Volume Title: Labor Statistics Measurement Issues

Volume Author/Editor: John Haltiwanger, Marilyn E. Manser and Robert Topel, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-31458-8

Volume URL: http://www.nber.org/books/halt98-1

Publication Date: January 1998

Chapter Title: The Worker-Establishment Characteristics Database

Chapter Author: Kenneth R. Troske

Chapter URL: http://www.nber.org/chapters/c8366

Chapter pages in book: (p. 371 - 404)

11 The Worker-Establishment Characteristics Database

Kenneth R. Troske

11.1 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 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" (1986, 688).[1]

The motivation behind the Rosen quote is that existing data sources have proved 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 the sorting of workers and employers in the labor market is required before we will begin to understand a number of current puzzles in labor economics such as rising wage inequality or the establishment size-wage

Kenneth R. Troske is assistant professor of economics at the University of Missouri, Columbia. This work was conducted while the author was an economist at the Center for Economic Studies, U.S. Bureau of the Census.

The author thanks 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 Bureau of the Census 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 Bureau of the Census.

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."

premium. As the Rosen quote makes clear, further understanding of the matching of workers and employers will only come about through the use of employer-employee matched data.

Employer-employee matched data would 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 possible biases resulting from treating labor as a unidimensional 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 United States. 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 employer-employee matched data and discuss 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 United States 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 provides both cross-sectional demographic information for workers such as age, sex, race, marital status, and earnings and 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 can replicate results obtained by previous researchers using alternative data sources.

Results from this analysis are somewhat mixed. On the positive side, several facts suggest that most WECD workers are matched to the correct establish-

2. The WECD is limited to manufacturing workers and plants for two reasons. First, preliminary analysis suggested that it would be impossible to match nonmanufacturing employers and employees given the limited place-of-work information, and second, the LRD only contains data for manufacturing plants. The availability of plant data depends on the year. In Census of Manufactures 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.

ments. 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 5 percent, and the two estimates are positively and significantly correlated. Second, establishments in the WECD have on average 16 percent 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 from alternative data sets.

On the negative side, only 6 percent of manufacturing workers in the SDF and 5 percent 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 SDF, and relative to all plants in the SSEL, the WECD contains a larger proportion of large, old, urban establishments and establishments located in the northeastern and midwestern regions of the country. 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.

Because the WECD does not contain a representative sample of workers and employers and 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 them 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 it only contains information for manufacturing workers and employers. To try to address this problem, and to make the data more representative, future versions of the WECD will be created from data with much more detailed place-of-work information. While these data were originally collected for workers in the decennial census, they were destroyed prior to the start of this project. However, in the future, this more de-

3. E.g., 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 is true for all workers in the labor market we would need to test this hypothesis on a random sample of workers.

tailed place-of-work information for workers will be saved, making it possible to 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 11.2 discusses the data sets used to match workers to establishments and outlines the matching process. Section 11.3 investigates the accuracy of the match. Section 11.4 presents examples of how these data can be used in empirical work to increase our understanding of the wage determination process. Section 11.5 summarizes and discusses preliminary plans for creating new versions of the WECD.

11.2 The Data and the Matching Algorithm

11.2.1 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] "At what location did this person work *last week?*" and "What kind of business or industry was this?"[5] The Census Bureau assigns geographic and industry codes to each person's record in the SDF based on the 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 select all respondents who indicated that they worked in manufacturing and worked in the previous week. This file contains approximately 3.18 million individual records.[6]

Each plant record in the 1990 SSEL includes a four-digit SIC code indicating the establishment's primary industry and geographic codes showing its location.[7] This information allows each plant in the United States to be assigned to a unique industry-location cell. For this project all 342,471 manufacturing establishments are selected from the 1990 SSEL.[8]

4. For a more complete discussion of data available from the 1990 decennial census, along with a copy of the long form, see Bureau of the Census (1992b). 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 fewer than 2,500 people a form was sent to one in two households, while in tracts with more than 2,500 housing units it was sent to one in eight households.

5. 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.

6. The estimated manufacturing workforce based on the 1990 census is 20.5 million, so the SDF sample of 3.18 million represents approximately 16 percent 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.

7. For a more complete description of the SSEL, see Bureau of the Census (1979).

8. The entire 1990 SSEL contains approximately 7.04 million nonagricultural establishments, of which 424,519 are 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 342,471 establishments.

11.2.2 The Matching Process

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.[9] The Census Bureau divides the entire country into a hierarchy of geographic areas and assigns codes to each area. The most aggregated 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 is assigned a unique geographic code, as is each county within a state. Thus each county in the United States has a unique state-county code combination. Counties are further divided into incorporated and unincorporated areas, and each incorporated area with a population of over 2,500 is assigned a unique place code.[10] Finally, highly populated places are further subdivided, with each separate physical block in a place assigned a unique block code.[11] 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 percent of the establishments in the 1990 SSEL.[12]

Industry codes must also be standardized since establishments in the 1990 SSEL are classified into industries using the SIC system, while workers in the

9. For a more complete description of geographic codes, see Bureau of the Census (1992b).

10. Portions of counties not in a qualifying place are assigned a place code of 9999.

11. In 1990 block codes were only available for addresses in Tape Address Register (TAR) areas. TAR areas roughly correspond to central cities or metropolitan statistical areas (MSAs).

12. 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.

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 percent of manufacturing establishments are located in an MSA. Thus I am missing block codes for only 4 percent of the establishments.

