The 1990 Decennial Employer-Employee Dataset

Kimberly Bayard, Joel Elvery, Judith Hellerstein and David Neumark*

September 2002

* Respectively, Board of Governors of the Federal Reserve, University of Maryland and U.S. Census Bureau, University of Maryland and NBER, Michigan State University and NBER We thank Jennifer Foster and Nicole Nestoriak for outstanding research assistance. We also thank James Monahan of the Office of the Chief Economist of the Bureau of the Census for extracting the Write-In data. This research was supported by NSF grant SBR95-10876 and the Russell Sage Foundation, through the NBER. The research in this paper was conducted while the authors were research associates at the Washington, DC, RDC. Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the Bureau of the Census.

Abstract:

We describe the construction and assessment of a new matched employer-employee data set, the 1990 Decennial Employer-Employee Dataset (1990 DEED). By using place of work name and address, we link workers from the 1990 Long Form Sample to their place of work in the 1990 Standard Statistical Establishment List. The resulting data set is much larger and more representative across regional and industry dimensions than previous matched data sets for the United States. The known strengths and limitations of the data set are discussed in detail.

Keywords: Matched employer-employee data, Decennial Census, Worker Establishment Characteristics Database (WECD)

Fifteen years ago, data sets matching employees with their employers were virtually nonexistent. The importance of these data sets was well understood, however, as highlighted by two authors in the original Handbook of Labor Economics (Ashenfelter and Layard, 1986). Robert Willis wrote that the study of wage determination "will hinge crucially on the development of data which links information on the individual characteristics of workers and their households with data on the firms who employ them" (1986, p. 589). And Sherwin Rosen wrote, that "on the empirical side ... the greatest potential for further progress rests in developing more suitable sources of data on the nature of selection and matching between workers and firms" (1986, p. 688).

Fortunately, since then matched employer-employee data sets have been created, first outside and then more recently in the United States. Indeed, by the time the more recent volumes of the Handbook of Labor Economics were published in 1999 (Ashenfelter and Card, 1999), there was enough research using these data sets to merit a full chapter (see Abowd and Kramarz, 1999).

This paper documents the construction and evaluation of a new U.S. matched employer-employee data set, based on the Decennial Census of Population for 1990. The key innovation in this data set – which we call the 1990 DEED (Decennial Employer-Employee Dataset) – is that we match workers to establishments by using the actual written worker responses to the question asking respondents to list the business address of their employer in the week prior to the Census. These responses are matched to a Census Bureau file containing business address information for all establishments in the United States.

The resulting 1990 DEED data set is very large, containing information on 3,839,904 workers matched to 1,166,571 establishments[1], which account for 22% of all workers who received the Long Form of the Decennial Census and 18% of active establishments in the Standard Statistical Establishment List (SSEL), an administrative database containing information for all business establishments operating in the U.S. in 1990. As it stands, it is the largest national matched employer-employee database covering the United States that contains detailed demographic information on workers, making it a rich source of

[1]These numbers are prior to sample restrictions imposed in the empirical analysis, as described below.

information for studying a variety of questions of interest to labor economists, demographers, and others.[2] It also serves as a model for constructing a similar database using the 2000 Decennial Census, which the authors plan to construct when the data is available.

Previous Matched Data Using the 1990 Decennial Census

In past research, we have used and/or created two more-limited matched data sets based on the 1990 Census of Population. The first data set we used covers manufacturing only, and is called the Worker-Establishment Characteristics Database (WECD). The second, which we created, covers all industries, and is called the New Worker-Establishment Characteristics Database (NWECD). The matched WECD and NWECD data sets are constructed from two data sources: the 1990 Sample Edited Detail File (SEDF), which contains all individual responses to the 1990 Decennial Census one-in-six Long Form; and the 1990 SSEL. The WECD and NWECD were created by using the detailed industry and location information for employers available in both the 1990 SEDF and the 1990 SSEL to link workers to their employers. For brevity's sake, the data we refer to is from 1990 unless otherwise noted.

The WECD and NWECD have proven very valuable. After describing the construction of these data sets, we briefly discuss some of the previous work we have conducted using them.[3] However, we also discuss some important limitations of the WECD and NWECD, and how they are ameliorated in the DEED.

Households receiving the 1990 Decennial Census Long Form were asked to report the name and address of the employer in the previous week for each employed member of the household. In addition, respondents were asked for the name and a brief (one or two word) description of the type of business or

[2]Another national matched employer-employee data set currently under construction at the U.S. Census Bureau is the Longitudinal Employer Household Database (LEHD). The LEHD is very rich in that it contains observations on all workers in covered establishments (not limited to the 1-in-6 sample of Long-Form respondents) and is longitudinal in nature. As of now, however, the LEHD does not contain detailed demographic information on workers, and only covers a handful of states (although some of the largest ones). In addition, it matches workers to firms rather than establishments, so that workers can only be matched to establishments when the establishment is not part of a multi-unit firm.

[3]See Troske (1998) for a more thorough discussion of the construction and representativeness of the WECD; and Bayard, et al. (2000) for an analogous description of the NWECD.

industry of the most recent employer for all members of the household. Based on the responses to these questions, the Census Bureau assigned geographic and industry codes to each record in the data and it is these codes that are available in the SEDF.

The SSEL is an annually-updated list of all business establishments with one or more employees operating in the U.S. The Census Bureau uses the SSEL as a sampling frame for its Economic Censuses and Surveys, and continuously updates the information it contains. Amoung other items, the SSEL contains the name and address of each establishment, geographic codes based on its location, its four-digit SIC code, and an identifier that allows the establishment to be linked to other establishments that are part of the same enterprise, and allows the SSEL data to be linked to other Census Bureau establishment- or firm-level data sets that contain more-detailed employer characteristics.[4,5]

Matching workers to employers to create the WECD and the NWECD proceeded in four steps. First, the geographic and industry codes in the SEDF and the SSEL were standardized. Next, all establishments that were unique in an industry-location cell were selected. Third, all workers who indicated they worked in the same industry-location cell as a unique establishment were matched to the establishment. Finally, all matches based on imputed data were eliminated . The WECD is also matched to data from the Census of Manufactures, which provides the ingredients necessary to estimate production functions, but restricts the data set to manufacturing plants.

Using the WECD, Hellerstein, et al. (1999) examine the relationships between productivity, wages, and worker characteristics in the manufacturing sector to test for discrimination and other deviations from equality between wages and marginal products. The unique contribution of the matched

---

[4]In both the SEDF and the SSEL the level of detail of the geographic codes depends on the location of the employer. In metropolitan areas, the Census Bureau assigns codes which identify an employer's state, county, place, tract, and block. A block is the smallest geographic unit defined by the Census in the SEDF and the SSEL. A typical block is that segment of a street that lies between two other streets, but could also be a street segment that lies between a street and a "natural" boundary such as a river or railroad tracks. A tract is a collection of blocks. In non-metropolitan areas, the Census Bureau defines tracts as "Block Numbering Areas" (BNAs), but for our purposes tracts and BNAs are equivalent. A Census designated place is a geographic area or township with a population of 2,500 or more.

[5]The discussion in this paper regarding the content of the SEDF and SSEL is intended to be an overview, not a comprehensive evaluation.

data in this research is to complement commonplace estimates of wage gaps by, e.g., race and sex, with production function estimates of productivity gaps by race and sex. This permits, for example, a test for sex discrimination in wages based on whether the wage gap exceeds the productivity gap (which it does).[6]

The WECD is also used in Hellerstein, et al. (2002) to examine the relationship between profitability and worker characteristics, to test the simple prediction of the neoclassical model of discrimination (Becker, 1971) that firms that hire more women or blacks are more profitable. Our paper also uses longitudinal data on establishments (but not workers) to examine the relationship between growth and workforce characteristics, to test whether non-discriminating employers appear to outcompete their rivals in product markets, consistent with the view that market competition roots out discrimination. The results from this paper indicate that firms that hire more women are indeed more profitable, consistent with discrimination. But among establishments with the largest market shares, which presumably operate in less-competitive product markets, discriminating firms are not "punished" by the market, suggesting that market competition alone is insufficient to counter discrimination.

Finally, Bayard, et al. (1999 and forthcoming) use the NWECD (covering all industries) to estimate the shares of racial, ethnic, and sex differences in wages that can be attributed to segregation across occupations, industries, establishments, and establishment-occupation (job) cells. This evidence speaks directly to the relative importance of equal pay and equal opportunity (including affirmative action) in breaking down pay gaps by race, sex, and ethnicity in U.S. labor markets.

While the WECD and NWECD have yielded unique results and new methods of studying labor market discrimination and other issues, there are a few shortcomings of these data sets that are of serious concern. Because the match is based on geographic and industry codes, in order to ensure that we link workers to the correct employers we only match workers to establishments that are unique in an industry-location cell. This substantially reduces the number of establishments available for matching. Of the 5.5 million establishments in the SSEL with positive employment, only 388,787 are unique in an industry-

[6]We have also pursued this question in data on manufacturing establishments in Israel, although with data that do not permit disaggregation among workers in an establishment (Hellerstein and Neumark, 1998 and 1999).

location cell. Once we match to workers, and impose a few other sample restrictions to improve the accuracy of the data, we end up with a data set covering 1,117,424 workers in 156,332 establishments, which covers 6% of all workers in the SEDF, and 2% of all establishments in the SSEL.[7] Second, although this is still a very large data set, matching on location and industry codes affects the representativeness of the resulting matched data. Establishments in the WECD and NWECD are larger and are more likely to be located outside of a metropolitan statistical area (MSA) than the typical establishment in the SSEL. In addition, relative to workers in the SEDF, workers in the matched data are more likely to be white and married, are slightly older, and have different patterns of education. Finally, because manufacturing establishments are more likely to be unique to an industry-location cell (consider a factory compared with a retail clothing outlet in a mall), they are considerably over-represented in the NWECD.

