WECD, respectively, while column (6) presents the cross-plant log difference in average worker earnings.

The Total row in Table 4 shows that of the 3,176,986 manufacturing workers in the SDF, 199,558 appear in the WECD, a match rate of 6%. The numbers in column 3 show that this match rate varies by industry. Tobacco, Paper, Leather, and Fabricated Metals, all have match rates of 10% or greater, while Lumber, Instruments, and Miscellaneous all have match rates of 3%. The numbers in column 6 show that matched workers average 10% higher wages than all SDF workers, but that the size and sign of this difference varies by industry. In three two-digit industries matched workers average lower wages than workers in the SDF. In 15 two-digit industries the absolute difference in earnings is less than 10%.

Table 5 presents the number and average employment for SSEL plants, unique plants, and WECD plants, for all plants in the data (the Total row) and by two-digit industry. Unique plants are plants that are unique in an industry-location cell. As mentioned earlier, only plants that are unique in an industry-location cell are matched to workers. Plants with workers matched to them are WECD plants. Columns (1)-(3) present the number of SSEL plants, unique plants, and WECD plants, respectively. Column (4) presents the proportion of plants that are unique (column 2 ÷ column 1), while column 5 presents the proportion of plants in the WECD (column 3 ÷ column 1). Columns (6)-(8) present the mean employment for SSEL plants, Unique plants, and WECD plants, respectively.

The Total row in Table 5 shows that of the 342,471 plants in the 1990 SSEL, 16,144 appear in the WECD, a match rate of 5%. This is almost identical to the match rate for workers. The numbers in column (5) show that this rate varies considerably across two-digit industries in a similar manner to the pattern seen in Table 4. Tobacco, Paper, Leather, and Fabricated Metals have the highest match rates while Lumber, Instruments, and Miscellaneous have the lowest.

The numbers in column (4) show that being unique in an industry-location cell does not guarantee that a plant appears in the final data. Overall, almost 20% of plants in the SSEL are unique, but only 5% appear in the WECD. The numbers in columns (6)-(8) show why this is the case. Comparing the average employment of unique plants with the average employment of SSEL plants shows that unique plants are much larger than SSEL plants. This is because it is much more likely that a large plant will be unique in an industry-location cell. Comparing the average employment of unique plants with the average employment of WECD plants shows that WECD plants are even larger than unique plants. This is the result of the sampling scheme of the Decennial Census long form. Since this form was sent to one-in-six households on average it is much more likely that a large establishment will contain a worker who received the form, and therefore, more likely that a large establishment will appear in the WECD.

The fact that WECD plants are larger than SSEL plants also explains why WECD workers have higher average wages than SDF workers. Previous research has found a positive correlation between plant size and worker wages (Brown and Medoff, 1989; Troske, 1994). Since WECD workers work in larger establishments than SDF workers they will in turn have higher average earnings.

Table 6 repeats the same analysis for workers found in Table 4 this time broken out by Census Division. One thing to notice in Table 6 is that the match rate is significantly lower in the

Mountain and Pacific Divisions. In the Pacific division only 2% of the workers in the SDF are matched to plants.

Table 7 repeats the same analysis for plants found in Table 5 this time broken out by plant size (Panel A) and Census Division (Panel B). The numbers in Panel A of Table 7 confirm the fact that large plants are both more likely to be unique and more likely to appear in the WECD. The Proportion Unique column shows that as plant size increases the probability that a plant is unique in a industry-location cell rises, from 0.15 for plants with 1-9 employees to 0.39 for plants with 1000 or more employees. However, the Proportion Matched column shows an even greater increase with size, rising from 0.02 in the smallest plants to 0.20 in the largest plants. In fact, the probability that a plant appears in the WECD, conditional on the plant being unique, rises from 0.12 for plants with 1-9 employees to 0.51 for plants with 1000 or more employees (not in table).[25]

Similar to Table 6, the numbers in Panel B show that the match rate for plants is significantly lower in the Mountain and Pacific divisions. While part of this is because plants in these divisions are less likely to be unique, this is not a complete explanation. Even conditional on being unique, plants in the Mountain and Pacific divisions are much less likely to appear in the WECD. The figures in columns (6)-(8) suggest one explanation for why this is the case. Plants in these divisions are smaller on average than plants in other divisions. As is shown in Panel A, small plants are not only less likely to be unique, they are also less likely to include workers who received a one-in-six long form in the Decennial Census.[26]

---

[25] This is computed as (WECD Plants ÷ Unique Plants).

[26] An alternative explanation could be that workers in these divisions are more likely to have imputed industry and location information. However, this is not the case. In fact, workers in the Mountain division are less likely to have imputed data than workers in the other divisions.

Tables 4-7 show that the success of the matching procedure varies by the industry and location of plants and workers, and by the size of the plant. Since the characteristics of workers and plants are not distributed randomly across industry, location, and plant size, this affects the representativeness of the WECD. In addition, work at the Census Bureau and elsewhere (Bates, Fay and Moore, 1991, Kulka, et.al, 1991) show that the probability that a household responded to the 1990 Decennial Census was correlated with the income and race of the household, the age and education of the head of the household, and whether the household contained related persons. Since the WECD only contains workers with non-imputed data this will also impact the representativeness of the WECD data.

These effects can be seen in Table 8 and Figure 1. Table 8 presents characteristics for all manufacturing workers in the SDF (column 1), for all manufacturing workers in the May 1988 Current Population Survey (column 2), and for all WECD workers (column 3). Figure 1 presents the educational distribution for SDF and WECD workers.[27] The numbers in Table 8 show that workers in the WECD are not a representative sample of the entire population of manufacturing workers. A much larger percentage of workers in the WECD are white, male, married, production workers than in either the SDF or CPS. Workers in the WECD are slightly older than workers in the SDF or CPS, and are more likely to be located in the Northeast and Midwest regions of the country. Table 8 also shows that, relative to workers in the SDF or the CPS, workers in the WECD worked more weeks, usually worked more hours per week, and averaged higher earnings and hourly wages. Finally, Figure 1 shows that, relative to workers in the SDF,

[27] Respondents to the CPS report the number of years of education completed. Respondents to the Decennial Census report the highest degree completed. Since these are not completely analogous concepts I do not include CPS workers in Figure 1.

workers in the WECD are more likely to have a high school diploma, and are less likely to have less than a high school diploma, a Bachelor's degree or an advanced degree. All of these results are very similar to the findings of Bates, Fay and Moore (1991), and Kulka, et.al (1991) and are exactly what we would expect given that large plants are over represented in the WECD.

To make estimates of characteristics based on the data in the WECD more closely match estimates of characteristics based on the SDF data, I produce weighted estimates of these characteristics using weights based on the conditional probability that a plant appears in the data. I now turn to discussing how I estimate these weights.

