Ethnicity, Language, and Workplace Segregation: Evidence from a New Matched Employer-Employee Data Set

Judith Hellerstein and David Neumark *

ABSTRACT. – We describe the construction and assessment of a new matched employer-employee data set (the Decennial Employer-Employee Dataset, or DEED) that we have undertaken as a part of a broad research agenda to study segregation in the U.S. labor market. In this paper we examine the role of segregation by Hispanic ethnicity and language proficiency, contributing new, previously unavailable descriptive information on segregation along these lines, and evidence on the wage premia or penalties associated with this segregation. The DEED is much larger and more representative across regional and industry dimensions than previous matched data sets for the United States, and improvements along both of these dimensions are essential to isolating the importance of segregation by language and ethnicity in the workplace.

Our empirical results reveal considerable segregation by Hispanic ethnicity and by English language proficiency. We find that Hispanic workers, but not white workers, suffer wage penalties from employment in a workplace with a large share of Hispanic workers, and even more so a large share of Hispanic workers with poor English language proficiency. In addition, we find that segregation of Hispanic workers among other Hispanics with similar English language proficiency does not reduce the penalties associated with poor own language skills.

L'ethnicité, la langue, et la ségrégation sur le lieu de travail: résultats d'une nouvelle base de données appariée employeurs-employés

Résumé. – Nous utilisons une nouvelle base de données appariée employeurs-employés pour examiner la ségrégation ethnique des hispaniques et le degré de maîtrise de la langue anglaise, ainsi que les écarts de salaire associés à cette ségrégation. Nous trouvons une ségrégation significative du fait de l'appartenance au groupe ethnique des hispaniques et du degré de maîtrise de la langue anglaise. Les employés hispaniques sont pénalisés au niveau des salaires quand il travaillent plus avec d'autres hispaniques que la moyenne des autres salariés – surtout quand ceux-ci ont de faibles compétences en langue anglaise. En revanche, la ségrégation augmentée par la maîtrise de la langue ne réduit pas les pénalités salariales associées aux compétences linguistiques individuelles.

* HELLERSTEIN is Associate Professor of Economics at the University of Maryland, and NEUMARK is Senior Fellow at the Public Policy Institute of California. Both are Research Associates of the NBER. We thank KIMBERLY BAYARD, JOEL ELVERY, JENNIFER FOSTER, and MEGAN BROOKS for outstanding research assistance, and ETIENNE WASMER, members of the University of Maryland Demography of Inequality Initiative, and an anonymous referee for helpful comments. 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 Census Bureau 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, nor reflect the views of the Public Policy Institute of California. This paper has been screened to insure that no confidential data are revealed. This paper was prepared for the ADRES/CEPR/Université du Maine Conference "Discrimination and Unequal Outcomes", Le Mans, France, January 17-20, 2002.

1 Introduction

There is vast evidence of segregation by race, sex, and ethnicity in the U.S. labor market, and much of this segregation is at the hard-to-observe level of the establishment. For example, in previous work using a matched employer-employee data set that we constructed to examine the role of segregation in the labor market, we found that randomly selected black workers worked in establishments that had 22-29% more blacks than in establishments of randomly selected white workers, and randomly selected Hispanic workers worked in establishments that had about 32% more Hispanics than in those of randomly selected white workers (BAYARD, *ET AL.* [1999]).

Even when controlling for individual characteristics of workers, part of the lower wages paid to blacks, women, and Hispanics remains attributable to segregation by industry, occupation, establishment, and job-cell (occupations within establishments). This adverse effect of segregation on wages has often been interpreted as evidence of labor market discrimination, where female and minority workers are "crowded" into a subset of jobs, resulting in lower wages for the workers in these jobs, who are disproportionately female and minority (BERGMANN [1974]).

While segregation per se, and the wage penalties associated with it, may stem from discrimination, alternative explanations are possible. Theoretically, paralleling Lang's [1986] model of language discrimination, productivity may rise when employers group like workers together if similarities across workers in an establishment lower transaction costs between workers. In such a case, there are incentives to segregate workers by type. And those workers "outside" of the majority may suffer wage penalties associated with the higher transaction costs that they impose, given that complete segregation is unlikely and some interaction is required.