Overview of the 1990 DEED

To address these deficiencies, we have developed an alternative method to match workers to employers that does not require establishments and workers to be located in unique industry-location cells. Instead, this method relies on matching the actual employer name and address information provided by respondents to the Decennial Census to name and address information available for employers in the SSEL. This methodology produces a matched data set that is much larger and more representative than the WECD or the NWECD.[8]

When the NWECD was created, the file that contained the keyed-in workplace name and address was unknown and unavailable to researchers. Subsequently, we were able to help track down this file and to participate in its conversion from an internal Census Bureau input/output language to a readable format. Because this name and address file had been used solely for internal processing purposes, it did

[7] Again, these numbers are prior to sample restrictions imposed in the analysis.

[8] Because the WECD contains only manufacturing establishments, while the DEED and the NWECD cover all industries, in the remaining discussion we focus only on comparing the latter two data sets.

not have an official name, but was informally known as the "Write-In" file. We have retained this moniker for reference purposes.

The Write-In file contains the information written on the questionnaires by Long-Form respondents, but not actually captured in the SEDF. For example, on the Long Form workers are asked to supply the name and address of their employers. In the SEDF, this information is retained as a set of geographic codes (state, county, place, tract, block), and the employer name and street address is omitted entirely. The Write-In file, however, contains the geographic codes as well as the employer's actual business name and address. Because name and address information is also available for virtually all employers in the SSEL, nearly all of the establishments in the SSEL that are classified as "active" by the Census Bureau are available for matching.

We can therefore use employer names and addresses for each worker in the Write-In file to match the Write-In file to the SSEL. Additionally, because both the Write-In file and the SEDF contain identical sets of unique individual identifiers, we can use these identifiers to link the Write-In file to the SEDF. This procedure potentially yields a much larger matched data set, and one whose representativeness is not compromised by the need to focus on establishments unique to industry-location cells.

Table 1 summarizes the type of information available on each file, and graphically displays the way the files are matched together and the resulting information contained in the DEED. As noted above, for virtually all establishments in the U.S., the SSEL contains basic establishment-level information including industry, geography, total employment, payroll, and an indicator for whether the establishment is a single-unit enterprise or part of a multi-unit firm. In addition, the SSEL contains a unique establishment identifier that can be used to match the establishment to any number of Census Bureau surveys that include the establishment. The SEDF contains an almost complete set of responses provided by all Long-Form respondents. The breadth and level of detail of the questions provide a broad source of information about each worker. Among the individual-level information contained in the Long Form are standard basic demographic characteristics (e.g., gender, age, race/ethnicity, education), earnings, hours

worked, industry, occupation, language proficiency, and immigrant status and cohort. In addition, the SEDF contains detailed geographic information about an individual's residence and workplace. Because the DEED links the SSEL and the SEDF together, we can assemble characteristics of the workforce of an establishment. The opportunity to compare and contrast the earnings and characteristics of workers both within and across employers is one of the most useful and unique features of matched employer-employee data sets in general, and the DEED in particular.

Before we can begin to link the three files together, we must first select valid observations from each file and organize them to facilitate matching. For workers, this is easy. We first match the Write-In files and the SEDF together based on the set of unique individual identifiers the two files have in common. As a practical matter, this is done on a state-by-state basis because the large SEDF and Write-In files are each comprised of 51 sub-files – one for each state and the District of Columbia.

We then select the records for all individuals who indicated that they worked, and who included any information about the identity of their employers. That is, even if the workers provide only an employer name and city, we still attempt to match the worker. Although we would increase the percentage of workers matched if we imposed stricter criteria on the individuals to be matched (e.g., requiring workers to include all address elements to be eligible for matching), we nonetheless attempt to match all possible workers, but impose strict criteria to make sure that workers who provide sparse information about the locations of their workplaces are matched correctly. Once we link the SEDF and Write-In files together and retain "matchable" observations, we output a new series of 51 state-specific files based on the location of each worker's employer. These 51 files contain the records that we attempt to match to the SSEL.

The selection of valid establishment observations from the SSEL is not as straightforward as the selection of worker records. The SSEL is the sampling frame for all establishment survey programs of the Census Bureau, and covers all businesses except those in private households and some government entities. "Businesses" are defined as legal or administrative entities that are assigned an Employer Identification Number (EIN) by the Internal Revenue Service; a single business may have many

establishments. Not all industries in the SSEL fall under the purview of Census Bureau surveys – those that do not are called "out-of-scope." Out-of-scope industries include many agricultural industries, urban transit, the U.S. Postal Service, private households, schools and universities, labor unions, religious and membership organizations, and government/public administration. The Census Bureau does not validate the quality of SSEL data for businesses in out-of-scope industries; for example, for some local governments the SSEL may contain only a single, consolidated observation that is intended to cover several establishments, while for others the coverage is more complete. We therefore eliminate all out-of-scope establishments, accounting for a 5.6% reduction in the total number of SSEL records. We also exclude establishments that are located outside of the United States, that are associated with an administrative entity, have zero or missing payroll, or have internal processing flags that indicate the record to be invalid.[9] The SSEL is maintained in two separate files: one for single-unit enterprises; and one for establishments that are part of multi-unit firms. We perform the relevant restrictions on each file, and when necessary rename relevant variables to maintain consistency across the two files. Finally, we combine the two files.

Matching Workers and Establishments

Once we have selected valid worker and establishment observations, we can begin to match worker records to their establishment counterparts. To match workers and establishments based on the Write-In file, we use MatchWare – a specialized record linkage program – to link the records. MatchWare is comprised of two parts: a name and address standardization mechanism (AutoStan); and a matching system (AutoMatch). This software has been used previously to link various Census Bureau data sets (Foster, et al., 1998; Miranda de Larra, 2002).

Our method to link records using MatchWare involves two basic steps. The first step is to use

[9] An additional issue was that there are occasionally multiple records for a given establishment. Often, these duplicate records occur because an establishment changed ownership during the year, so there is one SSEL record associated with each owner. Because we want to match a worker to only one establishment record, when we observe duplicate establishment records we select the record that is considered "active."

AutoStan to standardize employer names and addresses across the Write-In file and the SSEL. Standardization of addresses in the establishment and worker files helps to eliminate differences in how data are reported. For example, a worker may indicate that she works on "125 North Main Street," while her employer reports "125 No. Main Str." The standardization software considers myriad possibilities of different ways that common address and business terms can be written, and converts each to a single standard form.

Once the software standardizes the business names and addresses, each item is parsed into components. To see how this works, consider the case just mentioned above. The software will first standardize both the worker- and employer-provided addresses to something like "125 N Main St." Then AutoStan will dissect the standardized addresses and create new variables from the pieces. For example, the standardization software produces separate variables for the House Number ("125"), directional indicator ("N") , street name ("Main"), and street type ("St").[10] The value of parsing the addresses into multiple pieces is that we can match on various combinations of these components.

We supplemented the AutoStan software by creating an acronym for each company name, and added this variable to the list of matching components. We noticed that workers often included only the initials of the company for which they work (e.g., "ABC Corp"), while the business is more likely to include the official corporate name (e.g., "Albert, Bob, and Charlie Corporation").

The second step of the matching process is to select and implement the matching specifications. The AutoMatch software uses a probabilistic matching algorithm which accounts for missing information, misspellings, and even inaccurate information. This software also permits users to control which matching variables to use, how heavily to weight each matching variable, and how similar two addresses must be to be considered a match. AutoMatch is designed to compare match criteria in a succession of 'passes' through the data. Each pass is comprised of 'Block' and 'Match' statements. The Block statements list the variables that must match exactly in that pass in order for a record pair to be linked. In

[10]This example is provided for illustrative purposes only and does not demonstrate the full range of variables generated by the matching software. To learn more about the full range of possibilities, see the MatchWare documentation (MatchWare Technologies, Inc., 1997).

each pass, a worker record from the Write-In file is a candidate for linkage only if the Block variables agree completely with the set of designated Block variables on analogous establishment records in the SSEL. The Match statements contain a set of additional variables from each record to be compared. These variables need not agree completely for records to be linked, but are assigned weights based on their value and reliability.

As an example of how the Block and Match variables work, consider the case where we assign 'employer name' and 'city name' as Block variables, and assign 'street name' and 'house number' as Match variables. In this case, AutoMatch compares a worker record only to those establishment records with the same employer name and city name. All employer records meeting these criteria are then weighted by whether and how closely they agree with the worker record on the street name and house number Match specifications. The algorithm applies greater weights to items that appear infrequently. So, for example, if there are several establishments on "Main St." in a given town, but only one or two on "Mississippi St.," then the weight for 'street name' for someone who works on Mississippi St. will be greater than the 'street name' weight for a comparable Main St. worker. The employer record with the highest weight will be linked to the worker record conditional on the weight being above some chosen minimum weight. Worker records that cannot be matched to employer records based on the Block and Match criteria are considered residuals and we attempt to match these records on subsequent passes using different criteria.