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.[13]

The second step in the matching process is to eliminate nonunique establishments. To do this I first 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 when all establishments in the industry-place cell have valid block codes, or when an establishment is unique in an industry-place cell.[14] Eliminating nonunique establishments reduces the number of establishments available for matching from 342,471 to 63,949. Next, I assign workers and establishments to industry-location cells and match workers and establishments in the same cell. This is a two-step process. First, workers and establishments are assigned to industry-*block* cells and matched. Then 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.[15] 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.[16]

The resulting data set contains 199,557 worker records matched to 16,144

13. See Bureau of the Census (1992a). SIC codes are converted to census codes because the census codes are more aggregated than SIC codes.

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

15. E.g., 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, whether or not 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.

16. 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 the pay period including 12 March, which is when employment is recorded in the SSEL, and 1 April, the 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. The last two reasons are more likely to be problems with small establishments.

different plants.[17] The appendix provides a list of variables available for workers in the WECD and for establishments in the LRD.

11.3 Evaluating the Worker-Establishment Characteristics Database

11.3.1 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.[18] The matching algorithm only matches workers to establishments that are unique in an industry-location cell. Therefore, if workers and establishments have the correct geographic and industry codes, all workers in an industry-location cell that 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 means that the WECD will contain a random sample of workers in the plant.[19]

In spite of these assurances, some tests of the match are desirable. To begin, table 11.1 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 11.1. 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 employment in the pay period including 12 March 1990. The number in row 1 is 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 the worker data. Each worker in the SDF reports his or her 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 number in row 2 is then the average of this per worker earnings

17. 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 years of the Census of Manufactures and the Annual Survey of Manufactures. Therefore, the number of plants for which detailed data are available depends on the year (in particular, whether a survey or a census was conducted in a year). E.g., 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.

18. One large source for coding error is assigning an industry code to a worker's description of the primary industry of his or her 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.

19. This assumes that there is no systematic bias in response rates to the long form. See Bates, Fay, and Moore (1991) and Kulka et al. (1991) for a discussion of response rates to the 1990 decennial census.

Table 11.1 **Comparing Matched Plant and Worker Data**

	All Matched Workers and Plants (1)	Only Workers between Ages 18 and 65 Who Usually Worked 30–65 Hours a Week (2)	Only Plants with More than 10% of the Workforce Matched (3)
1. SSEL worker earnings	24,371.17	25,204.59	23,542.37
	(148.27)	(144.09)	(179.40)
2. SDF worker earnings	24,317.26	24,530.20	23,838.04
	(115.28)	(117.45)	(207.58)
3. Log difference (across plants)	−0.048	0.003	−0.006
	(0.005)	(0.005)	(0.008)
4. ρ (SSEL worker earnings, SDF worker earnings)	0.47	0.45	0.33
	(0.001)	(0.001)	(0.001)
5. Mean total employment in plants	151.43	156.29	105.74
	(4.32)	(4.48)	(4.70)
6. Mean proportion of workers matched to the plants	0.16	0.15	–
	(0.002)	(0.002)	
7. Number of plants	15,435	14,851	7,226

Note: Numbers in parentheses are standard errors except for row 4, where they are *p*-values.

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 between 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 11.1 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 week in 1989; and column (3) presents numbers for plants with more than 10 percent of the workforce matched to the plant.

The numbers in table 11.1 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 between the two estimates is less than 5 percent.[20] Further, when we con-

20. There are a number of reasons why these two estimates might differ. First, the estimate of earnings per worker based on plant data is an estimate of earnings paid to a worker by the plant, while the estimate based on worker data is total earnings paid to a worker by all employers. If some workers in a plant hold multiple jobs, 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 in the pay period including 12 March 1990. If the plant is growing over the year, the pay for workers added to the plant after 12 March will appear in the wage data but these workers will not appear in the employment figures. This will tend to make

sider the samples in columns (2) and (3), this difference falls to less than 1 percent and is statistically insignificant. The numbers in row 4 show that the SSEL and SDF worker earnings are positively and significantly correlated.[21] Finally, row 6 shows that on average 16 percent 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 11.2 breaks out the numbers in table 11.1, first by the size of the plant (panel A), and second by the nine census divisions (panel B). The numbers in table 11.2 are for workers who are between 18 and 65 years old and who usually worked between 30 and 65 hours a week in the previous year.[22]

The numbers in panel A reveal no systematic relationship between the difference between SSEL and SDF worker earnings and plant size. The largest difference, 14 percent, is found for plants with 1–9 employees, while the smallest difference, −0.1 percent, 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 percent of their workforce matched to the plant. However, the correlation between SSEL and SDF worker earnings in these plants is only .20. In contrast, plants with over 1,000 workers average 8 percent of their workforce matched to the plant, while the correlation between the two earnings measures is .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 five employees this means that the minimum percentage matched will be 20 percent. 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 1,000 employees, we are able to get a relatively accurate estimate of worker wages with only 8 percent of the workforce. Overall, while it appears that smaller plants have a much larger proportion of their workforce matched, larger plants appear to have a much more representative sample of workers matched.

SSEL worker earnings larger than SDF worker earnings. Also, if employment in a plant is seasonal and 12 March is a period of low (high) employment, SSEL earnings will appear higher (lower) than SDF earnings.

21. 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 ρ will in general be less than one 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.

22. I focus on these workers for two reasons. First, these workers have the strongest labor market attachments and therefore should have the most reliable earnings and hours worked data. Second, the log difference across plants (table 11.1, row 3) is small and insignificant for these workers.