As the labor force in the United States becomes increasingly heterogeneous across dimensions of race, ethnicity, and spoken language (including but not limited to English proficiency), the issue of how people interact in the labor market is going to become increasingly important. This is especially true at the establishment level, where workers need to interact with each other, with supervisors, and – in some types of establishments – with customers. There already is evidence that one's ability to communicate with others is an important determinant of labor market success. In BAYARD, *ET AL.* [1999], we found that workers who do not speak English well earn significantly less (between 16% and 30%) than workers who do speak English well, even controlling for race, ethnicity, and other typical human capital controls; TREJO [1997] reports a similar result.

In this paper, we describe the construction and assessment of a new matched employer-employee data set that we have undertaken as a part of a broad research agenda to examine labor market segregation in the United

States.[1] We then illustrate the usefulness of these data by examining the role of ethnicity and language in the labor market. The empirical analysis contributes new, previously unavailable descriptive information on labor market segregation in the United States, and its role in generating wage premia or penalties. In addition, although testing theories is not the main focus of this paper, we do provide evidence on models and hypotheses regarding the role of ethnicity and language in the labor market. Of course, using a data set that matches workers to establishments is key to determining the importance of segregation by ethnicity and language in the workplace, because so much of the segregation is at the establishment level (BAYARD, *ET AL.* [1999]), and because the most pertinent interactions between workers occur within establishments. While we have constructed and analyzed matched employer-employee data sets in the past, our new matched data set is much larger and more representative across regional and industry dimensions than previous matched data sets for the United States; the improvements along both of these dimensions are essential to isolating the importance of segregation by language and ethnicity in the workplace.

In order to narrow the definition of ethnicity, and because Hispanics are an important and rapidly growing segment of the labor force, in the empirical work that we present we focus on comparisons between white men and Hispanics.[2] We also narrow our definition of language and focus solely on English language proficiency, although there are clearly other possible definitions. We then examine the extent of establishment-level segregation by Hispanic ethnicity and by English language proficiency in our data, and estimate a variety of wage penalties or premiums associated with ethnic and language segregation.

Our empirical results reveal considerable segregation by Hispanic ethnicity in the U.S. labor market. White men who speak English very well and work with at least one Hispanic co-worker work in establishments that are on average 12% Hispanic, whereas Hispanics work in establishments that are on average 47% Hispanic. There is also a tremendous amount of segregation by English language proficiency. Of the Hispanic co-workers of white men who speak English very well, less than 2% of them speak English poorly or not at all; in contrast, of the Hispanic co-workers of Hispanics who do not speak English, 44% of them speak English poorly or not at all.

Turning to the effects of segregation on wages, we find that Hispanic workers, but not white workers, suffer wage penalties from employment in a workplace with a large share of Hispanic workers, and even more so a large share of Hispanic workers with poor English language proficiency. In addition, we find that segregation of Hispanic workers among other Hispanics with similar English language proficiency does not reduce the penalties associated with poor own language skills.

---

1. Part of our research agenda also focuses on the link between workplace segregation and residential segregation, since our worker data contains home address information that can be used to measure residential segregation.

2. We focus on non-black Hispanics only.

2 The Construction and Evaluation of the DEED Matched Employer-Employee Data Set

A. Introduction

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 section of the 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 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 DEED is very large, containing information on 3.2 million workers matched to nearly one million establishments, which account for 27% of workers in the Decennial Census and 19% of active establishments in the Standard Statistical Establishment List (SSEL), an administrative database containing information for all business establishments operating in the United States in 1990.[3] As it stands, it is the largest national matched employer-employee database covering the United States that contains detailed demographic information on workers,[4] making it a rich source of information for studying a variety of questions of interest to labor economists, demographers, and others.

3. These numbers are based on the data set after basic sample inclusion criteria have been imposed, as described below.

4. 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 one-in-six sample of Census 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.

B. **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 have 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. 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.[5] 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 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 1990 SEDF.

The SSEL is an annually-updated list of all business establishments with one or more employees operating in the United States. The Census Bureau uses the SSEL as a sampling frame for its Economic Censuses and Surveys, and continuously updates the information it contains. 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 other Census Bureau establishment- or firm-level data sets that contain more detailed employer characteristics.[6]

Matching workers to employers to create the WECD and the NWECD proceeded in four steps. First, we standardized the geographic and industry codes in the SEDF and the SSEL. Next, we selected all establishments that were unique in an industry-location cell. Third, all workers who indicated they

---

5. 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.