It is clear that different Block and Match specifications may produce different sets of matches. Matching criteria should be broad enough to cover as many potential matches as possible, but narrow enough to ensure that only high probability matches are linked. Because the AutoMatch algorithm is not exact there is always a range of quality of matches, and we are therefore cautious in how we accept linked record pairs.

Our general strategy was to impose the most stringent criteria in the earliest passes, and to loosen the criteria in subsequent passes. We did substantial experimentation with different matching algorithms, and visually inspected thousands of matches as a guide to help determine cutoff weights. In total, we ran

16 passes. As displayed in Appendix Table A1, we obtained most of our matches in the earliest passes.

Fine-Tuning the Matching

In order to assess the quality of the first version of our national matched data set, we embarked on a project to manually inspect and evaluate the quality of a large number of randomly-selected matches. We first selected random samples of 1,000 worker observations from each of the five most populous states (CA, NY, TX, FL, IL) plus three other states (PA, MD, CO), which were chosen either because they provided ethnic and geographic diversity or because researchers had familiarity with the labor markets and geography of those states. We also chose from these eight states a random sample of 300 establishments and their 8,088 corresponding matched worker observations. In total, then, we manually checked 16,088 employer-employee matches, of which 15,009 were matches to in-scope establishments[11].

For each observation selected, we retained identifying information from both the Decennial Census (SEDF) and the SSEL, such as employer name and address, and industry and zip code, along with the round and pass numbers in which the match had been made by AutoMatch. Two researchers independently ranked the quality of each of the matches by comparing information from the SEDF and the SSEL and assigning a numerical score to the match on a scale of one to five as follows: 1=definitely a correct match, 2=probably a correct match, 3=not sure, 4=probably not a correct match, 5=definitely not a correct match. To give a sense of what the matched addresses look like and how they were scored by hand-checkers, in Appendix A we present hypothetical examples of matched addresses from the SEDF and SSEL and their hand-checked scores. These closely resemble randomly selected hand-checked matches from the actual data; due to confidentiality restrictions, we cannot provide actual examples. The examples in Appendix A should make it clear that scores of 1 and 2 were given by the hand-checkers for

[11] As we were constructing the DEED, a working group at the Census Bureau was revising the list of out-of-scope industries. We obtained the updated list of the Census Bureau's out-of-scope industries after matching, and deleted matches that were in industries new to this updated list. Interestingly, we discovered that two industries (colleges and universities, and religious organizations) that we had initially included as in-scope and that are actually out-of-scope had match rates that we considered to be "bad" as defined below. We only report results for the hand-checked observations that were in-scope.

only high-quality matches, so that the matching criteria we set in AutoMatch worked to minimize type-two errors.

There are a number of ways to evaluate the quality of our matching process given the results of the hand-checking. First, in Table 2, we show a 2-way frequency table of the hand-checked scores for this version of the data set. This table illustrates that our matching procedure generally worked well. Over 66% of the hand-checked observations received scores of 1 from both hand-checkers, and over 88% of the observations received scores no lower than 2 from both researchers.[12] Only 0.62% of matches received scores of five from both hand-checkers.

In order to refine our match, we examined the hand-checked observations more carefully. We coded each observation as an acceptable or not acceptable match, where an acceptable match was conservatively defined to be one which received a score of 1 or 2 from *both* researchers. We then examined the distribution of acceptable matches over various dimensions of the data and in multiple ways. Table 3 contains the results of linear probability regressions for the probability that a hand-checked observation was deemed to be an *unacceptable* match against a series of demographic variables as well as a few other variables that may help determine whether the match is good or bad. The demographic variables include a worker's age, sex, race/ethnicity, education, full-time status, and English speaking ability. The geographic variables included state indicators, dummy variables for whether or not the worker's employer is located in an MSA, whether or not the block or tract code is allocated (these codes are allocated by the Census Bureau when there is not enough information to assign them with a great degree of certainty), and interactions between the block and tract allocation variables and the MSA indicator. We also include industry dummy variables, and in some specifications occupation dummy variables.[13]

The results of the regressions in Table 3 indicate that only a few variables or sets of variables are

[12]The hand-checking was done by five different researchers.

[13]We also experimented with including establishment size dummy variables in the regressions. Match quality does vary systematically by establishment size, with large establishments having fewer poor matches. We have not yet adjusted the matching algorithm to improve match quality for small establishments.

quantitatively and statistically significantly related to the probability of a bad match. Perhaps most noticeable are the differences by industry. For example, in column (5), the probability of a bad match is 0.06 lower in manufacturing than in services (the omitted industry, where the average probability of a bad match is 0.12).[14] Aside from differences by industry, blacks have a 0.05 to 0.07 higher probability of being poorly matched than whites, and those with advanced degrees are also more likely than others to be poorly matched.

The regression results in Table 3 prompted us to investigate further the systematic differences in match quality by industry. We noted that industry differences were even more dramatic at a level of aggregation finer than the two-digit controls included in the regressions. Figure 1 shows histograms of the distribution of error rates across industries (where industry is defined by the three-digit Census Industry classification on the SEDF) in the sample of hand-checked observations. Recall that workers are determined to be matched in error if at least one of two scorers assigns a score of 3, 4, or 5 to the worker-establishment match. We tallied up the percentage of workers in each industry who appeared to be matched in error and weighted the industries by their overall employment share. The upper histogram in Figure 1 shows that more than 55% of all industries (employment-weighted) have an error rate of 0.10 or less.[15] It is clear from the distribution shown on the histogram that there are very few industries where the error rates are greater than 0.25. It is also worth noting that our definition of "error" is quite conservative, and that a match is deemed to be in error if even one of the researchers rating the match was "not sure" about its quality. In order to better observe the distribution of error rates across industries in the left-hand tail, the lower histogram in Figure 1 examines the distribution of error rates for those industries with an error rate of less than 20%.

Given the information that match quality was so strongly associated with industry, we refined our

---

[14]The sample size is a bit lower than in the frequency table (Table 2) because these regressions exclude 11 people in the military and use averages of hand-checked scores for observations that were selected both in the worker sample and in the establishment sample.

[15]These histograms could be based on either worker-reported or establishment-reported industry. The qualitative conclusions were very similar; here we show the former.

matching procedure by developing criteria to reduce errors by industry. We identified those industries that (1) had an estimated error (bad match) rate of 10% or more, and (2) represented at least 1% of employment in the entire matched national data set. The 10% rate was chosen because there does seem to be a reasonable drop in the frequency of bad matches at around that point in the distribution. There are four possible tabulations for each industry because both the worker and the establishment are assigned industry codes, and because we manually checked two separate files (one worker- and one establishment-based) of randomly-selected matches. We considered an industry to be problematic if it met the two criteria in any of the four tabulations. After additional inspection of the problematic industries, we then imposed correction procedures (discussed below) that included deletion of observations from the matched data set if certain criteria were met.

Table 4 lists the industries that in any one of the four tabulations had an estimated error rate of at least 10% and also comprised at least 1% of employment in the first version of the matched data set. The table shows the case with the highest proportion of "unacceptable" matches of the four possible tabulations for each industry. Table 4 also indicates how many of the tabulations identified the industry as problematic. There were 14 industries that met both criteria for identification as problematic.

For each of these industries, we re-examined the data to determine what systematic reasons, if any, led the quality of the matches to be low, and to find a remedy to the problems. For 7 of the 14 industries, we decided to restrict good matches to be those for which the industry code in the SEDF matched the industry code in the SSEL. This eliminated bad matches such as the following hypothetical example:

SEDF Business Address:	matched to	SSEL Business Address:
General Hospital		Private Cafeteria of General Hospital
1 Medical Drive		1 Medical Drive
Anytown, USA		Anytown, USA
Industry: 831 (hospitals)		Industry: 641 (eating and drinking places)

In this example, the business name provided by hospital employees in the SEDF is quite similar to the business name in the SSEL of the hospital's cafeteria. The hospital's SSEL record (not shown above) uses the hospital's actual legal name which reflects the name of the parent hospital chain

headquarters; the hospital's address is perhaps also described by the location of the main headquarters rather than the physical location of the hospital in "Anytown." Therefore, the closest match to hospital workers (and note that many parts of the business address do match) is the privately-owned cafeteria located on the grounds of the hospital. This is clearly a bad match, and selecting only those observations where the SEDF and SSEL industries match exactly eliminates this problem. It should be noted that we did not impose on the entire data set the restriction that SSEL and SEDF industries match because industry can be miscoded on both the worker and establishment files (see Bayard, 2001). For five other industries we restricted good matches to be those for which the five-digit zip codes from the SEDF and SSEL matched exactly. This was important in certain industries like grocery stores and banks, where establishments with the same name had multiple establishments in similar locations (such as large cities) but in different zip codes. Finally, for the remaining two industries (physicians' offices and clinics, and legal services) we modified the AutoStan program to parse out words in the establishment names differently from the standard way, since employees in these industries often report different establishment names than employers in a way which the standard algorithm in AutoStan does not handle well (e.g., an employee will write the establishment name in the SEDF as "Jones & Smith" while the employer's name in the SSEL is "Law Offices of John Jones and Jane Smith").