Table 11.2 **Comparing Matched Plant and Worker Data by Size and Region**

	SSEL Worker Earnings (1)	SDF Worker Earnings (2)	Log Difference (3)	ρ(SSEL Earnings, SDF Earnings) (4)	Proportion Matched (5)	Number of Plants (6)
A. *Plant size (total employment)*						
1–9	24,146.61	22,173.18	0.142	0.20	0.40	2,277
	(381.37)	(453.24)	(0.02)	(0.0001)	(0.006)	
10–24	24,955.41	23,803.62	−0.001	0.32	0.16	2,718
	(436.68)	(302.33)	(0.01)	(0.0001)	(0.003)	
25–49	25,252.59	24,286.80	−0.040	0.41	0.10	2,542
	(425.09)	(304.20)	(0.01)	(0.0001)	(0.002)	
50–99	24,628.26	24,205.75	−0.025	0.52	0.09	2,746
	(289.74)	(182.88)	(0.009)	(0.0001)	(0.002)	
100–249	25,185.07	25,068.49	−0.014	0.60	0.08	2,640
	(237.41)	(174.12)	(0.020)	(0.0001)	(0.001)	
250–499	25,408.95	25,908.63	−0.033	0.68	0.08	1,079
	(306.91)	(274.49)	(0.010)	(0.0001)	(0.002)	
500–999	27,881.66	25,950.63	−0.026	0.76	0.08	520
	(428.18)	(427.73)	(0.011)	(0.0001)	(0.003)	
1,000+	34,280.33	35,850.85	−0.036	0.78	0.08	329
	(531.51)	(576.57)	(0.013)	(0.0001)	(0.004)	

B. *Census division*						
New England	27,432.81	26,314.58	0.032	0.41	0.12	1,429
	(520.59)	(496.30)	(0.015)	(0.0001)	(0.005)	
Middle Atlantic	26,446.22	25,092.65	0.009	0.46	0.14	3,391
	(357.98)	(231.26)	(0.010)	(0.0001)	(0.003)	
East–North Central	26,149.54	25,887.90	−0.012	0.44	0.14	4,224
	(268.08)	(208.37)	(0.009)	(0.0001)	(0.003)	
West–North Central	23,895.70	24,537.35	−0.037	0.46	0.16	1,198
	(434.34)	(438.11)	(0.018)	(0.0001)	(0.005)	
South Atlantic	23,132.80	22,138.76	0.020	0.43	0.14	1,732
	(323.94)	(310.25)	(0.014)	(0.0001)	(0.004)	
East–South Central	21,531.13	21,325.68	0.007	0.47	0.14	768
	(397.98)	(571.55)	(0.021)	(0.0001)	(0.006)	
West–South Central	21,570.96	21,555.19	−0.015	0.40	0.17	900
	(443.11)	(367.30)	(0.022)	(0.0001)	(0.007)	
Mountain	21,132.11	20,512.80	0.027	0.38	0.17	318
	(663.16)	(636.55)	(0.044)	(0.0001)	(0.011)	
Pacific	26,503.12	24,931.35	0.038	0.36	0.20	891
	(649.21)	(501.76)	(0.025)	(0.0001)	(0.009)	

Note: Numbers are for workers between ages 18 and 65 who usually worked 30–65 hours a week. Numbers in parentheses are standard errors except in column (4), where they are p-values.

The numbers in panel B show no systematic relationship between the difference in the two earnings measures and plant location. While the mean difference between the two earnings measures varies between −0.037 and 0.038, 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 11.3 breaks out the numbers presented in table 11.1 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 11.3 shows that the log difference between the measures of worker earnings varies from a high of 0.24 for tobacco to a low of −0.13 for petroleum refining. 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 percent 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 percent significance level. Viewed as a whole the numbers in tables 11.1, 11.2, and 11.3 suggest that workers are being matched to the correct establishments.

11.3.2 Examining the Representativeness of the Data

To begin examining whether the WECD data are representative of the underlying population of workers and plants, table 11.4 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 (col. [2]/col. [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 11.4 shows that of the 3,176,986 manufacturing workers in the SDF, 199,558 appear in the WECD, a match rate of 6 percent. The numbers in column (3) show that this match rate varies by industry. Tobacco, paper, leather, and primary metals all have match rates of 10 percent or greater, while lumber, instruments, and miscellaneous all have match rates of 3 percent. The numbers in column (6) show that matched workers average 10 percent higher wages than all SDF workers but that the size and sign of this difference varies by industry. In 3 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 percent.

Table 11.5 presents the number and average employment for all SSEL plants, unique plants, and WECD plants, for all plants in the data (the total row) and

Table 11.3 **Comparing Matched Plant and Worker Data by Industry**

Industry	SSEL Worker Earnings (1)	SDF Worker Earnings (2)	Log Difference (3)	ρ(SSEL Earnings, SDF Earnings) (4)	Proportion Matched (5)	Number of Plants (6)
Food	24,055.82	23,750.41	−0.01	0.48	0.12	1,665
	(347.16)	(421.18)	(0.01)	(0.0001)	(0.003)	
Tobacco	22,557.58	26,785.83	0.24	0.68	0.08	25
	(2502.03)	(2020.56)	(0.09)	(0.0002)	(0.01)	
Textile	20,419.94	20,618.58	−0.03	0.46	0.13	438
	(561.06)	(660.45)	(0.03)	(0.0001)	(0.01)	
Apparel	15,462.98	16,470.58	0.02	0.33	0.13	559
	(380.04)	(544.22)	(0.03)	(0.0001)	(0.01)	
Lumber	20,039.38	23,254.54	0.08	0.27	0.19	572
	(460.79)	(912.31)	(0.03)	(0.0001)	(0.01)	
Furniture	20,047.37	22,125.10	0.02	0.42	0.19	379
	(421.61)	(996.03)	(0.03)	(0.0001)	(0.01)	
Paper	26,981.37	27,280.02	−0.04	0.50	0.10	866
	(303.99)	(525.90)	(0.02)	(0.0001)	(0.004)	
Printing	19,348.33	21,666.39	0.09	0.44	0.16	1,228
	(313.51)	(362.91)	(0.02)	(0.0001)	(0.01)	
Chemicals	30,598.58	30,012.29	−0.03	0.28	0.17	1,165
	(641.66)	(501.74)	(0.02)	(0.0001)	(0.01)	
Petroleum refining	37,282.11	33,492.94	−0.13	0.07	0.17	161
	(1,434.79)	(1,502.55)	(0.05)	(0.38)	(0.02)	