As the discussion in Section II and the results in Tables 1-7 show, the probability that a plant appears in the data is first a function of whether the plant is unique in an industry-location cell, and second a function of whether the plant contains a worker who received and responded to the one-in-six long form in the 1990 Decennial Census. I assume that these two probabilities are independent and estimate the probability of these two events separately. The product of these two probabilities will then be an estimate of the conditional probability that a plant appears in the data.

The probability that a plant is unique is given by:

$$P(u) = X'\beta + u \qquad (1)$$

where P(u) is the probability that a plant is unique in an industry-location cell, **X** is a vector of plant characteristics, and u is a normally distributed random error term. Results from Tables 4-7 show that the probability that a plant is unique is related to plant size, industry and location. Therefore, **X** controls for (the log of) plant employment, two-digit industry, and Census division. In addition, since the geographic detail of a plant's location is related to whether or not the plant is

located in an urban area, **X** includes controls for whether or not the plant is located in a valid place (has a place code other than 9999) along with the total population and the population per square mile for the county where a plant is located.[28] Since I cannot directly observe P(u) but instead only observe $P^*(u)$ where:

$$P^*(u) \begin{cases} 1 \text{ if a plant is unique} \\ 0 \text{ otherwise} \end{cases}$$

equation (1) is estimated using a Probit model. Results from this estimation are available from the author.

The probability that a plant is matched, conditional on being unique, is given by:

$$P(m|u) = Y'\gamma + \varepsilon \qquad (2)$$

where P(m|u) is the probability that, conditional on being unique, a plant appears in the WECD, **Y** is a vector of plant characteristics and , is a normally distributed random error term. The results in Tables 4-7 show that plant size also affects whether a plant contains matched workers. Therefore, (the log of) plant employment is included in **Y**. Since the sampling frame of the SDF varied with the population of an area, **Y** includes controls for the population per square mile and the total population for a plant's county. County level measures of median age, median education of individuals over 25 and its square, the density of nonminority whites, and the density of family households, are also included in **Y** to control for variation in response rates with age, education and household type. To control for the fact that more detailed geographic information is available for workers in urban areas, **Y** includes a control for whether the plant is located in a valid place.

[28] These latter two numbers are based on the 1980 Decennial Census.

Finally, **Y** includes controls for Census Division and two-digit industry. Again, since I do not directly observe P(m|u) but instead observe $P^*(m|u)$ where:

$$P^*(m|u) \begin{cases} 1 & \text{if a plant is matched} \\ 0 & \text{otherwise} \end{cases}$$

equation (2) is estimated using a Probit model. Results from this estimation are available from the author.

Column (4) in Table 8 presents estimates of the characteristics of workers in the WECD weighted by the inverse of the estimated probability that a worker's plant appears in the data. Figure 1 includes the weighted educational distribution for WECD workers. The numbers in Table 8 show that weighted estimates of worker characteristics are much closer to estimates of these characteristics based on the SDF data. The weighted cross-worker mean of age, sex, race, marital status, occupation and location are all much closer to the cross-worker mean of these characteristics found in the SDF. The weighted mean of number of weeks worked, usual hours worked last year, wage or salary income and hourly wage are also much closer to the values found in the SDF. Finally, Figure 1 shows that the weighted educational distribution for WECD workers is quite similar to the educational distribution for SDF workers.

To examine how representative plants in the WECD are of the entire population of plants, Table 9 presents various characteristics for: all manufacturing plants in the SDF (column 1), unique plants (column 2), unique plants weighted by the inverse of the estimated probability of being unique (column 3), all plants in the WECD (column 4), and all WECD plants weighted by the inverse of the estimated probability that they appear in the WECD (column 5). The unweighted numbers show that neither unique nor WECD plants are representative of the entire population of manufacturing plants. As shown in previous tables,

unique plants and WECD plants are much larger and are more likely to be located in the Northeast and Midwest regions. The Plant Age variable shows that a much larger percentage of unique and WECD plants are more than 10 years old, while the place and multi-unit variables show that unique and WECD plants are more likely to be located in a place and to be part of a multi-establishment firm. However, columns (3) and (5) show that the weighted cross-plant means of these characteristics more closely resemble the means for all manufacturing plants in the SSEL.

C. *Replicating Previous Findings*

While the results in Tables 8 and 9 are encouraging, they are in some ways incomplete. Given that the primary use of these data is to study relationships in a regression framework, a more complete test of these data involves examining whether regression results using these data can replicate results found in the original data and results found by previous researchers using alternative data sources. This is what is done in Tables 10 and 11. Table 10 presents the results from regressions of (log) worker wages on a standard set of worker characteristics. Column (1) presents results based on all workers in the SDF controlling for whether a worker is matched to a plant. Column (2) presents the results from the identical regression excluding this control. Column (3) presents the results for the identical regression in column (2) using only data for workers in the WECD, while column (4) presents the results from the same regression where the WECD data are weighted by the estimated probability that a worker appears in the matched data.

The coefficient on the match variable shows that workers matched to plants earn 3% higher wages than nonmatched workers. However, comparing the coefficients on the rest of the variables across the four columns shows that there is almost no difference in the relationship between these characteristics and the wages

of matched and nonmatched workers. The only major difference in the four columns is the relationship between education and wages. The coefficients on the education variables in column (2) show a much stronger relationship between education and wages than the coefficients on education in either columns (3) or (4). However, all four regressions show a very strong positive relationship between education and wages. The most likely explanation for this finding, and the positive coefficient on the Match variable in column (1) is that the matched workers work in larger plants than the nonmatched workers. Results in Troske (1994) show that workers in large establishments earn higher wages and that part of the observed education premium is the result of more educated workers working in larger establishments.

The estimated relationships seen in Table 10 are similar to previously reported relationships between experience, sex, marital status, race, education, and wages (Cain, 1986; Korenman and Neumark, 1991; and Mincer, 1974). For example, the coefficients on Female, Black, Married, Female*Married, and Female*Black show that women earn 17% less than men, black men earn 4-6% less than nonblack men, married men earn 13% more than single men, married women earn about the same as single women, and black women earn about the same as white women.

Table 11 presents the results from regressions of (log) average annual earnings in a plant on various plant characteristics, for all plants in the SSEL (column 1), unique plants (column 2), unique plants weighted by the probability of being unique (column 3), WECD plants (column 4) and WECD plants weighted by the probability of appearing in the WECD (column 5). Analogous to Table 10, the coefficients on the various variables in Table 11 are similar across the five regressions. The major differences occur for the location variables. The coefficient on Place in column (1) is positive while the coefficients on Place in the other four regressions are all negative (although never

significantly different from zero). The coefficients on the three region variables also vary in sign and magnitude across the five regressions (although in all five regression plants in the Northeast region pay the highest wages). The most likely explanation for these differences is that almost all unique plants and WECD plants are located in a place, and very few these plants are located in the West region.