We applied each industry's restrictions and passed all of the data again through the AutoMatch procedure. From this second version of the national data set, we selected random samples from the 14 problematic industries of 100 workers and 30 establishments (all the workers matched to each establishment) in the same eight states examined earlier. In total, we reviewed 3,659 records for this second inspection. As before, two researchers independently scored each observation for match quality based on the scale given earlier. The results of this second round of checks indicated that we had substantially reduced the error rate in 8 of the 14 industries, but 6 industries still had error rates over 10% and comprised at least 1% of overall employment in the matched data. These industries are: Grocery stores (601); Eating and drinking places (641); Banking (700); Insurance (711); Real estate, including real estate-insurance offices (712); and Offices and clinics of physicians (812). Because we used a much

smaller sample in the second round, the error rates are less reliable. After examination of the second version of the national data set, there were no obvious correction procedures to reduce error rates in the 6 industries, and so we decided to retain this version of the data set as final. Data users should exercise caution when using matches in the 6 industries listed above; these industries account for 14% of the workers in the final version of the DEED and 17% of the establishments.

Evaluating the Representativeness of the Matched Data

To evaluate the representativeness of the matched DEED data set, it is useful to compare basic descriptive statistics from the DEED with their counterparts from the SEDF. In addition, to measure the degree to which the DEED is an improvement over the earlier data set, the NWECD, it is useful to examine basic statistics for this data set as well.

Table 5 displays comparisons of the means and standard deviations of an extended set of demographic characteristics from the SEDF, the DEED, and the NWECD. The first three columns show the means and standard deviations for workers in all data sets who are not excluded by the basic restrictions.[16] Column (4) displays the level differences between means for the DEED and the SEDF, while column (5) displays the level differences between means for the NWECD and the SEDF.

Out of all 12,143,183 workers in the SEDF who met the basic criteria, 3,291,213 (approximately 27%) are also in the DEED, a substantial improvement over the NWECD, which contains 904,589 workers who met similar criteria, or only 7% of all possible matches. The means of the demographic variables in both matched data sets are quite close to the means in the SEDF. For example, female workers comprise 46% of the SEDF, and 47% of both matched data sets. The distribution of workers across races and ethnicities is also relatively similar across the data sets. In the SEDF, white, Hispanic,

[16]We exclude individuals from the SEDF who did not work in the year prior to the survey year (1989) or were self-employed. We also dropped workers employed in an industry that was considered "out-of-scope" in the SSEL (see the earlier discussion of "out-of-scope"). These restrictions were more stringent than those used in the construction of the base sample of the NWECD, which is why the sample size for the NWECD in Table 5 is slightly smaller than reported in our previous work with the NWECD (e.g., Bayard, et al., 1999). The sample rules used, and their effects on sample size, are detailed in Appendix B.

and black workers account for 82, 7, and 8% of the total, respectively. The comparable figures for the DEED are 86, 5, and 5%; and in the NWECD, 87, 4, and 7%. Similarly, there is also a close parallel among the distributions of workers across education categories in all data sets.

The distribution of workers across industries paints a different picture. Because of the matching algorithm used, the NWECD was heavily over-representative of workers in manufacturing, and under-representative of retail workers. The DEED is not limited in the same way. Approximately 25% of all workers in the SEDF are employed in the manufacturing sector, and although this number is somewhat greater in the DEED (33%), it is substantially higher in the NWECD (49%). Retail workers comprise 20% of all workers in the SEDF, and 17% in the DEED, but only 9% of all NWECD workers.

The second half of the table, columns (6) through (10), displays summary statistics for full-time workers in the SEDF, DEED, and NWECD. The results are very similar to those for all workers, with means across demographic characteristics fairly similar across all three data sets, while the distribution of workers across industries in the DEED is much more similar to the underlying SEDF than is the distribution in the NWECD.

In addition to comparing worker-based means in all three data sets, it is useful to examine the similarities across establishments in the SSEL, the DEED, and the NWECD. Table 6 shows descriptive statistics for establishments in each data set as well as the level differences between the SSEL means and those from the matched data sets. There are 5,237,592 establishments in the SSEL; of these, 972,436 (19%) also appear in the DEED, and 137,735 (slightly more than 2.5%) are in the NWECD. Because only workers who are sent Decennial Census Long Forms are eligible for matching to their employers, it is far more likely that at least one worker in large establishments will be sent a Long Form, and consequently that that establishment is included in either the DEED or the NWECD. One can see evidence of the bias towards larger employers in both data sets by comparing the means across data sets for total employment. An average establishment in the SSEL has 18 employees, while the average establishment in the DEED has 53 workers, and establishments in the NWECD have, on average, 62 employees.

The distributions of establishments across industries in the DEED and NWECD relative to the SSEL are similar to those in the worker sample in the sense that the DEED is much closer to the SSEL. For example, although there is roughly the same share of Service establishments in all three data sets (28% in the SSEL, 26% in the DEED, and 26% in the NWECD), there is a far greater representation of manufacturing establishments in the NWECD (29%) than in the SSEL (6%) or the DEED (13%).

Examining the distribution of establishments across geographic areas also reveals that the DEED is more representative of the SSEL than is the NWECD. In both the SSEL and the DEED, just over 81% of establishments are in an MSA, while this is true for only 61% of NWECD establishments. Additionally, the distribution of establishments across Census regions is very similar in the SSEL and the DEED, while the NWECD distribution is not as similar to the SSEL.

Figure 2 displays a histogram that highlights one measure of the quality of the matched data set. We construct dual measures of average earnings per worker - one that uses information available in the SSEL and the other that uses SEDF-based information - and then show a histogram of the differences in these two measures. The SSEL-derived average earnings value is simply the log of each plant's payroll divided by its total employment. The SEDF-based average earnings measure is the sum of earnings for all workers matched to a given plant divided by the total number of matched workers; we then take logs of this ratio. The Figure shows that these two measures match up very closely. The bulk of the nearly symmetric distribution is tightly centered around zero, and the tails are quite short. The distribution of the difference in log average earnings provides encouraging evidence that workers are correctly matched to their employers. It also provides evidence that the workers in the DEED are representative of the other workers at the establishment who could not be matched[17].

Another important way to compare the representativeness of the matched data is to go beyond differences in means in the two data sets, and to examine regression relationships in the data sets to see whether the conditional relationships between variables are as similar as the unconditional summary

[17]We have constructed a similar histogram using levels rather than logs. The comments regarding Fig. 2 are equally true for the distribution of level differences in average earnings.

statistics. Table 7 presents results from a regression of the log of hourly wages on a set of demographic characteristics. We run two sets of regressions for each of the three data sets. Coefficient estimates and standard errors for all workers are shown in the first three columns, and for full-time workers only in the last three columns. Across all dimensions, the DEED coefficients are uniformly of the same sign, consistently close, and in some cases nearly identical to the SEDF coefficients. Although the NWECD estimates are also quite similar, there are a few cases where these coefficients diverge fairly notably from the SEDF (such as education, working in an MSA, female, and some industries and occupations).

Table 8 presents establishment-based regressions for the SSEL, the DEED, and the NWECD. We run two sets of regressions of average earnings per worker in an establishment on a set of plant characteristics. The first set includes one-digit industry effects, and the second set controls for industry at the three-digit level. Looking across all columns, we see that the coefficient estimates from the DEED match up fairly well with those from the SSEL. One interesting result is that the coefficient estimates for whether a plant is a single-unit establishment are more similar in the SSEL and the NWECD than in the DEED.[18] The DEED estimates for the industry dummies are generally much more similar to their SSEL counterparts than are the NWECD coefficients.

VI. Conclusions and Future Research

In this paper we document the construction the 1990 Decennial Employer-Employee Dataset, a large, new matched employer-employee data set for the United States. We have described in detail the matching process and outlined our efforts to refine the process to ensure that the resulting data set is research-quality. We have also shown that the DEED offers substantial improvements over its predecessor, the NWECD, in terms of the raw number of workers and establishments it contains as well as its representativeness.

The creation of the DEED allows researchers to address topics that previously could not be

[18]This result is reversed when we interact the single-unit dummy with the log of total employment and the log of squared total employment, suggesting that the "single-unit" effect is closely tied to the establishment size.

explored using the NWECD, the WECD, or even most other existing matched employer-employee data sets. For example, in Hellerstein and Neumark (2002), we use the DEED to examine the role of workplace segregation by Hispanic ethnicity and English proficiency in determining wages. In addition, the DEED is uniquely suited to examining the link between residential and workplace segregation, because as a match between a household data set and an establishment data set, the DEED contains information on residential address and workplace address of all workers. We plan to examine the relationship between residential and workplace segregation, and to examine the impact of this relationship on labor market outcomes (such as wages and employment) for workers of different races and ethnicities.

We have laboriously and carefully constructed the DEED in order to be able to further our research agenda and to assist the Census Bureau in meeting some of its objectives, and we have plans to construct the corresponding 2000 version of the DEED when the Long-Form information from the 2000 Decennial Census becomes available at the Census Bureau.[19] We are confident that the quality and scope of the data we have constructed will allow us to gain new insights into employer-employee relations, including the mechanisms and importance of segregation in the labor market.