(*continued*)

Table 11.3 (continued)

Industry	SSEL Worker Earnings (1)	SDF Worker Earnings (2)	Log Difference (3)	ρ(SSEL Earnings, SDF Earnings) (4)	Proportion Matched (5)	Number of Plants (6)
Rubber	23,691.93	24,052.27	−0.03	0.45	0.12	717
	(467.37)	(467.37)	(0.02)	(0.0001)	(0.01)	
Leather	16,662.93	17,503.39	0.05	0.46	0.14	178
	(754.53)	(777.90)	(0.05)	(0.0001)	(0.01)	
Stone	26,068.61	25,288.76	−0.06	0.41	0.14	853
	(409.75)	(528.45)	(0.02)	(0.0001)	(0.01)	
Primary metals	26,942.87	27,624.96	−0.02	0.45	0.12	898
	(372.66)	(702.90)	(0.02)	(0.0001)	(0.005)	
Fabricated metals	26,287.79	26,299.20	−0.04	0.33	0.14	1,490
	(500.68)	(484.06)	(0.02)	(0.0001)	(0.005)	
Machinery	27,216.31	28,512.74	0.02	0.34	0.19	1,421
	(324.71)	(576.73)	(0.02)	(0.0001)	(0.01)	
Electrical equipment	23,467.39	25,601.72	0.06	0.40	0.13	726
	(394.61)	(608.20)	(0.02)	(0.0001)	(0.01)	
Transportation	26,112.19	26,212.33	0.01	0.52	0.17	715
	(455.76)	(534.98)	(0.02)	(0.0001)	(0.01)	
Instruments	28,540.42	29,043.37	0.02	0.18	0.17	257
	(1,049.58)	(950.43)	(0.05)	(0.0041)	(0.02)	
Miscellaneous	20,423.02	22,959.16	0.07	0.26	0.17	538
	(427.49)	(696.47)	(0.03)	(0.0001)	(0.01)	

Note: Numbers are for workers between ages 18 and 65 who usually worked between 30–65 hours a week. Numbers in parentheses are standard errors, except in col. (4), where they are *p*-values.

Table 11.4 **Number and Mean Earnings of SDF and WECD Workers by Industry**

Industry	SDF Workers (1)	WECD Workers (2)	Proportion Matched (3)	Mean Earnings of SDF Workers (4)	Mean Earnings of WECD Workers (5)	Log Difference (6)
Food	231,420	20,597	0.09	22,131	23,619	0.07
Tobacco	7,393	1,379	0.19	35,899	35,890	0.00
Textile	121,159	6,485	0.05	18,307	19,228	0.05
Apparel	161,014	6,255	0.04	13,946	14,722	0.05
Lumber	134,031	3,856	0.03	18,214	26,448	0.37
Furniture	92,274	3,217	0.04	18,576	20,482	0.10
Paper	106,615	14,411	0.14	29,322	31,217	0.06
Printing	282,069	11,510	0.04	23,143	21,154	−0.09
Chemicals	176,282	12,089	0.07	33,342	33,183	0.00
Petroleum	27,194	1,913	0.07	36,301	37,633	0.04
Rubber	109,594	8,608	0.08	23,484	25,854	0.10
Leather	24,484	2,442	0.10	16,025	16,606	0.04
Stone	88,855	6,666	0.08	24,271	26,167	0.08
Primary metals	126,963	17,224	0.14	28,897	31,854	0.10
Fabricated metals	185,281	13,435	0.07	25,108	27,417	0.09
Machinery	373,079	17,313	0.05	28,804	31,515	0.09
Electrical equipment	281,519	14,633	0.05	27,810	25,342	−0.09
Transportation	379,002	30,622	0.08	32,035	35,379	0.10
Instrument	92,684	2,406	0.03	29,057	29,868	0.03
Miscellaneous	176,074	4,442	0.03	21,693	21,264	−0.02
Total	3,176,986	199,558	0.06	25,558	28,107	0.10

Table 11.5 **Number, Proportion, and Average Total Employment of All, Unique, and Matched Plants by Industry**

Industry	All SSEL Plants (1)	Unique Plants (2)	WECD Plants (3)	Proportion Unique (4)	Proportion Matched (5)	Average SSEL Plant Employment (6)	Average Unique Plant Employment (7)	Average WECD Plant Employment (8)
Food	19,117	6,598	1,801	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	5,838	1,804	466	0.31	0.08	112.0	124.4	161.4
Apparel	21,275	2,858	643	0.13	0.03	47.9	76.7	110.5
Lumber	31,573	3,845	657	0.12	0.02	22.2	31.3	52.5
Furniture	11,168	1,612	421	0.14	0.04	45.3	50.8	64.5
Paper	6,126	2,342	888	0.38	0.15	103.1	123.7	163.5
Printing	58,803	5,514	1,491	0.09	0.03	26.3	39.3	75.3
Chemicals	11,659	3,914	1,230	0.34	0.11	74.3	82.5	126.9
Petroleum refining	2,161	922	165	0.43	0.08	53.4	67.3	130.8
Rubber	14,435	2,884	752	0.20	0.05	60.8	93.1	155.0
Leather	1,897	767	198	0.40	0.10	62.2	76.0	118.1
Stone	15,245	4,368	931	0.29	0.06	34.2	44.4	80.0
Primary metals	6,548	2,843	934	0.43	0.14	109.7	130.9	222.1
Fabricated metals	35,513	6,742	1,580	0.19	0.04	41.7	61.3	121.6
Machinery	49,097	6,255	1,514	0.13	0.03	39.1	68.5	127.8
Electrical equipment	15,941	2,887	757	0.18	0.05	97.4	142.3	240.0
Transportation	10,002	3,170	762	0.32	0.08	180.7	241.9	448.4
Instrument	9,688	1,851	283	0.19	0.03	99.6	123.6	229.4
Miscellaneous	16,251	2,698	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