6. 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 that 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.

worked in the same industry-location cell as a unique establishment were matched to the establishment. Finally, we eliminated all matches based on imputed data. 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 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).[7]

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. This latter 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 2003] 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 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 new methods of studying labor market discrimination and other issues, and unique results, there are a few shortcomings of these data sets that are of serious concern. Because the match is based on the 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 1990 SSEL with positive employment, only 388,787 are unique in an industry-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 including about 900,000 workers in 138,000 establishments, which covers

7. 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]).

7% of all workers in the SEDF, and 3% of all establishments in the SSEL.[8] 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. For focusing on Hispanics and language proficiency, under-representation of small establishments in urban areas is particularly problematic, rendering the WECD and NWECD even less ideal.

C. Overview of the 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.[9]

When the NWECD was created, the specific name and address files for Long-Form respondents were unknown and unavailable to researchers. Subsequently, we were able to help track down the name and address files and to participate in their 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 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 employer. 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.

8. Again, these numbers are prior to sample restrictions imposed in the analysis.

9. 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.

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 in 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 United States, the SSEL contains basic establishment-level information including geography, industry, total employment, payroll, and an indicator for whether the establishment is a single-unit enterprise or part of a multi-unit firm. The SEDF contains the full set of responses provided by all Long-Form respondents. 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 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 (described below) 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 considered legal or administrative entities assigned an Employer Identification Number (EIN) by the Internal Revenue Service; a sin-

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, 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

gle 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 establish-

ments, 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.[10] 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.

D. 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. 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]).

Our method to link records using MatchWare involves two basic steps. The first step is to use 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 a wide variety 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).[11] 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").

10. 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."

11. 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]).

The second step of the matching process is to select and implement the matching specifications. The AutoMatch software uses a probabilistic matching algorithm that 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 appear in order 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 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.

For example, we might 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. 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.

E. **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 sam-

ples of 1,000 worker observations from each of the five most populous states (CA, NY, TX, PA, IL) plus three other states (FL, 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.[12]

For each observation selected, we retained identifying information from both the SEDF (Decennial Census) 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; and 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 Table A2 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 Table A2 should make it clear that scores of 1 and 2 were given by the hand-checkers for 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.[13] 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 that 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 vari-

---

12. 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.

13. The hand-checking was done by five different researchers who were randomly assigned to matches.

TABLE 2

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

	Score B					
Score A	1	2	3	4	5	Row total
1	9,930 *66.16*	2,229 *14.85*	291 *1.94*	56 *0.37*	79 *0.53*	12,585 *83.85*
2		1,126 *7.50*	406 *2.71*	95 *0.63*	30 *0.20*	1,657 *11.04*
3			158 *1.05*	123 *0.82*	252 *1.68*	533 *3.55*
4				40 *0.27*	101 *0.67*	141 *0.94*
5					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.

ables include a worker's age, sex, race/ethnicity, education, full-time status, and English speaking ability. The geographic variables include state indicators, dummy variables for whether or not the location of the worker's employer is 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.[14]

The results of the regressions in Table 3 indicate that only a few variables or sets of variables are 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).[15] Aside from differences by industry in the probability of a bad match, blacks are 0.05-0.07 more likely to be poorly matched than whites, and those with advanced degrees are also more likely than others to be poorly matched.

14. 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.

15. The sample size is a bit lower than in the frequency table (Table 2) because these regressions exclude 5 people in the military and use the worst two of the four hand-checked scores for observations that were selected both in the worker sample and in the establishment sample.

TABLE 3

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

	Coefficient (1)	Std. error (2)	Coefficient (3)	Std. error (4)	Coefficient (5)	Std. error (6)	Coefficient (7)	Std. error (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)

	Coefficient (1)	Std. error (2)	Coefficient (3)	Std. error (4)	Coefficient (5)	Std. error (6)	Coefficient (7)	Std. error (8)
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: Illinois (state), services (industry), and support (occupation)

The characterization of match quality as varying systematically by industry was also the consensus of the researchers who had done the hand-checking, where industry differences were noted even more dramatically at a level of industry finer than the two-digit controls included in the regressions in Table 3. Figure 1 shows a histogram of the distribution of error rates across industries (where industry is defined by the 3-digit classification in 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.[16] 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 individuals 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 0.20.