The estimated relationships seen in Table 11 are also similar to previously reported relationships between plant characteristics and average wages. The coefficient on log plant employment shows that large employers pay higher average wages (Brown and Medoff 1989; Dunne and Schmitz 1992), while the coefficients on the plant age variables show that older plants also pay higher wages (Brown and Medoff 1994; Dunne and Roberts, 1990).

The results in Tables 8 and 9 show that while the unweighted data are not a representative sample of either the underlying population of workers and plants, it is possible to use weights based on the probability that a plant appears in the data to produce estimates of characteristics that are similar to estimates from the SDF and SSEL data. Even more encouraging, the results in Tables 10 and 11 show that these data are capable of replicating both the relationships found in data for the underlying population, and the relationships found by previous researchers using alternative data sources. Thus, it appears that these data are useful for addressing certain empirical questions. Just what some of these questions are is what I turn to next.

IV. **What Can We Learn From the WECD?**

A. *The Establishment Size-Wage Premium*[29]

One question that has long interested labor economists is why large employers pay higher wages than small employers -- what is referred to as the employer size-wage premium. Despite this long interest, previous attempts to account for the employer size-wage premium in terms of observable worker or employer characteristics have met with limited success. The reason for this lack of success is that, while most theoretical explanations for the employer size-wage premium stress the matching of workers and employers (eg., Oi 1983, 1990; Hamermesh 1980, 1993; Dunne and Schmitz 1992), previous empirical work has relied on either worker surveys with little information about the characteristics of a worker's employer, or establishment surveys with little information about the characteristics of workers in a plant. Obviously the WECD, which contains information for both workers and employers, is an ideal source for investigating the employer size-wage premium.

One explanation for why large employers pay higher wages is that large employers hire more skilled workers. Two models for why this might be true are the capital-skill complementarity model (e.g., Griliches 1969; Hamermesh 1980, 1993) and the Dunne and Schmitz (1992) model. In the capital-skill complementarity model large plants employ more skilled workers because they also employ more capital. In this model the size of the capital stock in a worker's plant should be positively correlated with wages. In the Dunne and Schmitz model large employers employ more skilled workers because they also employ skill-biased advanced technology capital. This model predicts that the use of advanced technology capital in a worker's plant will be positively correlated with worker wages. Because the WECD contains

[29] For a complete discussion of the issues in this section see Troske (1994).

information on both of these variables it can be used to investigate these hypotheses.

Consider the results in Troske (1994). This paper contains a series of worker wage regressions which include measures for both total capital stock and the amount of advanced technology capital in the plant in addition to total employment. These regressions show that workers who work in plants with higher capital-labor ratios or who work in plants with larger amounts of advanced capital relative to labor, receive higher wages. These results are consistent with both the capital-skill complementarity and the Dunne and Schmitz models for why large employers hire more skilled workers. More importantly, these results also represent the first successful attempt to account for the establishment size-wage premium in terms of worker or employer characteristics.

B. *Wages, Productivity and Worker Characteristics*[30]

Models of wage determination such as: life cycle wage models, models of race or sex discrimination, returns to education, productivity effects of marriage, models of job-specific human capital accumulation, industry rents, etc., all hinge on the relationship between wages, productivity and worker characteristics. However, direct measures of worker productivity are hard to obtain, so economists usually must rely on proxies for worker productivity when conducting empirical research. The difficulty with this approach is that whether these proxies reflect productivity differences is always in doubt, making it difficult to distinguish between competing models. However, data such as the WECD, by combining worker and plant data, avoids these difficulties by allowing researchers to directly compare

[30] For a complete discussion of the analysis discussed in this section see Hellerstein, Neumark, and Troske (1994).

estimates of the relative wages of workers with estimates of workers' relative marginal productivity.

As an example, consider Hellerstein, Neumark and Troske (1994). This paper uses a production function approach, where workers with different characteristics are treated as substitute labor inputs in the plant, to directly estimate the marginal product of workers. These estimates are then compared with estimated wage differentials among groups of workers. This analysis represents a departure from most of the existing empirical literature on wage determination because the authors directly compare estimates of workers' relative wages with estimates of workers' marginal products. Two of the findings from this analysis are: 1) there is no significant difference in the marginal product and marginal wages of married workers, and 2) the marginal wages of women appear to be significantly less than their marginal product. While these results are tentative they suggest two things. First, explanations for the observed marriage premium should focus on whether marriage is a signal for inherent productivity differences between married and single men or whether marriage in some way makes men more productive. Second, explanations for the gender wage gap should focus on why women receive lower wages than men and not on why women are less productive than men. However, the primary importance of these results is again the new insight into the wage determination process that we gain using employer-employee matched data.

C. *Technology Use and Worker Wages*[31]

While there has been growing interest among both economists and policy makers regarding the importance of skill-biased technical change in determining both the rate of return to

[31] For a complete discussion of the issues in the section see Doms, Dunne, and Troske (1994).

education and the increasing wage differential between skilled and unskilled workers, there have been few micro-level studies which contain direct evidence on the effects of technical change on worker wages. One of the principle reasons why, is the lack of data linking a plant's use of advanced technology and the plant's demand for skilled labor. Linking the WECD with the plant-level data from Census Bureau's Survey of Manufacturing Technology, which asks manufacturers about their use of advanced manufacturing technology in the plant, creates a data set which contains direct measures of a plant's use of technology, along with information on the characteristics of workers in the plant. These data can then be used to examine the effect of technology use on the wages and skill mix of workers in the plant.

As an example of this, consider Doms, Dunne and Troske (1994). Results in this paper show that plants that use advanced technology capital in production, and that produce more complex products, do pay workers higher wages. However, these authors also show that a significant portion of this premium is accounted for once they control for cross-plant differences in worker skill. These results are consistent with the hypothesis that much of the recent increase in the dispersion of wages is the result of skill-biased technical change. However, these results also represent one of the first successful attempts to show that worker skill varies systematically with employer characteristics.

V. **Summary**

Results from examining the quality of the WECD are mixed. The results from Section III shows that while a rather small percentage of workers and plants appear in the WECD, it does appear that workers are being matched to the correct establishments. The results from Tables 8 and 9 show that, while

the WECD data is not a representative sample of the underlying population of workers and plants, it is possible to construct weights so that estimates of characteristics using these data more closely resemble estimates of these characteristics from data on the underlying population. Even more importantly, the results in Table 10 show that these data are capable of replicating relationships found in both the original data and in previous research based on alternative data sources. This latter finding in particular suggests that these data are capable of investigating hypothesized relationships between worker and plant characteristics that are derived from theoretical models. Evidence on this point is found in Section IV where I present examples of how these data have been used to investigate hypotheses regarding the determination of worker wages. I should point out however, that these data will offer only limited support for theories. They can show whether or not the hypothesized relationships are present in a select sample of workers and plants -- they may not generalize to the entire population. However, given the uniqueness of these data, even with these limitations, these data should prove to be a valuable research tool.