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), (2), and (3) present the number of SSEL plants, unique plants, and WECD plants, respectively. Column (4) presents the proportion of plants that are unique (col. [2]/col. [1]), while column (5) presents the proportion of plants in the WECD (col. [3]/col. [1]). Columns (6), (7), and (8) present the mean employment for all SSEL plants, unique plants, and WECD plants, respectively.

The total row in table 11.5 shows that of the 342,471 plants in the 1990 SSEL, 16,144 appear in the WECD, a match rate of 5 percent. 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 manner similar to the pattern seen in table 11.4. Tobacco, paper, leather, and primary 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 percent of plants in the SSEL are unique, but only 5 percent appear in the WECD. The numbers in columns (6), (7), and (8) show why this is the case. Comparing the average employment of unique plants with the average employment of all SSEL plants shows that unique plants are much larger than all 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, in press). Since WECD workers work in larger establishments than SDF workers they will in turn have higher average earnings.

Table 11.6 repeats the same analysis for workers found in table 11.4, this time broken out by census division. One thing to notice in table 11.6 is that the match rate is significantly lower in the Mountain and Pacific divisions. In the Pacific division only 2 percent of the workers in the SDF are matched to plants.

Table 11.7 repeats the same analysis for plants found in table 11.5, this time broken out by plant size (panel A) and census division (panel B). The numbers in panel A of table 11.7 confirm the fact that large plants are both more likely to be unique and more likely to appear in the WECD. Column (4) shows that

Table 11.6 **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 of SDF Workers (4)	Mean Earnings of WECD Workers (5)	Log Difference (6)
New England	189,131	17,673	0.09	28,781.95	22,822.79	0.00
Middle Atlantic	469,899	37,820	0.08	27,559.07	27,151.79	0.01
East–North Central	772,079	69,986	0.09	27,362.52	30,617.08	−0.05
West–North Central	276,567	18,682	0.07	23,049.96	26,582.73	−0.06
South Atlantic	479,648	20,263	0.04	22,508.84	25,788.60	−0.06
East–South Central	234,695	11,066	0.05	20,469.50	23,810.22	−0.07
West–South Central	293,049	12,234	0.04	23,764.57	23,212.54	0.01
Mountain	105,588	3,408	0.03	24,224.02	23,400.80	0.02
Pacific	356,322	8,426	0.02	28,571.62	33,644.64	−0.07

Table 11.7 **Number, Proportion, and Average Total Employment of SDF, Unique, and Matched Plants by Plant Size and Census Division**

	All SSEL Plants (1)	Unique Plants (2)	WECD Plants (3)	Proportion Unique (4)	Proportion Matched (5)	Average SSEL Plant Employment (6)	Average Unique Plant Employment (7)	Average WECD Plant Employment (8)
A. *Plant size (total employment)*								
1–9	161,192	24,765	2,924	0.15	0.02	4.1	4.1	5.0
10–24	74,981	12,944	3,088	0.17	0.04	15.5	15.7	16.2
25–49	41,796	8,415	2,687	0.20	0.06	34.9	35.2	35.9
50–99	28,877	7,014	2,821	0.24	0.10	70.1	70.8	71.2
100–249	22,599	6,401	2,673	0.28	0.12	154.2	155.8	156.3
250–499	7,973	2,259	1,091	0.28	0.14	345.8	347.9	346.5
500–999	3,378	1,197	526	0.35	0.16	679.3	680.2	683.3
1,000+	1,675	654	334	0.39	0.20	2,411.6	2,450.2	2,527.3
B. *Census division*								
New England	23,616	5,416	1,560	0.23	0.07	48.8	67.8	153.2
Middle Atlantic	54,657	12,063	3,667	0.22	0.07	46.4	70.4	116.2
East–North Central	65,381	13,629	4,526	0.21	0.07	59.3	95.6	165.8
West–North Central	23,252	5,478	1,308	0.24	0.06	56.2	84.7	153.5
South Atlantic	50,336	8,013	1,866	0.16	0.04	58.9	108.5	178.6
East–South Central	19,235	3,847	815	0.20	0.04	69.9	113.9	169.9
West–South Central	34,872	5,831	1,025	0.17	0.03	47.4	72.9	123.2
Mountain	15,868	2,553	385	0.16	0.02	38.6	63.7	111.7
Pacific	55,254	7,119	992	0.13	0.02	44.1	73.6	104.5

as plant size increases the probability that a plant is unique in an industry-location cell rises, from 0.15 for plants with 1–9 employees to 0.39 for plants with 1,000 or more employees. However, column (5) 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's being unique, rises from 0.12 for plants with 1–9 employees to 0.51 for plants with 1,000 or more employees (not in table).[23]

Similar to table 11.6, 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), (7), and (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.[24]

Tables 11.4 through 11.7 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 et al. 1991; Kulka et al. 1991) shows 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 nonimputed data this will also affect the representativeness of the WECD data.