Given the information that match quality was so strongly associated with industry, we embarked upon a plan to refine our matching procedure by developing criteria to reduce errors by industry. We identified those industries that (1) had an estimated error (bad match) rate of 0.10 or more, and (2) represented at least 1% of employment in the entire matched national data set. The 0.10 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. 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, there are four possible tabulations for each industry. 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 0.10 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.

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

FIGURE 1

Histograms for Industries by Error Rates Based on First Round of Hand-Checking

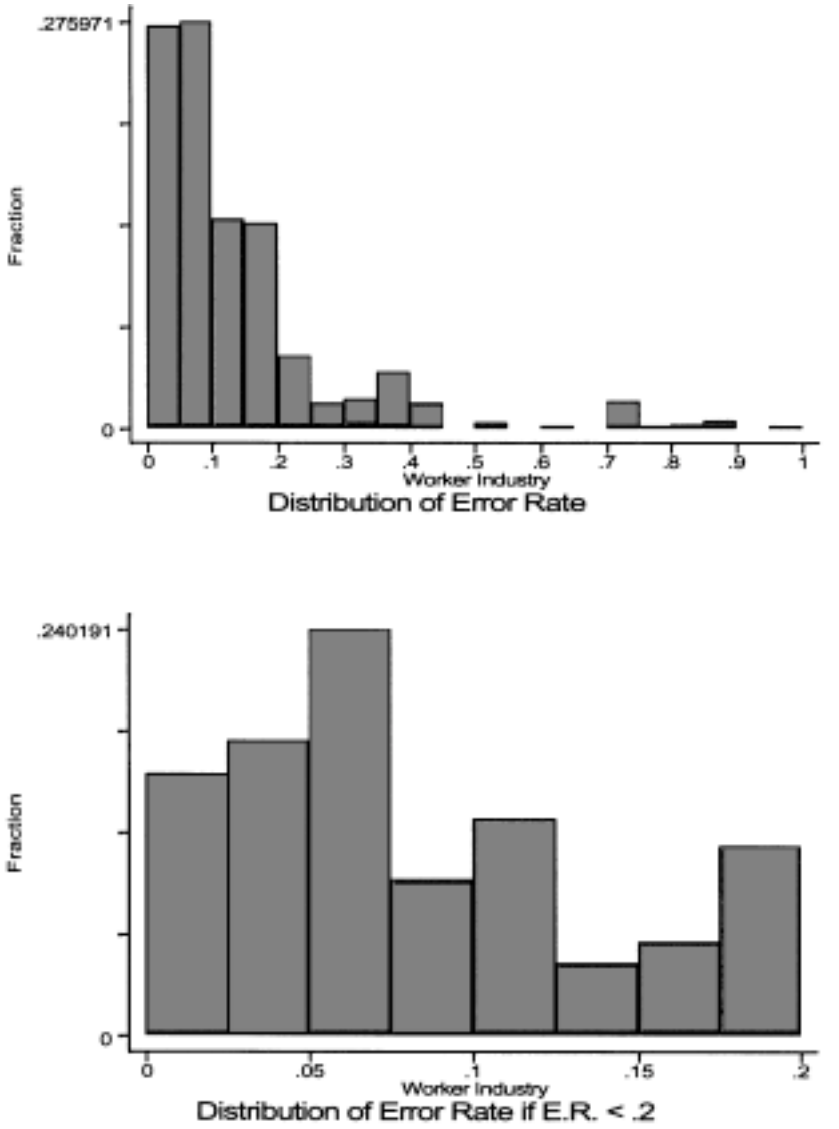


TABLE 4
Scored Match Rates for "Problem" Industries

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

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 seven 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 1 Medical Drive Anytown, USA Industry: 831 (hospitals)		Private Cafeteria of General Hospital 1 Medical Drive Anytown, USA Industry: 641 (eating and drinking places)