VI. **Future Plans**

One of the strongest conclusions that emerges from this analysis is that creating employer-employee matched databases requires very detailed information upon which to base the match. The two major weaknesses of the WECD are a direct result of not having detailed place-of-work information. First, the numbers in Table 5 show that, even with the relatively detailed industry and location information available, over 80% of manufacturing plants are not unique in an industry-location cell, and therefore, cannot be uniquely matched to workers. This in turn causes the

WECD to be a nonrandom sample. Second, preliminary analysis suggested that it would have been impossible to accurately match nonmanufacturing workers and employers given the available location information. Thus, the WECD only contain data for manufacturing workers and employers. Obviously, if we hope to produce larger more representative employer-employee matched databases containing workers and employers from all sectors of the economy we will need more detailed information to link workers to employers.

Luckily, it appears that more detailed information is available. Currently, the SSEL contains the name and street address for all establishments in the United States. In addition, the name and address of a worker's employer is collected on the long-form of the Decennial Census. Unfortunately, it was not possible to use this information when constructing the WECD, because this information was destroyed prior to starting the WECD project. However, in the future this information will be kept and made available to researchers at the Census Bureau. This more detailed information, in conjunction with business name and address matching algorithms, should allow us to construct larger, more representative employer-employee matched data bases, and allow us to extend these databases to nonmanufacturing workers and employers.

Preliminary tests of this hypothesis could occur quite soon. The Census Bureau is currently planning on replacing the long form questionnaire with the "Continuous Measurement" (CM) survey, which will be a monthly household survey designed to collect long form data continuously. The Center for Economic Studies has begun working with the individuals in charge of processing the CM data to ensure that the data are processed in ways that will helps us link workers and employers. Preliminary tests of the CM survey are scheduled for January 1995. Data from this test will be made available to the Center for Economic Studies shortly

afterwards. The data from this test will allow us to determine whether having more detailed name and address information does produce a more representative sample. If these tests are successful, and if the CM program becomes fully operational in 1999 as is currently planned, then we should be able to start producing yearly versions of the WECD as soon as 2000.

References

Bates, Nancy, Robert E. Fay, and Jeffery C. Moore. "Lower Mail Response in the 1990 Census: A Preliminary Interpretation." In *1991 Annual Research Conference Proceedings* Washington D.C.: U.S. Bureau of the Census (1991).

Brown, Charles and James Medoff. "The Employer Size-Wage Effect." *Journal of Political Economy* 97 (November 1989): 1027-1059.

_____. "Firm Age and Wages." unpublished paper, University of Michigan (March 1995).

Cain, Glen. "The Economic Analysis of Labor Market Discrimination: A Survey." In *Handbook of Labor Economics* edited by Orley C. Ashenfelter and Richard Layard. Amsterdam: North Holland, 1986.

Doms, Mark, Timothy Dunne, and Kenneth R. Troske. "Workers, Wages, and Technology." mimeo, Center for Economic Studies, U.S. Bureau of the Census, Washington D.C. (December 1994).

Dunne, Timothy and James A. Schmitz. "Wages, Employment Structure and Employer Size-Wage Premia: Their Relationship to Advanced-Technology Usage at U.S. Manufacturing Establishments." Discussion Paper No. 92-15 Washington D.C.: Bureau of the Census, Center for Economic Studies (December 1992).

_____ and Mark Roberts. "Plant, Firm and Industry Wage Variation." unpublished paper, University of Oklahoma (December 1990).

Griliches, Zvi. "Capital-Skill Complementarity." *Review of Economics and Statistics* vol. 51. (1969): 465-68.

_____. "Notes on the Role of Education in Production Functions and Growth Accounting." In *Education, Income, and Human Capital*, edited by W. Lee Hansen. New York: NBER 1970.

Hamermesh, Daniel S. "Commentary." In *The Economics of Firm Size, Market Structure and Social Performance*, edited by John J. Siegfried. Washington D.C.: Federal Trade Commission, 1980.

_____. *Labor Demand*. Princeton, NJ: Princeton University Press, 1993.

Hellerstein, Judith K., David Neumark, and Kenneth R. Troske. "Wages, Productivity, and Worker Characteristics."

unpublished paper, Center for Economic Studies, U.S. Bureau of the Census, Washington D.C. (November 1994).

Korenman, Sanders and David Neumark. "Does Marriage Really Make Men More Productive?" *Journal of Human Resources* vol. XXVI no. 2 (Spring 1991): 282-307.

Kulka, Richard A., Nicholas A. Holt, Woody Carter, and Kathryn L. Dowd. "Self-Reports of Time Pressures, Concerns for Privacy, and Participation in the 1990 Mail Census." In *1991 Annual Research Conference Proceedings* Washington D.C.: U.S. Bureau of the Census (1991).

Mincer, Jacob. *Schooling, Experience, and Earnings*. New York: NBER 1974.

McGuckin, Robert and George Pascoe. "The Longitudinal Research Database (LRD): Status and Research Possibilities." *Survey of Current Business* (Nov. 1988): 30-37.

Oi, Walter Y.. "The Fixed Employment Costs of Specialized Labor." In *The Measurement of Labor Costs*, edited by Jack E. Triplett. Chicago: University of Chicago Press (for NBER), 1983.

_____. "Employment Relations in Dual Labor Markets (``It's Nice Work If You Can Get It'')." *Journal of Labor Economics* 8 (January 1990): S124-S149.

Rosen, Sherwin. "The Theory of Equalizing Differences." In *Handbook of Labor Economics*, edited by O. Ashenfelter and R. Layard. Amsterdam: North Holland,

1986.

Willis, Robert. "Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions." in *Handbook of Labor Economics*, edited by O. Ashenfelter and R. Layard. Amsterdam:North Holland, 1986.

Troske, Kenneth R. "Evidence on the Employer Size-Wage Premium From Worker-Establishment Matched Data." Discussion Paper No. 94-10 Washington D.C.: U.S. Bureau of the Census, Center for Economic Studies (August 1994).

U.S. Bureau of the Census. "The Standard Statistical Establishment List Program." Technical Paper no. 44. Washington D.C., 1979.