These effects can be seen in table 11.8 and figure 11.1. Table 11.8 presents characteristics for all manufacturing workers in the SDF (col. [1]), for all manufacturing workers in the May 1988 Current Population Survey (CPS; col. [2]), and for all WECD workers (col. [3]). Figure 11.1 presents the educational distribution for SDF and WECD workers.[25] The numbers in table 11.8 show that workers in the WECD are not a representative sample of the entire population of manufacturing workers. A larger percentage of workers in the WECD are white, male, married, production workers than in either the SDF or the CPS.

23. This is computed as WECD plants/unique plants (col. [3]/col. [2]).

24. 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.

25. 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 fig. 11.1.

Table 11.8 Comparing the Characteristics of SDF, CPS, and WECD Workers

Characteristic	SDF Workers (1)	1988 May CPS Workers, Manufacturing (2)	WECD Workers (3)	WECD Workers Weighted (4)
Percentage male	66.9	65.4	70.1	66.9
Percentage non-Hispanic white	85.2	88.8	89.6	88.3
Percentage now married	67.3	66.7	71.0	67.7
Percentage in occupation				
Manager and professional	18.2	18.6	16.4	19.2
Technical, clerical. and sales	21.6	20.8	19.7	21.4
Production worker	60.2	60.6	64.0	59.4
Percentage 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	38.3	39.9	38.8
	(37)	(37)	(39)	(39)
Mean number of weeks worked[a]	47.5	–	48.9	48.2
	(52)		(52)	(52)
Mean usual hours worked per week[a]	41.2	41.0	41.7	41.3
	(40)	(40)	(40)	(40)
Mean wage or salary income[a]	25,558.1	–	28,106.7	25,676.8
	(21,000)		(25,000)	(25,000)
Mean hourly wage[a,b]	13.25	10.30	13.87	12.90
	(10.58)	(9.08)	(11.96)	(11.96)
N	3,176,986	4,757	199,558	1,639,556.2

Note: Numbers in parentheses are the medians of the distribution.

[a]Reference period is the previous year (1989) for SDF and WECD workers and the previous week for the CPS workers.

[b]For the SDF and WECD workers, hourly wage is estimated as: (wage or salary income / number of weeks worked) / usual hours worked per week.

Workers in the WECD are slightly older than workers in the SDF or the CPS and are more likely to be located in the Northeast and Midwest regions of the country. Table 11.8 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 11.1 shows that, relative to workers in the SDF, 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 et al. (1991) and Kulka et al. (1991) and are exactly what we would expect given that large plants are overrepresented in the WECD.

To make estimates of characteristics based on the data in the WECD more

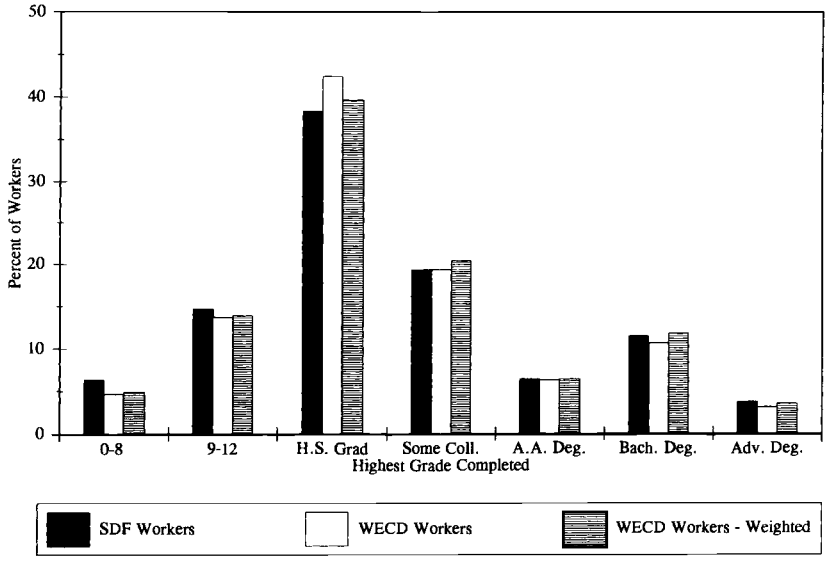

Figure 11.1 Educational distribution of SDF and WECD workers

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. First, I will discuss how I construct these weights.

As the discussion in section 11.2 shows, the probability that a plant appears in the data is a function of whether the plant is unique in an industry-location cell and 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 the two events separately. The product of the 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

$$P(u) = \mathbf{X}'\beta + u, \tag{1}$$

where $P(u)$ is the probability that a plant is unique in an industry-location cell, $\mathbf{X}$ is a vector of plant characteristics, and u is a normally distributed random error term. Results from tables 11.4 through 11.7 show that the probability that a plant is unique is related to plant size, industry, and location. Therefore, $\mathbf{X}$ includes 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 the plant is located in an urban area, $\mathbf{X}$ includes controls for whether the plant is located in a valid place (has a place code other than 9999) and the total population and the population per square mile for the county

where a plant is located.[26] Since I cannot directly observe $P(u)$ but instead only observe $P^*(u)$, where

$$P^*(u) = \begin{cases} 1 & \text{if a plant is unique,} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

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

$$P(m \mid u) = \mathbf{Y}'\gamma + \varepsilon, \tag{3}$$

where $P(m \mid u)$ is the probability that, conditional on being unique, a plant appears in the WECD, $\mathbf{Y}$ is a vector of plant characteristics, and ε is a normally distributed random error term. The results in tables 11.4 through 11.7 show that plant size also affects whether a plant contains matched workers. Therefore, (the log of) plant employment is included in $\mathbf{Y}$. Since the sampling frame of the SDF varied with the population of an area, $\mathbf{Y}$ 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 age 25 and its square, density of nonminority whites, and density of family households are also included in $\mathbf{Y}$ 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, $\mathbf{Y}$ includes a control for whether the plant is located in a valid place. Finally, $\mathbf{Y}$ includes controls for census division and two-digit industry. Again, since I do not directly observe $P(m \mid u)$ but instead observe $P^*(m \mid u)$, where

$$P^*(m \mid u) = \begin{cases} 1 & \text{if a plant is matched,} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

equation (3) is estimated using a probit model. Results from this estimation are available from the author.