In this example, hospital employees in the SEDF and the business name of the hospital's cafeteria in the SSEL refer to the hospital by a common name, while – in another entry on the SSEL with an industry code of 831 – the actual hospital uses the hospital's legal name and perhaps also the address of the parent hospital chain headquarters, rather than the physical location of the hospital in Anytown. Therefore, the closest match to the hospital workers in the SEDF (and note that many of the 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 exactly 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 (like in large cities) but in different zip codes. Finally, for the remaining two industries (physicians' offices and clinics, and legal services) we modified the AutoMatch 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 that the standard algorithm in AutoMatch 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 17 problematic industries of 100 workers and 30 establishments (all the workers matched to each establishment) in the same eight states examined earlier. 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 eight of the 14 industries, but six industries still had error rates over 0.10 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 six 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 six industries listed above; these industries account for 14% of the workers in the final version of the DEED and 17% of the establishments.

F. **Evaluating the Representativeness of the Matched Data**

To evaluate the representativeness of the matched DEED, 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 comparable 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

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	37.51	38.61	0.42	1.52	37.69	37.87	38.96	0.18	1.26
	(12.78)	(12.23)	(12.23)			(11.27)	(11.06)	(11.10)		
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	1.53	1.84	– 0.04	0.27	1.57	1.51	1.83	– 0.05	0.27
	(1.62)	(1.55)	(1.65)			(1.59)	(1.53)	(1.63)		
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	2.30	2.24	0.10	0.03	2.31	2.38	2.31	0.06	– 0.01
	(0.70)	(0.65)	(0.64)			(0.58)	(0.57)	(0.54)		
Hourly wage	12.10	12.89	11.76	0.79	– 0.34	12.22	12.98	11.73	0.76	– 0.50
	(82.19)	(37.07)	(25.77)			(11.27)	(12.07)	(11.73)		

	All workers					Full-time workers				
	SEDF	DEED	NWECD	DEED – SEDF	NWECD – SEDF	SEDF	DEED	NWECD	DEED – SEDF	NWECD – SEDF
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Hours worked in 1989	39.51	40.42	39.91	0.92	0.41	42.11	42.41	41.89	0.30	– 0.21
	(11.44)	(10.37)	(10.18)			(6.12)	(6.03)	(5.56)		
Weeks worked in 1989	46.67	48.21	47.95	1.54	1.28	50.33	50.71	50.65	0.37	0.32
	(11.05)	(9.35)	(9.70)			(4.25)	(3.71)	(3.79)		
Earnings in 1989	22576	25581	22485	3005.4	– 90.8	26465	28559	25280	2093.5	– 1185.2
	(26760)	(29475)	(21232)			(26852)	(29336)	(20804)		
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 workers in all data sets who are not excluded by the basic restrictions.[17] 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, 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%.[18] Similarly, there is a close parallel among the distributions of workers across education categories in all data sets.

The distributions of workers across industries paint 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, while only 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

17. We exclude individuals from the SEDF who did not work in the year prior to the survey year [1989], worked in public administration, 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). 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]).

18. We argued earlier that the DEED was especially useful in capturing Hispanic workers, who work disproportionately in small urban establishments. The share Hispanic in the DEED is higher by 33% (5.2% vs. 3.9%, which round to the 5% and 4% figures reported in the text). Moreover, in the DEED, 9.6% of Hispanics work in central city establishments with fewer than 15 workers; 7.1% of whites work in establishments fitting these criteria.

TABLE 6
Means for Establishments

	SSEL (1)	DEED (2)	NWECD (3)	DEED – SSEL (4)	NWECD–SSEL (5)
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 (1385.12)	24.24 (111.79)	18.56 (76.08)	3.22	– 2.46
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 worker's reported place of work was used to determine MSA status.

sent a Long Form, and consequently more likely that such establishments are 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 are 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 distributions 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.

Finally, an 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 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 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).

Thus, the DEED offers substantial improvements in providing a matched employer-employee data set for the United States that is larger and more representative of the actual population of workers and establishments.

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

	All			Full-time		
	SEDF (1)	DEED (2)	NWECD (3)	SEDF (4)	DEED (5)	NWECD (6)
FIRE	0.022	0.027	– 0.013	0.023	0.030	– 0.018
	(0.001)	(0.001)	(0.005)	(0.001)	(0.001)	(0.005)
Services	– 0.069	– 0.042	0.026	– 0.066	– 0.040	0.012
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(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 are reported in parentheses.