[19]The 1990 DEED is the property of the United States Census Bureau, and while proprietary, is available to other researchers who meet the Census Bureau criteria for restricted-use data.

References

Ashenfelter, Orley C., and David Card, eds. 1999. Handbook of Labor Economics, Vols. 3A-3C (Amsterdam: Elsevier Science Publishers).

Ashenfelter, Orley C., and Richard Layard, eds. 1986. Handbook of Labor Economics, Vols. 1-2 (Amsterdam: Elsevier Science Publishers).

Abowd, John M., and Francis Kramarz. 1999. "The Analysis of Labor Markets Using Matched Employer-Employee Data." In Orley C. Ashenfelter and David Card, eds., Handbook of Labor Economics, Vol. 3B (Amsterdam: Elsevier Science Publishers), pp. 2629-710.

Bayard, Kimberly. 2001. "Measurement Error and Inter-Industry Wage Differentials." Mimeograph, Board of Governors of the Federal Reserve System.

Bayard, Kimberly, Judith Hellerstein, David Neumark, and Kenneth Troske. 2000. "The New Worker-Establishment Characteristics Database." Proceedings of the Second International Conference on Establishment Surveys (American Statistical Association), pp. 981-90.

Bayard, Kimberly, Judith Hellerstein, David Neumark, and Kenneth Troske. "New Evidence on Sex Segregation and Sex Differences in Wages from Matched Employer-Employee Data." Forthcoming in Journal of Labor Economics.

Bayard, Kimberly, Judith Hellerstein, David Neumark, and Kenneth Troske. 1999. "Why Are Racial and Ethnic Wage Gaps Larger for Men than for Women? Exploring the Role of Segregation Using the New Worker-Establishment Characteristics Database." In John C. Haltiwanger, Julia I. Lane, James R. Spletzer, Jules J.M. Theeuwes, and Kenneth R. Troske, eds. The Creation and Analysis of Employer-Employee Matched Data (Amsterdam: Elsevier Science B.V.), pp. 175-203.

Becker, Gary S. 1971. The Economics of Discrimination, Second Edition (Chicago: University of Chicago Press).

Foster, Lucia, John Haltiwanger, and C.J. Krizan. 1998. "Aggregate Productivity Growth: Lessons from Microeconomic Evidence." NBER Working Paper No. 6803.

Hellerstein, Judith and David Neumark. 2002. "Ethnicity, Language, and Workplace Segregation: Evidence from a New Matched Employer-Employee Data Set". NBER Working Paper 9037.

Hellerstein, Judith, and David Neumark. 1999. "Sex, Wages, and Productivity: An Empirical Analysis of Israeli Firm-Level Data." International Economic Review, Vol. 40, No. 1, February, pp. 95-123.

Hellerstein, Judith, and David Neumark. 1998. "Wage Discrimination, Segregation, and Sex Differences in Wages and Productivity Within and Between Plants." Industrial Relations, Vol. 37, No. 2, April, pp. 232-60.

Hellerstein, Judith K., David Neumark, and Kenneth R. Troske. 2002. "Market Forces and Sex Discrimination." Journal of Human Resources. Vol. 37, No.2, Spring, pp. 353-80.

Hellerstein, Judith K., David Neumark, and Kenneth R. Troske. 1999. "Wages, Productivity, and Worker Characteristics: Evidence from Plant-Level Production Functions and Wage Equations." Journal of Labor Economics, Vol. 17, No. 3, July, pp. 409-46.

Kain, John. 1968. “Housing Segregation, Negro Employment, and Metropolitan Decentralization.” Quarterly Journal of Economics, Vol. 82, May, 1968, pp. 175-97.

Lang, Kevin. 1986. “A Language Theory of Discrimination.” Quarterly Journal of Economics, Vol. 101, No. 2, May, pp. 363-82.

Lazear, Edward P. 1999. “Culture and Language.” Journal of Political Economy, Vol. 107, No. 6, Part 2, December, pp. S95-126.

MatchWare Technologies, Inc. 1997. AutoMatch 4.2 User Manual (Burtonsville, MD).

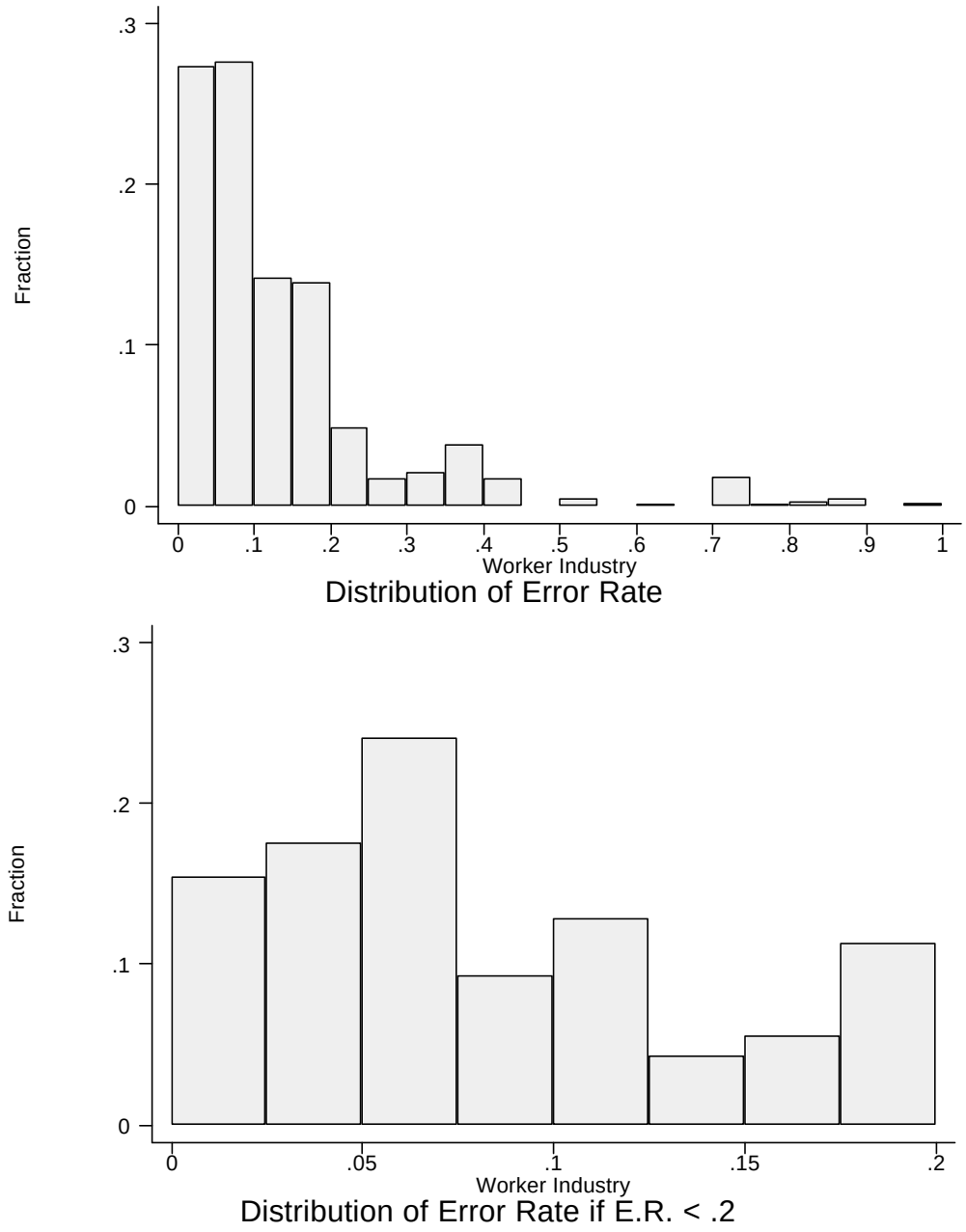
Miranda de Larra, Javier. 2002. “LBD Documentation: Auto Repair Shop Industry: Evaluation of Probabilistic Linking Algorithms”. Center for Economic Studies Technical Note 2002-02.

Rosen, Sherwin. 1986. “The Theory of Equalizing Differences.” In Orley C. Ashenfelter and Richard Layard, eds., Handbook of Labor Economics, Vol. 1 (Amsterdam: Elsevier Science Publishers), pp. 641-92.

Troske, Kenneth. 1998. “The Worker-Establishment Characteristics Database.” In John Haltiwanger, Marilyn E. Manser, and Robert Topel, eds. Labor Statistics Measurement Issues (Chicago: The University of Chicago Press), pp. 371-404.