Column (4) in table 11.8 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 11.1 includes the weighted educational distribution for WECD workers. The numbers in table 11.8 show that weighted estimates of worker characteristics are much closer to estimates of these characteristics based on the SDF data. The weighted cross-worker means of age, sex, race, marital status, occupation, and location are all much closer to the cross-worker means of these characteristics found in the SDF. The weighted means of number of weeks worked, usual hours worked last year, wage or

26. The latter two numbers are based on the 1980 decennial census.

Table 11.9 **Characteristics of SSEL Plants, Unique Plants, and WECD Plants**

Characteristic	All SSEL Plants (1)	Unique Plants (2)	Unique Plants Weighted (3)	WECD Plants (4)	WECD Plants Weighted (5)
Mean employment	52.2	84.5	60.2	146.3	63.3
	(11)	(16)	(16)	(43)	(43)
Mean annual payroll	1,414,237	2,377,177	1,688,294	4,411,189	1,731,777
	(199,000)	(312,000)	(312,000)	(943,000)	(943,000)
Average earnings	21,496	21,917	21,819	24,088	22,540
	(18,686)	(19,500)	(19,500)	(22,531)	(22,531)
Percentage in place	74.1	87.7	61.1	89.1	61.1
Percentage multiunit	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
Percentage 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
N	342,524	63,949	381,309.22	16,144	317,440.76

Note: Numbers in parentheses are medians of the distribution.

salary income, and hourly wage are also much closer to the values found in the SDF. Finally, figure 11.1 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 11.9 presents various characteristics for all manufacturing plants in the SSEL (col. [1]), unique plants (col. [2]), unique plants weighted by the inverse of the estimated probability of being unique (col. [3]), all plants in the WECD (col. [4]), and all WECD plants weighted by the inverse of the estimated probability that they appear in the WECD (col. [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 multiunit variables show that unique and WECD plants are more likely to be located in a place and to be part of a multiestablishment 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.

11.3.3 Replicating Previous Findings

While the results in tables 11.8 and 11.9 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 11.10 and 11.11. Table 11.10 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 percent 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 among 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 column (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 (in press) 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.[27]

The estimated relationships seen in table 11.10 are similar to previously reported relationships between experience, sex, marital status, race, education, and wages (Cain 1986; Korenman and Neumark 1991; Mincer 1974). For example, the coefficients on the female, black, and married variables and the female-married and female-black interactions show that women earn 17 percent less than men, black men earn 4 to 6 percent less than nonblack men, married men earn 13 percent more than single men, married women earn about the same as single women, and black women earn about the same as white women.

Table 11.11 presents the results from regressions of (log) average annual earnings in a plant on various plant characteristics, for all plants in the SSEL

27. Further evidence that this is true is given by the fact that the coefficients on the education variables are the only coefficients to change significantly between cols. (3) and (4).

Table 11.10 **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)
Experience	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)			
Adjusted R^2	0.50	0.50	0.51	0.47
N	704,373	704,373	185,186	185,007

Notes: These regressions only include workers who are between ages 18 and 65, who usually work 30–65 hours a week, and who have average wages between $2.50 and $100.00 an hour. Numbers in parentheses are standard errors.

(col. [1]), unique plants (col. [2]), unique plants weighted by the probability of being unique (col. [3]), WECD plants (col. [4]), and WECD plants weighted by the probability of appearing in the WECD (col. [5]). As in table 11.10, the coefficients on the various variables in table 11.11 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

Table 11.11 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)
Multiunit	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					
Adjusted R^2	0.23	0.26	0.23	0.32	0.31
N	234,694	49,735	49,698	15,138	15,137

Note: Numbers in parentheses are standard errors.

from zero). The coefficients on the three region variables also vary in sign and magnitude across the five regressions (although in all five regressions 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 of these plants are located in the West region.[28]

28. The fact that only the coefficients on the size and location variables change between the weighted and unweighted regressions provides further evidence that these characteristics are significant determinants of whether a plant appears in the WECD. Obviously, given that I am controlling for these characteristics in the unweighted regressions, the coefficients on the other variables should be unbiased estimates of the effect of these characteristics on wages and therefore will not change when estimating the weighted regression (assuming that they are uncorrelated with size or location).

The estimated relationships seen in table 11.11 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 1995), while the coefficients on the plant age variables show that older plants also pay higher wages (Brown and Medoff 1995; Dunne and Roberts 1990).

The results in tables 11.8 and 11.9 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 11.10 and 11.11 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.

11.4 What Can We Learn from the Worker-Establishment Characteristics Database?

11.4.1 The Establishment Size-Wage Premium

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.[29] 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 (e.g., Oi 1983, 1990; Hamermesh 1980, 1993; Dunne and Schmitz 1995), 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.

Consider the results in Troske (in press). Using the WECD data this paper examines a number of possible explanations for the employer size-wage premium. The main conclusion is that, while a significant portion of the size-wage premium is reduced once the fact that large plants are more capital intensive and employ more skilled workers has been controlled for, a majority of the premium remains unexplained. However, the primary importance of these results is that they represent the first attempt to account for the establishment size-wage premium in terms of both worker and employer characteristics.