_____. "1990 Census of Population and Housing--Guide Part A. Text." Washington D.C., 1992a.

_____. "1990 Census of Population and Housing--Classified Index of Industries and Occupations." Washington D.C., 1992b.

Table 1

Comparing Matched Plant and Worker Data

		All Matched Workers and Plants (1)	Only Workers Between Age 18 and 65 Who Usually Worked 30-65 Hours a Week (2)	Only Plants with More than 10% of the Workforce Matched (3)
SSEL Worker Earnings	(1)	24371.17 (148.27)	25204.59 (144.09)	23542.37 (179.40)
SDF Worker Earnings	(2)	24317.26 (115.28)	24530.20 (117.45)	23838.04 (207.58)
Log Difference (across plants)	(3)	-0.048 (0.005)	0.003 (0.005)	-0.006 (0.008)
D (SSEL Worker Earnings, SDF Worker Earnings)	(4)	0.47 (0.001)	0.45 (0.001)	0.33 (0.001)
Mean Total Employment in Plants	(5)	151.43 (4.32)	156.29 (4.48)	105.74 (4.70)
Mean Proportion of Workers Matched to the Plants	(6)	0.16 (0.002)	0.15 (0.002)	---
Number of Plants	(7)	15,435	14,851	7226

Note: The numbers in parentheses are standard errors except for row (4) where they are p-values.

Table 2

Comparing Matched Plant and Worker Data by Size and Region
For Workers Between 18 and 65 Who Usually Worked Between 30 and 65 Hours a Week

	SSEL Worker Earnings	SDF Worker Earnings	Log Difference	D(SSEL Earnings, SDF Earnings)	Proportion Matched	Number of Plants
			Panel A - Plant Size (Total Employment)			
1-9	24146.61	22173.18	0.142	0.20	0.40	2277
	(381.37)	(453.24)	(0.02)	(0.0001)	(0.006)	
10-24	24955.41	23803.62	-0.001	0.32	0.16	2718
	(436.68)	(302.33)	(0.01)	(0.0001)	(0.003)	
25-49	25252.59	24286.80	-0.040	0.41	0.10	2542
	(425.09)	(304.20)	(0.01)	(0.0001)	(0.002)	
50-99	24628.26	24205.75	-0.025	0.52	0.09	2746
	(289.74)	(182.88)	(0.009)	(0.0001)	(0.002)	
100-249	25185.07	25068.49	-0.014	0.60	0.08	2640
	(237.41)	(174.12)	(0.020)	(0.0001)	(0.001)	
250-499	25408.95	25908.63	-0.033	0.68	0.08	1079
	(306.91)	(274.49)	(0.010)	(0.0001)	(0.002)	
500-999	27881.66	25950.63	-0.026	0.76	0.08	520
	(428.18)	(427.73)	(0.011)	(0.0001)	(0.003)	
1000+	34280.33	35850.85	-0.036	0.78	0.08	329
	(531.51)	(576.57)	(0.013)	(0.0001)	(0.004)	
			Panel B - Census Division			
New England	27432.81	26314.58	0.032	0.41	0.12	1429
	(520.59)	(496.30)	(0.015)	(0.0001)	(0.005)	
Mid. Atlantic	26446.22	25092.65	0.009	0.46	0.14	3391
	(357.98)	(231.26)	(0.010)	(0.0001)	(0.003)	
East-No. Central	26149.54	25887.90	-0.012	0.44	0.14	4224
	(268.08)	(208.37)	(0.009)	(0.0001)	(0.003)	
West-No. Central	23895.70	24537.35	-0.037	0.46	0.16	1198
	(434.34)	(438.11)	(0.018)	(0.0001)	(0.005)	
South Atlantic	23132.80	22138.76	0.020	0.43	0.14	768
	(323.94)	(310.25)	(0.014)	(0.0001)	(0.004)	
East-So. Central	21531.13	21325.68	0.007	0.47	0.14	768
	(397.98)	(571.55)	(0.021)	(0.0001)	(0.006)	
West-So. Central	21570.96	21555.19	-0.015	.040	0.17	900
	(443.11)	(367.30)	(0.022)	(0.0001)	(0.007)	
Mountain	21132.11	20512.80	0.027	0.38	0.17	318
	(663.16)	(636.55)	(0.044)	(0.0001)	(0.011)	
Pacific	26503.12	24931.35	0.038	0.36	0.20	891
	(649.21)	(501.76)	(0.025)	(0.0001)	(0.009)	

Note: The numbers in parentheses are standard errors except in column (5) where they are p-values.

Table 3

Comparing Matched Plant and Worker Data by Industry
For Workers Between 18 and 65 Who Usually Worked Between 30 and 65 Hours a Week

Industry	SSEL Worker Earnings (1)	SDF Worker Earnings (2)	Log Difference (3)	D(SSEL Earnings, SDF Earnings) (4)	Proportion Matched (5)	Number of Plants (6)
Food	24055.82	23750.41	-0.01	0.48	0.12	1665
	(347.16)	(421.18)	(0.01)	(0.0001)	(0.003)	
Tobacco	22557.58	26785.83	0.24	0.68	0.08	25
	(2502.03)	(2020.56)	(0.09)	(0.0002)	(0.01)	
Textile	20419.94	20618.58	-0.03	0.46	0.13	438
	(561.06)	(660.45)	(0.03)	(0.0001)	(0.01)	
Apparel	15462.98	16470.58	0.02	0.33	0.13	559
	(380.04)	(544.22)	(0.03)	(0.0001)	(0.01)	
Lumber	20039.38	23254.54	0.08	0.27	0.19	572
	(460.79)	(912.31)	(0.03)	(0.0001)	(0.01)	
Furniture	20047.37	22125.10	0.02	0.42	0.19	379
	(421.61)	(996.03)	(0.03)	(0.0001)	(0.01)	
Paper	26981.37	27280.02	-0.04	0.50	0.10	866
	(303.99)	(525.90)	(0.02)	(0.0001)	(0.004)	
Printing	19348.33	21666.39	0.09	0.44	0.16	1228
	(313.51)	(362.91)	(0.02)	(0.0001)	(0.01)	
Chemicals	30598.58	30012.29	-0.03	0.28	0.17	1165
	(641.66)	(501.74)	(0.02)	(0.0001)	(0.01)	
Petroleum Refining	37282.11	33492.94	-0.13	0.07	0.17	161
	(1434.79)	(1502.55)	(0.05)	(0.38)	(0.02)	
Rubber	23691.93	24052.27	-0.03	0.45	0.12	717
	(467.37)	(467.37)	(0.02)	(0.0001)	(0.01)	
Leather	16662.93	17503.39	0.05	0.46	0.14	178
	(754.53)	(777.90)	(0.05)	(0.0001)	(0.01)	
Stone	26068.61	25288.76	-0.06	0.41	0.14	853
	(409.75)	(528.45)	(0.02)	(0.0001)	(0.01)	
Primary Metals	26942.87	27624.96	-0.02	0.45	0.12	898
	(372.66)	(702.90)	(0.02)	(0.0001)	(0.005)	
Fabricated Metals	26287.79	26299.20	-0.04	0.33	0.14	1490
	(500.68)	(484.06)	(0.02)	(0.0001)	(0.005)	
Machinery	27216.31	28512.74	0.02	0.34	0.19	1421
	(324.71)	(576.73)	(0.02)	(0.0001)	(0.01)	

Table 3 - cont.