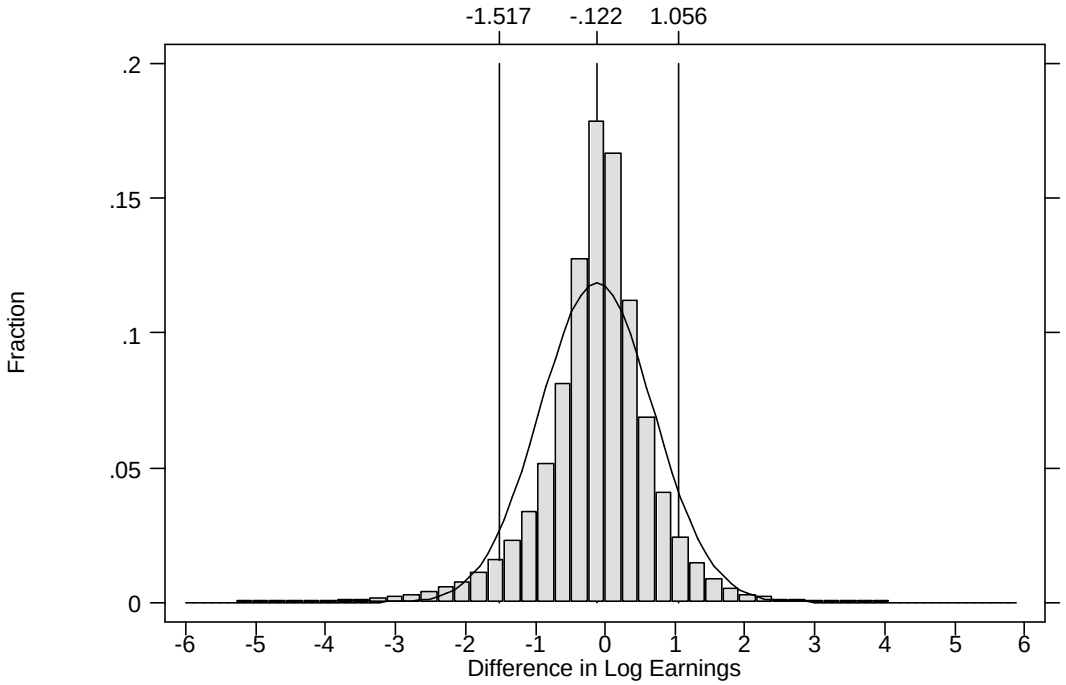
Willis, Robert J. 1986. “Wage Determinants: A Survey and Reinterpretation of Human Capital Earnings Functions.” In Orley C. Ashenfelter and Richard Layard, eds., Handbook of Labor Economics, Vol. 1 (Amsterdam: Elsevier Science Publishers), pp. 525-602.

Figure 1
Distribution of Industry Average Error Rates Based on First Round of Hand-Checking



Notes: These figures depict the distribution of industry average error rate weighted by the number of workers employed in the industry. The industry reported by the workers was used for these calculations.

Figure 2
Difference in Log Establishment Average Earnings Calculated from SSEL and DEED



Note: The graph shows the difference for each establishment between log average annual earnings calculated from the SSEL (payroll/total number of employees) and log average annual earnings calculated as the mean for all workers matched to the establishment in the DEED. The vertical lines indicate (from left to right) the lower 5% tail, the mean, and the upper 5% tail. These numbers were calculated for the whole sample of 972,436 establishments. There were 153 establishments where the mean difference in log earnings was either less than -6 or greater than 6 and are not on the graph.

Table 1

Linking the Three Files: Information Available in Each File

SSEL	Write-In File	SEDF
Business name and address <==> Business name and address		
	Unique person identifier <==>	Unique person identifier
Many characteristics: Industry Geographic location Total employment Payroll Indicator for whether the establishment is a single-unit enterprise or part of a multi-unit firm. Unique establishment identifier (can be used to match to other Census Bureau establishment- or firm- based data sets)	Limited demographic information for individual workers including each worker's: Occupation Industry	Demographic, and household, and labor market information, including: Sex Age Race/Ethnicity Education English language proficiency Earnings Hours Occupation and industry Immigration Similar information for other individuals in the household Detailed geographic information on worker's residence and workplace

Table 2
Two-Way Frequency of Hand-Checked Scores for All Hand-Checked Data from First Version of DEED

	Score B					
Score A	Definite Match	Likely Match	Not Sure	Unlikely Match	Not A Match	Row total
Definite Match	9,930 *66.16*	2,229 *14.85*	291 *1.94*	56 *0.37*	79 *0.53*	12,585 *83.85*
Likely Match		1,126 *7.50*	406 *2.71*	95 *0.63*	30 *0.20*	1,657 *11.04*
Not Sure			158 *1.05*	123 *0.82*	252 *1.68*	533 *3.55*
Unlikely Match				40 *0.27*	101 *0.67*	141 *0.94*
Not A Match					93 *0.62*	93 *0.62*
Column total	9,930 *66.16*	3,355 *22.35*	855 *5.70*	314 *2.09*	555 *3.70*	15,009 *100*

Note: Percent of sample in cell is reported in italics in the second entry of each box. We have recorded all non-matching scores above the diagonal.

Table 3
Linear Probability Estimates for Bad Match Quality as Functions of Worker Characteristics

	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error	Coefficient	Std. error
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	0.135	(0.032)	0.122	(0.032)	0.149	(0.033)	0.140	(0.033)
Age	-0.002	(0.001)	-0.001	(0.001)	0.000	(0.001)	0.000	(0.001)
$Age^2/100$	0.002	(0.002)	0.001	(0.002)	0.000	(0.002)	0.000	(0.002)
Full-time	-0.019	(0.007)	-0.020	(0.007)	-0.018	(0.007)	-0.017	(0.007)
Female	0.028	(0.005)	0.029	(0.005)	0.008	(0.005)	0.009	(0.006)
Black	0.069	(0.012)	0.055	(0.012)	0.050	(0.012)	0.047	(0.012)
Hispanic	-0.004	(0.011)	-0.003	(0.011)	-0.002	(0.011)	-0.002	(0.011)
Less than high school	-0.006	(0.010)	-0.008	(0.009)	0.001	(0.009)	-0.001	(0.009)
Some college	0.016	(0.007)	0.013	(0.007)	0.006	(0.007)	0.008	(0.007)
B.A.	0.022	(0.008)	0.017	(0.008)	0.001	(0.008)	0.004	(0.009)
Advanced degree	0.055	(0.010)	0.044	(0.010)	0.018	(0.011)	0.023	(0.011)
Speak English:								
Well	-0.007	(0.017)	-0.011	(0.017)	-0.005	(0.017)	-0.007	(0.017)
Poorly	0.009	(0.023)	0.005	(0.023)	0.019	(0.023)	0.016	(0.023)
Not at all	-0.048	(0.042)	-0.054	(0.042)	-0.035	(0.042)	-0.037	(0.042)
Work in MSA	0.021	(0.021)	0.007	(0.021)	0.001	(0.020)	0.002	(0.020)
No block	0.009	(0.049)	0.005	(0.049)	0.001	(0.048)	0.001	(0.048)
No tract	-0.023	(0.045)	-0.027	(0.045)	-0.021	(0.045)	-0.021	(0.045)
No tract × MSA	0.041	(0.060)	0.053	(0.060)	0.054	(0.059)	0.054	(0.059)
No block × MSA	-0.067	(0.062)	-0.053	(0.061)	-0.036	(0.061)	-0.037	(0.061)
State:								
California			-0.002	(0.011)	-0.007	(0.011)	-0.007	(0.011)
Colorado			-0.020	(0.011)	-0.024	(0.010)	-0.024	(0.010)
Florida			0.052	(0.011)	0.041	(0.011)	0.041	(0.011)
Maryland			0.046	(0.011)	0.038	(0.011)	0.039	(0.011)
New York			0.072	(0.011)	0.061	(0.011)	0.061	(0.011)
Pennsylvania			-0.030	(0.010)	-0.042	(0.010)	-0.042	(0.010)
Texas			0.007	(0.011)	0.003	(0.011)	0.004	(0.011)
Industry:								
Mining					-0.020	(0.029)	-0.017	(0.029)
Construction					-0.078	(0.013)	-0.075	(0.013)
Manufacturing					-0.062	(0.008)	-0.060	(0.008)
Transportation					0.004	(0.012)	0.005	(0.013)
Wholesale					-0.072	(0.011)	-0.070	(0.012)
Retail					-0.026	(0.008)	-0.025	(0.008)
FIRE					0.088	(0.009)	0.090	(0.009)
Occupation:								
Manager							-0.003	(0.007)
Service							0.020	(0.011)
Farming							0.122	(0.063)
Production							-0.001	(0.010)
Laborer							0.007	(0.009)
R^2	0.0104		0.0221		0.0428		0.0433	

Note: There are 14,954 observations. Sample includes all hand-checked observations on individuals employed and at work in 1990 in the United States. The omitted categories are: services (industry) and support (occupation).