29. For a complete discussion of the issues in this section, see Troske (in press).

11.4.2 Wages, Productivity, and Worker Characteristics

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, and the like, all hinge on the relationship between wages, productivity, and worker characteristics.[30] 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 problem 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, avoid these difficulties by allowing researchers to directly compare estimates of the relative wages of workers with estimates of workers' relative marginal productivity.

As an example, consider Hellerstein et al. (in press). 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 that (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. Although 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.

11.4.3 Technology Use and Worker Wages

While there has been growing interest among both economists and policymakers regarding the importance of skill-biased technical change in determining both the rate of return to education and the increasing wage differential between skilled and unskilled workers, there have been few microlevel studies that contain direct evidence on the effects of technical change on worker wages.[31] One of the principal reasons for this is the lack of data linking

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

31. For a complete discussion of the issues in the section, see Doms, Dunne, and Troske (1997).

a plant's use of advanced technology and the plant's demand for skilled labor. Linking the WECD with the plant-level data from the Census Bureau's Survey of Manufacturing Technology, which asks manufacturers about their use of advanced manufacturing technology in the plant, creates a data set that 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 et al. (1997). Results in this paper show that plants that use advanced technology capital in production 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.

11.5 Summary and Concluding Remarks

Results from examining the quality of the WECD are mixed. The results from section 11.3 show that, while a rather small percentage of workers and plants appear in the WECD, it does seem that workers are being matched to the correct establishments. The results from tables 11.8 and 11.9 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 important, the results in table 11.10 show that these data are capable of replicating relationships found in both the original data and in previous research based on alternative data sources. The latter finding in particular suggests that these data allow investigation of hypothesized relationships between worker and plant characteristics that are derived from theoretical models. Evidence on this point is found in section 11.4, 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 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 they should prove to be a valuable research tool.

One of the strongest conclusions that emerges from this analysis is that creating employer-employee matched databases requires very detailed information on which to base the match. The two major weaknesses of the WECD, the fact that it is a nonrandom sample and the fact that it only contains data for

manufacturing workers and employers, are a direct result of not having detailed place-of-work information. 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.

While more detailed name and address information for both workers and employers was collected, it was not possible to use this information when constructing the WECD because the name and address information for workers' employers was destroyed prior to starting the WECD project. However, in the future this information will be saved 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 databases and to extend these databases to nonmanufacturing workers and employers.

Appendix

Worker Variables Available from the Worker-Establishment Characteristics Database

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 1 April, 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)
Weeks 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 Longitudinal Research Database

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 man-hours
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

References

Bates, Nancy, Robert E. Fay, and Jeffery C. Moore. 1991. 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.

Brown, Charles, and James Medoff. 1989. The employer size-wage effect. *Journal of Political Economy* 97 (November): 1027–59.

———. 1995. Firm age and wages. Ann Arbor, University of Michigan, March. Unpublished paper.

Cain, Glen. 1986. The economic analysis of labor market discrimination: A survey. In *Handbook of labor economics*, ed. Orley C. Ashenfelter and Richard Layard. Amsterdam: North-Holland.

Doms, Mark, Timothy Dunne, and Kenneth R. Troske. 1997. Workers, wages, and technology. *Quarterly Journal of Economics* 112 (February): 253–90.

Dunne, Timothy, and Mark Roberts. 1990. Plant, firm and industry wage variation. Norman: University of Oklahoma, December. Unpublished paper.

Dunne, Timothy, and James A. Schmitz. 1995. Wages, employment structure and employer size-wage premia: Their relationship to advanced-technology usage at U.S. manufacturing establishments. *Economica* 62:89–107.

Griliches, Zvi. 1969. Capital-skill complementarity. *Review of Economics and Statistics* 51:465–68.

———. 1970. Notes on the role of education in production functions and growth accounting. In *Education, income, and human capital*, ed. W. Lee Hansen. New York: Columbia University Press.

Hamermesh, Daniel S. 1980. Commentary. In *The economics of firm size, market structure and social performance*, ed. John J. Siegfried. Washington, D.C.: Federal Trade Commission.

———. 1993. *Labor demand.* Princeton, N.J.: Princeton University Press.

Hellerstein, Judith K., David Neumark, and Kenneth R. Troske. In press. Wages, productivity, and worker characteristics: Evidence from plant-level production functions and wage equations. *Journal of Labor Economics.*

Korenman, Sanders, and David Neumark. 1991. Does marriage really make men more productive? *Journal of Human Resources* 26 (2): 282–307.

Kulka, Richard A., Nicholas A. Holt, Woody Carter, and Kathryn L. Dowd. 1991. 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.

Mincer, Jacob. 1974. *Schooling, experience, and earnings.* New York: Columbia University Press.

Oi, Walter Y. 1983. The fixed employment costs of specialized labor. In *The measurement of labor costs,* ed. Jack E. Triplett. Chicago: University of Chicago Press.

———. 1990. Employment relations in dual labor markets ("It's nice work if you can get it"). *Journal of Labor Economics* 8 (January): S124–S149.

Rosen, Sherwin. 1986. The theory of equalizing differences. In *Handbook of labor economics,* ed. Orley C. Ashenfelter and Richard Layard. Amsterdam: North-Holland.

Troske, Kenneth R. In press. Evidence on the employer size-wage premium from worker-establishment matched data. *Review of Economics and Statistics.*

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

———. 1992a. 1990 Census of Population and Housing—Classified index of industries and occupations. Washington, D.C.: U.S. Bureau of the Census.

———. 1992b. 1990 Census of Population and Housing—Guide part A. Text. Washington, D.C.: U.S. Bureau of the Census.

Willis, Robert. 1986. Wage determinants: A survey and reinterpretation of human capital earnings functions. In *Handbook of labor economics,* ed. Orley C. Ashenfelter and Richard Layard. Amsterdam: North-Holland.