Industry	SSEL Worker Earnings (1)	SDF Worker Earnings (2)	Log Difference (3)	D(SSEL Earnings, SDF Earnings) (4)	Proportion Matched (5)	Number of Plants (6)
Electrical Equipment	23467.39 (394.61)	25601.72 (608.20)	0.06 (0.02)	0.40 (0.0001)	0.13 (0.01)	726
Transpor-tation	26112.19 (455.76)	26212.33 (534.98)	0.01 (0.02)	0.52 (0.0001)	0.17 (0.01)	715
Instruments	28540.42 (1049.58)	29043.37 (950.43)	0.02 (0.05)	0.18 (0.0041)	0.17 (0.02)	257
Misc.	20423.02 (427.49)	22959.16 (696.47)	0.07 (0.03)	0.26 (0.0001)	0.17 (0.01)	538

Note: The numbers in parentheses are standard errors, except in column (4) where they are p-values.

Table 4

Number and Mean Earnings of SDF and WECD Workers By Industry

Industry	SDF Workers (1)	WECD Workers (2)	Proportion Matched (3)	Mean Earnings SDF Workers (4)	Mean Earnings WECD Workers (5)	Log Difference (6)
Food	231420	20597	0.09	22,131	23,619	0.07
Tobacco	7393	1379	0.19	35,899	35,890	0.00
Textile	121159	6485	0.05	18,307	19,228	0.05
Apparel	161014	6255	0.04	13,946	14,722	0.05
Lumber	134031	3856	0.03	18,214	26,448	0.37
Furniture	92274	3217	0.04	18,576	20,482	0.10
Paper	106615	14411	0.14	29,322	31,217	0.06
Printing	282069	11510	0.04	23,143	21,154	-0.09
Chemicals	176282	12089	0.07	33,342	33,183	0.00
Petroleum	27194	1913	0.07	36,301	37,633	0.04
Rubber	109594	8608	0.08	23,484	25,854	0.10
Leather	24484	2442	0.10	16,025	16,606	0.04
Stone	88855	6666	0.08	24,271	26,167	0.08
Primary Metals	126963	17224	0.14	28,897	31,854	0.10
Fabricated Metals	185281	13435	0.07	25,108	27,417	0.09
Machinery	373079	17313	0.05	28,804	31,515	0.09
Electrical Equipment	281519	14633	0.05	27,810	25,342	-0.09
Transpor-tation	379002	30622	0.08	32,035	35,379	0.10
Instrument	92684	2406	0.03	29,057	29,868	0.03
Misc.	176074	4442	0.03	21,693	21,264	-0.02
Total	3,176,986	199,558	0.06	25,558	28,107	0.10

Table 5

Number, Proportion and Average Total Employment of All, Unique, and Matched Plants By Industry

Industry	SSEL Plants (1)	Unique Plants (2)	WECD Plants (3)	Proportion Unique (4)	Proportion Matched (5)	SSEL Plant Emp (6)	Unique Plant Emp (7)	WECD Plant Emp (8)
Food	19117	6598	1801	0.35	0.09	75.6	89.9	143.4
Tobacco	134	75	25	0.56	0.19	297.4	417.5	844.0
Textile	5838	1804	466	0.31	0.08	112.0	124.4	161.4
Apparel	21275	2858	643	0.13	0.03	47.9	76.7	110.5
Lumber	31573	3845	657	0.12	0.02	22.2	31.3	52.5
Furniture	11168	1612	421	0.14	0.04	45.3	50.8	64.5
Paper	6126	2342	888	0.38	0.15	103.1	123.7	163.5
Printing	58803	5514	1491	0.09	0.03	26.3	39.3	75.3
Chemicals	11659	3914	1230	0.34	0.11	74.3	82.5	126.9
Petroleum Refining	2161	922	165	0.43	0.08	53.4	67.3	130.8
Rubber	14435	2884	752	0.20	0.05	60.8	93.1	155.0
Leather	1897	767	198	0.40	0.10	62.2	76.0	118.1
Stone	15245	4368	931	0.29	0.06	34.2	44.4	80.0
Primary Metals	6548	2843	934	0.43	0.14	109.7	130.9	222.1
Fabricated Metals	35513	6742	1580	0.19	0.04	41.7	61.3	121.6
Machinery	49097	6255	1514	0.13	0.03	39.1	68.5	127.8
Electrical Equipment	15941	2887	757	0.18	0.05	97.4	142.3	240.0
Transpor-tation	10002	3170	762	0.32	0.08	180.7	241.9	448.4
Instrument	9688	1851	283	0.19	0.03	99.6	123.6	229.4
Misc.	16251	2698	646	0.17	0.04	24.2	36.7	66.6
Total	342,471	63,949	16,144	0.19	0.05	52.2	84.5	146.3

Table 6

Number and Mean Earnings of SDF and WECD Workers By Census Division

Census Division	Number of SDF Workers (1)	Number of WECD Workers (2)	Proportion Matched (3)	Mean Earnings SDF Workers (4)	Mean Earnings WECD Workers (5)	Log Difference (6)
New England	189131	17673	0.09	28781.95	28822.79	0.00
Mid. Atlantic	469899	37820	0.08	27559.07	27151.79	0.01
East-No. Central	772079	69986	0.09	27362.52	30617.08	-0.05
West-No. Central	276567	18682	0.07	23049.96	26582.73	-0.06
South Atlantic	479648	20263	0.04	22508.84	25788.60	-0.06
East-So. Central	234695	11066	0.05	20469.50	23810.22	-0.07
West-So. Central	293049	12234	0.04	23764.57	23212.54	0.01
Mountain	105588	3408	0.03	24224.02	23400.80	0.02
Pacific	356322	8426	0.02	28571.62	33644.64	-0.07

Table 7

Number, Proportion and Average Total Employment of SDF, Unique, and Matched Plants
By Plant Size and Census Division

	SSEL Plants (1)	Unique Plants (2)	WECD Plants (3)	Proportion Unique (4)	Proportion Matched (5)	SSEL Plant Emp (6)	Unique Plant Emp (7)	WECD Plant Emp (8)
				Panel A - Plant Size (Total Employment)				
1-9	161192	24765	2924	0.15	0.02	4.1	4.1	5.0
10-24	74981	12944	3088	0.17	0.04	15.5	15.7	16.2
25-49	41796	8415	2687	0.20	0.06	34.9	35.2	35.9
50-99	28877	7014	2821	0.24	0.10	70.1	70.8	71.2
100-249	22599	6401	2673	0.28	0.12	154.2	155.8	156.3
250-499	7973	2259	1091	0.28	0.14	345.8	347.9	346.5
500-999	3378	1197	526	0.35	0.16	679.3	680.2	683.3
1000+	1675	654	334	0.39	0.20	2411.6	2450.2	2527.3
				Panel B - Census Division				
New England	23616	5416	1560	0.23	0.07	48.8	67.8	153.2
Mid. Atlantic	54657	12063	3667	0.22	0.07	46.4	70.4	116.2
East-No. Central	65381	13629	4526	0.21	0.07	59.3	95.6	165.8
West-No. Central	23252	5478	1308	0.24	0.06	56.2	84.7	153.5
South Atlantic	50336	8013	1866	0.16	0.04	58.9	108.5	178.6
East-So. Central	19235	3847	815	0.20	0.04	69.9	113.9	169.9
West-So. Central	34872	5831	1025	0.17	0.03	47.4	72.9	123.2
Mountain	15868	2553	385	0.16	0.02	38.6	63.7	111.7
Pacific	55254	7119	992	0.13	0.02	44.1	73.6	104.5