Table 4

Scored Match Rates for "Problem" Industries

Industry number	Industry name	Proportion of matches coded "unacceptable"	Share of employment in the DEED	Number of times industry met "bad" criteria
641	Eating and drinking places	0.576	5.639	3
712	Real estate, including real estate-insurance offices	0.502	1.404	4
700	Banking	0.462	2.995	4
710	Security, commodity brokerage, and investment companies	0.344	1.153	2
812	Offices and clinics of physicians	0.268	1.839	4
601	Grocery stores	0.264	2.731	4
831	Hospitals	0.185	7.566	3
410	Trucking service	0.172	1.545	2
441	Telephone communications	0.154	1.041	2
711	Insurance	0.133	2.986	2
841	Legal services	0.127	1.380	2
832	Nursing and personal care facilities	0.118	1.482	1
591	Department stores	0.104	1.703	1
510	Professional and commercial equipment and supplies	0.100	1.001	1

Table 5

Means of Worker Characteristics

	All workers					Full-time workers				
	SEDF (1)	DEED (2)	NWECD (3)	DEED - SEDF (4)	NWECD - SEDF (5)	SEDF (6)	DEED (7)	NWECD (8)	DEED - SEDF (9)	NWECD - SEDF (10)
Age	37.08 (12.78)	37.51 (12.23)	38.61 (12.23)	0.42	1.52	37.69 (11.27)	37.87 (11.06)	38.96 (11.10)	0.18	1.26
Female	0.46	0.47	0.47	0.02	0.02	0.42	0.44	0.44	0.02	0.02
Married	0.60	0.65	0.68	0.04	0.07	0.64	0.67	0.70	0.03	0.06
Female× married	0.25	0.28	0.30	0.02	0.04	0.24	0.26	0.28	0.02	0.04
White	0.82	0.86	0.87	0.04	0.05	0.82	0.87	0.87	0.04	0.05
Hispanic	0.07	0.05	0.04	-0.01	-0.03	0.06	0.05	0.04	-0.01	-0.03
Black	0.08	0.05	0.07	-0.03	-0.01	0.08	0.05	0.07	-0.03	-0.01
Full-time	0.77	0.83	0.82	0.06	0.05	--	--	--	--	--
Number of kids (if female)	1.57 (1.62)	1.53 (1.55)	1.84 (1.65)	-0.04	0.27	1.57 (1.59)	1.51 (1.53)	1.83 (1.63)	-0.05	0.27
High school diploma	0.34	0.33	0.39	-0.01	0.05	0.35	0.34	0.40	-0.01	0.05
Some college	0.30	0.32	0.30	0.02	0.00	0.31	0.33	0.30	0.02	-0.01
B.A.	0.13	0.16	0.11	0.03	-0.02	0.14	0.17	0.11	0.02	-0.03
Advanced degree	0.05	0.05	0.04	0.01	-0.01	0.05	0.06	0.04	0.01	-0.01
Ln(hourly wage)	2.21 (0.70)	2.30 (0.65)	2.24 (0.64)	0.10	0.03	2.31 (0.58)	2.38 (0.57)	2.31 (0.54)	0.06	-0.01
Hourly wage	12.10 (82.19)	12.89 (37.07)	11.76 (25.77)	0.79	-0.34	12.22 (11.27)	12.98 (12.07)	11.73 (11.73)	0.76	-0.50
Hours worked in 1989	39.51 (11.44)	40.42 (10.37)	39.91 (10.18)	0.92	0.41	42.11 (6.12)	42.41 (6.03)	41.89 (5.56)	0.30	-0.21
Weeks worked in 1989	46.67 (11.05)	48.21 (9.35)	47.95 (9.70)	1.54	1.28	50.33 (4.25)	50.71 (3.71)	50.65 (3.79)	0.37	0.32
Earnings in 1989	22576 (26760)	25581 (29475	22485 (21232)	3005.4	-90.8	26465 (26852)	28559 (29336)	25280 (20804)	2093.5	-1185.2
Industry:										
Mining	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.01	0.00	0.00
Construction	0.07	0.04	0.00	-0.03	-0.07	0.07	0.04	0.00	-0.03	-0.07
Manufacturing	0.25	0.33	0.49	0.08	0.24	0.28	0.36	0.54	0.08	0.26
Transportation	0.08	0.05	0.07	-0.02	-0.03	0.08	0.06	0.08	-0.03	0.00
Wholesale	0.05	0.07	0.03	0.02	-0.02	0.06	0.08	0.03	0.02	-0.03
Retail	0.20	0.17	0.09	-0.03	-0.11	0.16	0.14	0.07	0.00	-0.07
FIRE	0.08	0.08	0.02	0.00	-0.07	0.09	0.09	0.02	0.00	-0.07
Services	0.26	0.24	0.28	-0.02	0.03	0.24	0.23	0.25	-0.01	0.01
Observations	12,143,183	3,291,213	904,589			9,375,086	2,725,599	742,188		

Note: Standard deviations are reported in parentheses. For columns 1-5, the sample corresponds to row d. in Appendix B.

For columns 6-10, the sample corresponds to row e. in Appendix B.

Table 6

Means for Establishments

	SSEL	DEED	NWECD	DEED - SSEL	NWECD - SSEL
Total employment	17.57	52.68	61.64	35.11	44.07
	(253.75)	(577.39)	(276.14)		
Establishment size:					
1 - 25	0.88	0.65	0.68	-0.24	-0.20
26 - 50	0.06	0.15	0.11	0.09	0.05
51 - 100	0.03	0.10	0.09	0.07	0.06
101 +	0.03	0.10	0.12	0.07	0.10
Industry:					
Mining	0.00	0.01	0.01	0.00	0.01
Construction	0.09	0.07	0.00	-0.02	-0.09
Manufacturing	0.06	0.13	0.29	0.07	0.23
Transportation	0.04	0.05	0.09	0.01	0.06
Wholesale	0.08	0.11	0.09	0.03	0.01
Retail	0.25	0.24	0.21	-0.01	-0.04
FIRE	0.09	0.10	0.04	0.01	-0.05
Services	0.28	0.26	0.26	-0.03	-0.02
In MSA	0.81	0.82	0.61	0.00	-0.21
Census Region:					
North East	0.06	0.06	0.04	0.00	-0.02
Mid Atlantic	0.16	0.15	0.14	0.00	-0.01
East North Central	0.16	0.20	0.23	0.04	0.07
West North Central	0.07	0.08	0.12	0.01	0.04
South Atlantic	0.18	0.16	0.14	-0.02	-0.04
East South Central	0.05	0.05	0.08	0.00	0.03
West South Central	0.10	0.10	0.11	0.00	0.01
Mountain	0.06	0.05	0.05	-0.01	-0.01
Pacific	0.16	0.15	0.10	-0.02	-0.07
Payroll ($1000)	397	1358	1519	961	1122
	(5064)	(10329)	(11155)		
Payroll/total employment	21.02	24.24	18.56	3.22	-2.46
	(1385.12)	(111.79)	(76.08)		
Share of employees matched	--	0.17	0.29	--	--
Multi-unit establishment	0.23	0.42	0.36	0.19	0.13
N	5,237,592	972,436	137,735		

Note: 55 establishments in the DEED sample do not have valid county data from the SSEL. For these 55, the workers reported place of work was used to determine MSA status. The sample corresponds to row d. of Appendix B for the DEED and NWECD and row e. for the SSEL.

Table 7

Log Wage Regressions with Aggregated Industry and Occupation Dummies

	All			Full-time		
	SEDF (1)	DEED (2)	NWECD (3)	SEDF (4)	DEED (5)	NWECD (6)
Intercept	0.746 (0.002)	0.721 (0.003)	0.676 (0.006)	0.681 (0.002)	0.687 (0.003)	0.709 (0.006)
Age	0.041 (.0001)	0.045 (.0001)	0.043 (.0003)	0.056 (.0001)	0.057 (.0002)	0.054 (.0003)
$Age^2/100$	-0.039 (.0001)	-0.042 (.0002)	-0.040 (.0003)	-0.055 (.0001)	-0.056 (.0002)	-0.053 (.0004)
Black	-0.061 (0.001)	-0.058 (0.001)	-0.063 (0.002)	-0.070 (0.001)	-0.062 (0.001)	-0.069 (0.002)
Hispanic	-0.093 (0.001)	-0.083 (0.001)	-0.073 (0.003)	-0.098 (0.001)	-0.083 (0.001)	-0.079 (0.003)
Married	0.080 0.000	0.071 (0.001)	0.079 (0.001)	0.083 0.000	0.071 (0.001)	0.073 (0.001)
High school diploma	0.104 (0.001)	0.101 (0.001)	0.124 (0.002)	0.129 (0.001)	0.122 (0.001)	0.136 (0.002)
Some college	0.184 (0.001)	0.182 (0.001)	0.201 (0.002)	0.208 (0.001)	0.202 (0.001)	0.214 (0.002)
Associates degree	0.257 (0.001)	0.250 (0.001)	0.312 (0.002)	0.268 (0.001)	0.259 (0.001)	0.307 (0.002)
B.A.	0.400 (0.001)	0.392 (0.001)	0.402 (0.002)	0.426 (0.001)	0.417 (0.001)	0.415 (0.002)
Advanced degree	0.575 (0.001)	0.575 (0.002)	0.531 (0.003)	0.599 (0.001)	0.602 (0.002)	0.552 (0.003)
Work in MSA	0.198 0.000	0.194 (0.001)	0.165 (0.001)	0.202 0.000	0.197 (0.001)	0.162 (0.001)
Female	-0.171 (0.001)	-0.187 (0.001)	-0.137 (0.003)	-0.295 0.000	-0.316 (0.001)	-0.312 (0.001)
Full-time	0.237 (0.001)	0.219 (0.001)	0.256 (0.002)	--	--	--
Female × full-time	-0.126 (0.001)	-0.131 (0.002)	-0.180 (0.003)	--	--	--
Industry:						
Mining	0.197 (0.002)	0.142 (0.004)	0.239 (0.007)	0.189 (0.002)	0.143 (0.003)	0.233 (0.006)
Construction	0.019 (0.001)	0.025 (0.002)	-0.047 (0.024)	0.016 (0.001)	0.023 (0.002)	-0.077 (0.022)
Manufacturing	0.061 (0.001)	0.049 (0.001)	0.127 (0.003)	0.062 (0.001)	0.049 (0.001)	0.126 (0.003)
Transportation	0.117 (0.001)	0.103 (0.002)	0.177 (0.004)	0.112 (0.001)	0.102 (0.001)	0.183 (0.003)
Retail	-0.173 (0.001)	-0.172 (0.001)	-0.160 (0.004)	-0.181 (0.001)	-0.172 (0.001)	-0.175 (0.003)
FIRE	0.022 (0.001)	0.027 (0.001)	-0.013 (0.005)	0.023 (0.001)	0.030 (0.001)	-0.018 (0.005)
Services	-0.069 (0.001)	-0.042 (0.001)	0.026 (0.003)	-0.066 (0.001)	-0.040 (0.001)	0.012 (0.003)