Table 8

Comparing the Characteristics of SDF, CPS, and WECD Workers

	SDF Workers (1)	1988 May CPS Workers Manufacturing (2)	WECD Workers (3)	WECD Workers Weighted (4)
Percent Male	66.9	65.4	70.1	66.9
Percent Non-Hispanic White	85.2	88.8	89.6	88.3
Percent Now Married	67.3	66.7	71.0	67.7
Occupation				
Percent Manager and Professional	18.2	18.6	16.4	19.2
Percent Technical, Clerical, and Sales	21.6	20.8	19.7	21.4
Percent Production Worker	60.2	60.6	64.0	59.4
Percent in Region				
Northeast	20.8	27.6	27.9	19.9
Midwest	33.0	28.4	44.5	33.3
South	31.7	32.5	21.8	33.8
West	14.5	11.5	5.9	11.8
Mean Age	38.9 (37)	38.3 (37)	39.9 (39)	38.8 (39)
Mean Number of Weeks Worked	47.5 (52)	--	48.9 (52)	48.2 (52)
Mean Usual Hours Worked Per Week	41.2 (40)	41.0 (40)	41.7 (40)	41.3 (40)
Mean Wage or Salary Income	25558.1 (21000)	--	28106.7 (25000)	25676.8 (25000)
Mean Hourly Wage	13.25 (10.58)	10.30 (9.08)	13.87 (11.96)	12.90 (11.96)
Number of Workers	3,176,986	4757	199,558	1,639,556.2

Note: Numbers in parentheses are the medians of the distribution. The reference period for Number of Weeks Worked, Usual Hours Worked Per Week, Wage or Salary Income and Hourly Wage is the previous year (1989) for SDF and WECD workers, and the previous week for the CPS workers. For the SDF and WECD workers Hourly Wage is estimated as: (Wage or Salary Income ÷ Number of Weeks Worked) ÷
Usual Hours Worked Per Week.

Table 9

Characteristics of SDF Plants, Unique Plants, and WECD Plants

	SSEL Plants (1)	Unique Plants (2)	Unique Plants Weighted (3)	WECD Plants (4)	WECD Plants Weighted (5)
Mean Employment	52.20 (11)	84.50 (16)	60.21 (16)	146.27 (43)	63.3 (43)
Mean Annual Payroll	1,414,237 (199,000)	2,377,177 (312,000)	1,688,294 (312,000)	4,411,189 (943,000)	1,731,777 (943,000)
Average Earnings	21,495.54 (18,686.36)	21,916.90 (19,500.00)	21,819.17 (19,500.00)	24,088.37 (22,531.25)	22,540.01 (22,531.25)
Percent in Place	74.1	87.7	61.1	89.1	61.1
Percent Multi-unit	20.0	31.1	20.1	44.3	22.7
Plant Age					
0-4	26.8	19.4	21.4	5.9	8.8
5-9	18.5	22.7	25.4	20.8	27.3
10-14	23.3	30.2	24.7	41.5	29.3
15+	26.9	27.8	28.5	31.9	34.6
Percent in Region					
Northeast	22.9	27.4	22.0	32.4	23.4
Midwest	25.9	29.9	24.3	36.1	25.9
South	30.5	27.6	31.2	22.9	33.3
West	20.5	15.1	12.6	8.5	17.3
Number	342,524	63,949	381,309.22	16,144	317,440.76

Note: The numbers in parentheses are medians of the distribution.

Table 10

Regression of Worker Wages for SDF and WECD Workers

	SDF Workers with Match (1)	SDF Workers without Match (2)	WECD Workers (3)	WECD Workers Weighted (4)
Intercept	1.55	1.55	1.41	1.34
	(0.008)	(0.008)	(0.031)	(0.02)
Exp	0.06	0.06	0.06	0.07
	(0.001)	(0.001)	(0.002)	(0.002)
Exp^2 *10	-0.02	-0.02	-0.02	-0.03
	(0.001)	(0.001)	(0.001)	(0.001)
Exp^3*1000	0.05	0.05	0.05	0.07
	(0.002)	(0.002)	(0.004)	(0.004)
Exp^4*10000	-0.05	-0.05	-0.04	-0.06
	(0.002)	(0.002)	(0.004)	(0.004)
Female	-0.17	-0.17	-0.17	-0.18
	(0.002)	(0.002)	(0.004)	(0.004)
Married	0.13	0.13	0.12	0.13
	(0.001)	(0.001)	(0.002)	(0.003)
Black	-0.06	-0.06	-0.04	-0.05
	(0.002)	(0.002)	(0.004)	(0.005)
Female*Married	-0.14	-0.14	-0.13	-0.14
	(0.002)	(0.002)	(0.004)	(0.004)
Female*Black	0.08	0.08	0.07	0.08
	(0.004)	(0.004)	(0.007)	(0.008)
Educ1	----	----	----	----
Educ2	0.13	0.13	0.10	0.12
	(0.001)	(0.001)	(0.002)	(0.003)
Educ3	0.21	0.21	0.17	0.18
	(0.002)	(0.002)	(0.003)	(0.003)
Educ4	0.41	0.41	0.36	0.39
	(0.002)	(0.002)	(0.004)	(0.004)
Educ5	0.55	0.55	0.49	0.47
	(0.003)	(0.003)	(0.006)	(0.006)
Match	0.03	----	----	----
	(0.002)			
Adj. R^2	0.50	0.50	0.51	0.47
Number Obs.	704,373	704,373	185,186	185,007

Note: These regressions only include workers who are between 18 and 65 years old, who usually work between 30 and 65 hours per week, and who have average wages between $2.50 and $100.00 an hour.