Occupation:						
Manager	0.301	0.322	0.288	0.290	0.305	0.272
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)
Support	0.116	0.114	0.066	0.116	0.112	0.065
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)
Service	-0.050	-0.063	-0.094	-0.087	-0.095	-0.110
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)
Farmer	-0.116	-0.132	-0.156	-0.121	-0.139	-0.159
	(0.003)	(0.007)	(0.010)	(0.003)	(0.006)	(0.010)
Production	0.137	0.139	0.138	0.136	0.135	0.130
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)
R^2	0.358	0.396	0.369	0.426	0.448	0.442
N	12,143,183	3,291,213	904,589	9,375,086	2,725,599	742,188

Note: The dependent variable is the log of hourly wages. Standard errors in parentheses. For columns 1-3, the sample corresponds to row d. in Appendix B. For columns 4-6, the sample corresponds to row e. in Appendix B.

Table 8

Log Average Payroll with Employment and Aggregated Industry Effects

	One Digit Industry Effects			Three Digit Industry Effects		
	SSEL (1)	DEED (2)	NWECD (3)	SSEL (4)	DEED (5)	NWECD (6)
Intercept	9.5186 (0.001)	9.7037 (0.003)	9.1824 (0.008)	9.7499 (0.002)	9.9386 (0.004)	9.4996 (0.051)
Log(total employment)	0.0805 (0.001)	0.0525 (0.002)	0.0992 (0.004)	0.0748 (0.001)	0.0959 (0.002)	0.1296 (0.004)
Log(total employment)^2	-0.0060 (0.0002)	-0.0042 (0.0003)	-0.0039 (0.001)	-0.0027 (0.0002)	-0.0075 (0.0002)	-0.0088 (0.001)
In a Census-Designated Place	0.1276 (0.001)	0.1000 (0.002)	0.1001 (0.004)	0.0931 (0.001)	0.0746 (0.001)	0.1178 (0.004)
Single Unit	-0.2857 (0.001)	-0.0456 (0.001)	-0.2401 (0.004)	-0.2876 (0.001)	-0.1164 (0.001)	-0.2193 (0.004)
Mining	0.4337 (0.005)	0.4167 (0.009)	0.7039 (0.018)	.	.	.
Construction	0.2102 (0.001)	0.2401 (0.003)	0.3880 (0.056)	.	.	.
Manufacturing	0.2234 (0.002)	0.1135 (0.002)	0.4170 (0.005)	.	.	.
Transportation	0.2335 (0.002)	0.2079 (0.003)	0.5820 (0.007)	.	.	.
Wholesale	0.3943 (0.001)	0.2553 (0.002)	0.5037 (0.007)	.	.	.
Retail	-0.4670 (0.001)	-0.5362 (0.002)	-0.0892 (0.005)	.	.	.
FIRE	0.1571 (0.001)	0.1334 (0.002)	0.3610 (0.010)	.	.	.
R^2	0.115	0.168	0.217	0.245	0.353	0.376
N	5,238,631	972,266	137,534	5,238,631	972,266	137,534

Note: The dependent variable is the log of average earnings per worker in an establishment. Standard Errors are in parentheses. The sample corresponds to row d. of Appendix B for the DEED and NWECD and row e. for the SSEL, with the additional restriction that average payroll per worker had to be greater than $200 and less than $750,000.

Appendix A

Hypothetical Matched Observations and Hand-Check Scores

Worker-supplied information:	SSEL information:
Score A =1 & Score B=1:	
Tiles ‘R’ Us	Tiles ‘R’ Us, Inc.
2440 Main St.	2440 S Main
Shelbyville, SW 11111	Shelbyville, SW 11111
Industry=703	Industry=703
Score A =2 & Score B=2:	
Tiles ‘R’ Us	Tiles ‘R’ Us, Inc.
2240 E Main St.	2440 S Main
Gotham, SW 11111	Gotham, SW 11111
Industry=703	Industry=703
or	
Tiles ‘R’ Us	Tiles ‘R’ Us, Inc
Shopping Plaza	PO Box 222
Shelbyville, SW 11111	Shelbyville, SW 11111
Industry=703	Industry=703
Score A =3 & Score B=3:	
Grocery Store Chain Name	Grocery Store Chain Name
Grocery Store Chain Name	2440 S Main
Shelbyville, SW 11111	Shelbyville, SW 11111
Industry=601	Industry=601
or	
Tiles ‘R’ Us	Tiles ‘R’ Us, Inc
2440 S Main St	2400 US Highway 10
Gotham, SW 11111	Gotham, SW 11110
Industry=703	Industry=703
Score A =4 & Score B=4:	
Shelbyville Hose	Shelbyville Manufacturing
Main	2440 S Main
Shelbyville, SW 110011	Shelbyville, SW 11111
Industry=121	Industry=200
or	
Bank of Gotham	Bank of Gotham
2440 Main St	300 Fenwick R
Gotham, SW 11111	Gotham, SW 11111
Industry=700	Industry=700

Score A =5 & Score B=5:

Gotham Shop & Save	Gotham Engine Repair Co.
2440 Main St	2400 Peaceful St
Shelbyville, SW 11111	Shelbyville, SW 11111
Industry=603	Industry=751

or

Reliable Car Repair	Reliable Dry Cleaners
200 Main St	2440 Main St
Shelbyville, SW 11111	Shelbyville, SW 11111
Industry=751	Industry=771

Score A =1 & Score B=2:

Shelbyville Hospital	Shelbyville Hospital
Main	2440 Main St
Shelbyville, SW 11101	Shelbyville, SW 11111
Industry=831	Industry=831

Score A =1 & Score B=3:

Shelbyville Gas Works	Shelbyville Gas Works
2440 Main St.	2440 Main St.
Shelbyville, SW 11111	Shelbyville, SW 11111
Industry=201	Industry=641 (Eating Place industry code)

or

Chuck & Dave's Bait	Chuck & Dave's
Highway 10	2440 Highway 10
Shelbyville, SW 11111	Shelbyville, SW 11111
Industry=601	Industry=601

Score A =2 & Score B=5:

A1 Manufacturing	A1 Manufacturing Credit Union
2440 Main St.	2440 Main St.
Gotham, SW 11111	Gotham, SW 11111
Industry=201	Industry=702

Score A =2 & Score B=3:

Gotham Bank	Gotham Bank
Gotham Bank	2440 Main St.
Gotham, SW	Gotham, SW 11111
Industry=700	Industry=700

Note: In these examples, Shelbyville is a small city or a town and Gotham is a major city.

Appendix Table A1: Distribution of Matches

by Pass and Round

	Total %		Total %
Pass 1	490,408 14.9	Pass 9	60,987 1.85
Pass 2	385,627 11.72	Pass 10	19,215 0.58
Pass 3	702,355 21.34	Pass 11	7,027 0.21
Pass 4	523,534 15.91	Pass 12	25,737 0.78
Pass 5	227,557 6.91	Pass 13	85,646 2.60
Pass 6	376,436 11.44	Pass 14	55,098 1.67
Pass 7	165,877 5.04	Pass 15	35,040 1.06
Pass 8	119,877 3.64	Pass 16	10,792 0.33
Total			3,291,213 100

Appendix B

Observation Counts for DEED and Source Datasets

		DEED		SEDF	SSEL
		Workers	Establishments	Workers	Establishments
a.	All valid observations			17,311,211	6,351,658
b.	All with industry not Public Administration	3,839,904	1,166,571	16,497,515	6,351,652
c.	& in US, non-missing wages, not self-employed, and (if DEED) number matched < total employment	3,435,354	1,029,712	14,228,835	--
d.	& not in out of scope industry	3,291,213	972,436	12,143,183	6,045,800
e.	(if SSEL) & valid county & total employment > 0	--	--	--	5,237,592
f.	& aged 18-65, usually worked 30-65 hours per week, worked at least 30 weeks in 1989, and earned $2.50 - $500 per hour	2,725,599	826,815	9,375,086	--

Note: The number of establishments for the DEED column is the set of establishments whose matched employees meet the worker side restriction listed. The SEDF is the Long Form Sample from the 1990 Decennial Census and the SSEL is the 1990 Business Register.