Table 11

Plant Level Regression of Log Average Earnings in the Plant

	SSEL Plants (1)	Unique Plants (2)	Unique Plants Weighted (3)	WECD Plants (4)	WECD Plants Weighted (5)
Intercept	2.34	2.46	2.35	2.51	2.25
	(0.017)	(0.04)	(0.035)	(0.114)	(0.082)
Log Employment	0.18	0.15	0.17	0.13	0.23
	(0.003)	(0.005)	(0.005)	(0.010)	(0.010)
Log Employment Squared	-0.02	-0.01	-0.01	-0.01	-0.02
	(0.000)	(0.001)	(0.001)	(0.001)	(0.002)
Place	0.03	-0.10	-0.07	-0.18	-0.10
	(0.013)	(0.036)	(0.025)	(0.082)	(0.056)
Multi-unit	0.16	0.16	0.14	0.13	0.13
	(0.004)	(0.008)	(0.009)	(0.011)	(0.015)
Plant Age					
0-4	-0.19	-0.19	-0.17	-0.18	-0.18
	(0.004)	(0.009)	(0.009)	(0.015)	(0.015)
5-9	-0.10	-0.11	-0.11	-0.10	-0.08
	(0.003)	(0.007)	(0.008)	(0.012)	(0.013)
10-14	-0.05	-0.06	-0.04	-0.06	-0.07
	(0.004)	(0.008)	(0.008)	(0.011)	(0.013)
15+	----	----	----	----	----
Region					
Northeast	0.13	0.14	0.22	0.08	0.32
	(0.012)	(0.036)	(0.020)	(0.103)	(0.051)
Midwest	-0.01	0.05	0.10	0.005	0.15
	(0.012)	(0.036)	(0.021)	(0.103)	(0.051)
South	-0.02	0.05	0.15	-0.02	0.19
	(0.012)	(0.036)	(0.020)	(0.103)	(0.050)
West	----	----	----	----	----
Adj. R^2	0.23	0.25	0.22	0.30	0.29
Number	234,694	49,735	49,698	15,138	15,137

Note: Standard Errors are in parenthesis

Figure 1 - Educational Distribution of SDF and WECD Workers

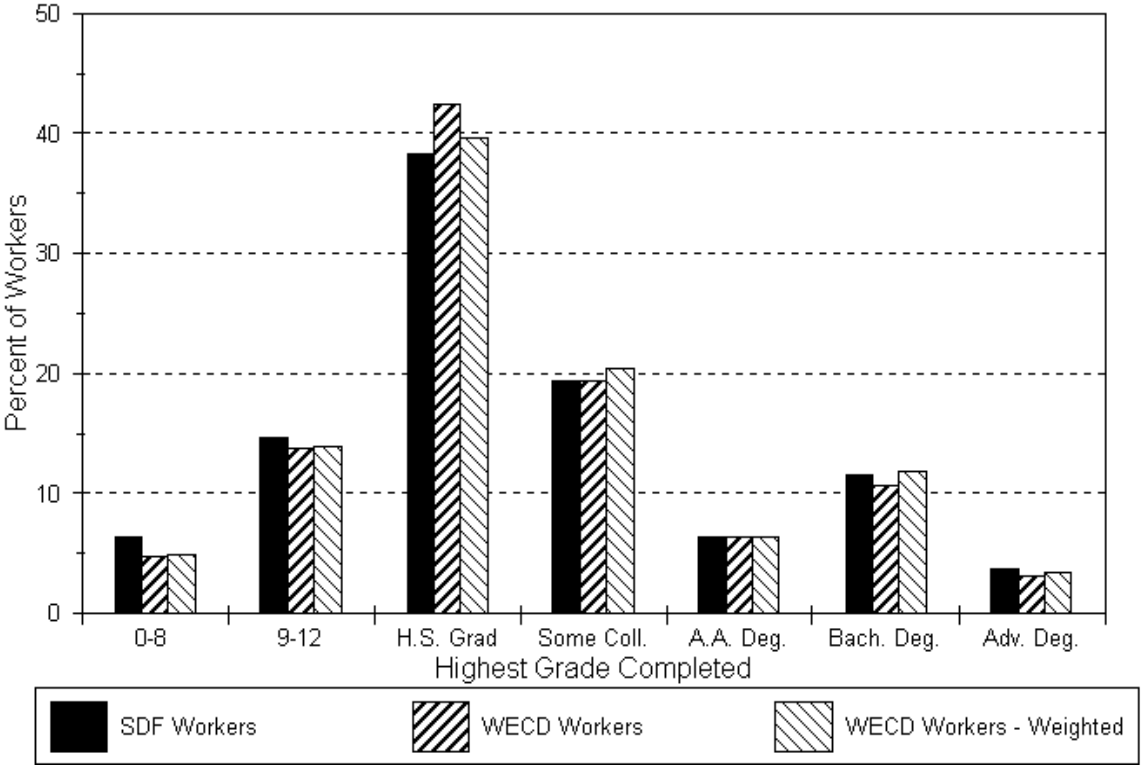

Table A.1

Worker Variables Available from the WECD

Place of Residence: State Code
Place of Residence: County Code
Place of Residence: Place Code
Place of Residence: Block Code
Sex
Detailed Race Code (Three-Digit Race Code)
Age
Marital Status
Person Weight
Place of Birth
Citizenship
Year of Entry
School Enrollment
Highest Degree Completed
Ancestry (Six-Digit Code)
Mobility Status (Where Lived on April 1, 1985)
Language Other Than English at Home
English Ability
Military Service
Work Limitation Status
Mobility Limitation
Personal Care Limitation
Number of Children Ever Born
Hours Worked Last Week
Principal Means of Transportation to Work
Time of Departure for Work
Travel Time to Work
Occupation (Three Digit Code)
Class of Worker
Worked Last Year (1989)
Week Worked Last Year (1989)
Usual Hours Worked Last Year (1989)
Wage or Salary Income (1989)
Nonfarm Self-Employment Income (1989)
Farm Self-Employment Income (1989)
Interest, Dividends, and Net Rental Income (1989)
Social Security Income (1989)
Public Assistance Income (1989)
Retirement Income (1989)
All Other Income (1989)

Establishment Variables Available in the LRD

Total Value of Shipments
Four-Digit SIC Code
Establishment State Code
Establishment County Code
Establishment Place Code
Value Added
Value of Resales
Receipts for Contract Work
Miscellaneous Receipts
Total Employment
Total Employment-Production Workers
Total Production Worker Manhours
Total Salary and Wages
Total Production Worker Wages
Total Supplemental Labor Costs
Legally Required Supplemental Labor Costs
Cost of Materials
Cost of Resales
Cost of Fuels
Cost of Purchased Electricity
Cost of Contract Work
Beginning of Year Inventory: Finished Goods
Beginning of Year Inventory: Work-In-Progress
Beginning of Year Inventory: Materials
Beginning of Year Inventory: Total
End of Year Inventory: Finished Goods
End of Year Inventory: Work-In-Progress
End of Year Inventory: Materials
End of Year Inventory: Total
New Building Expenditure
New Machinery Expenditures
Used Capital Expenditures
Beginning of Year: Building Assets
Beginning of Year: Machinery Assets
End of Year: Building Assets
End of Year: Machinery Assets
Building Depreciation
Machinery Depreciation
Building Retirements
Machinery Retirements
Material Code
Product Code