3 Examining the Extent of Segregation by Hispanic Ethnicity and English Language Proficiency

There are multiple reasons why segregation may exist in the labor market. One reason may be discrimination of the sort that causes labor market inefficiencies, such as in Bergmann's [1974] model where employers "crowd" some types of workers into establishments or occupations, lowering these workers' marginal products and therefore their wages. Conversely, segregation may exist for efficiency-enhancing reasons that are based on lowering the transaction costs of communication in the labor market. For example, one way to think about the importance of communication in an establishment is in the framework of Lang's [1986] model of language discrimination. In this model, there are transaction costs within a firm when some members of the firm have to learn the language of other members; this can also refer to differences in proficiency in a common language, of course. Because of this, the model predicts that in the short run workers of different languages will be segregated in the workplace and, given that some interaction is required, workers who do not

speak the language of the economically-dominant majority will earn lower wages. In the long run, the results of Lang's model are more ambiguous. If workers learn each others' languages, segregation will be reduced and the wage gap across groups will decrease. Alternatively, members of the minority group may become owners of firms, which will perpetuate segregation by language by reducing the wage gap across language groups. "Language" in Lang's model can be literal, or it can refer to cultural differences across groups that lead to transaction costs in communication.

Because of its size and depth as a matched employer-employee data set, the DEED is uniquely capable of measuring the extent of segregation at the establishment level by ethnicity and language. However, because the SEDF contains such rich information about individuals, there are multiple ways of categorizing ethnicity and language. For example, English language proficiency and Household Language Spoken are both measures of literal language, whereas differing Country of Origin may reflect cultural language differences across individuals who speak the same literal language. In this paper, we use English language proficiency as the definition of language, taking Lang's model most literally in that we use the true language of the majority (English) as the benchmark. Clearly, however, in examining differences between whites and Hispanics, we are indirectly also examining the possibility that there are "language" differences across these ethnic groups even among those who speak English well.

In Table 8, we report summary statistics from the sample of white and Hispanic men in the DEED who work in establishments where we match at least one Hispanic worker.[19] We make this restriction for two reasons. First, a vast majority of white men in the DEED are matched to establishments where we do not observe them working with any Hispanic co-workers; incorporating these men into the data set would result in our estimates being driven by a sample of workers who we did not observe to interact with any Hispanics in the workplace. Second, in a part of our analysis we examine wage gaps between Hispanics and whites by incorporating establishment fixed effects, where identification of the wage gap is driven by the existence in the establishment of at least one Hispanic worker (and one white worker, but this is less of an issue in the data). We also restrict our sample of white men to those who speak English very well, so that we do not have to consider language differences across white men; for ease of exposition throughout the paper we refer to this sample of white men who speak English very well simply as "white men." In short, we consider white men who speak English well and work with at least one Hispanic to be the benchmark, majority group, to which Hispanics are compared. As a result of this sample selection, our sample contains information on just over 327,000 white men who speak English well, and just over

19. We restrict the sample to white men so that we do not have to consider wage differences between sexes or races. We also restrict the sample to workers who report working full-time (at least 30 hours per week and at least 30 weeks) in the previous year. We do this for two reasons. First, these workers are more attached to the labor force in general and we thus worry less about capturing previous labor market experience differentials between part-time and full-time workers. Second, because of the nature of the DEED construction, we match workers to their employer in the previous week, but the wage data come from the previous year. Full-time workers are more likely than part-timers to have earned last year's wages at the previous week's establishment.

TABLE 8

Means for Men from DEED in Establishments with At Least One Matched Hispanic Worker

			Hispanic and speak English:			
	White (1)	Hispanic (2)	Very well (3)	Well (4)	Poorly (5)	Not at all (6)
Log(wage)	2.748	2.284	2.392	2.249	2.020	1.825
	(0.538)	(0.536)	(0.539)	(0.483)	(0.450)	(0.409)
Age	39.604	35.542	34.768	37.165	36.274	36.164
	(10.691)	(10.882)	(10.443)	(11.065)	(11.508)	(12.665)
$Age^2/100$	16.828	13.817	13.179	15.037	14.482	14.682
	(8.896)	(8.592)	(8.138)	(8.942)	(9.165)	(10.082)
Age – 18	21.604	17.542	16.768	19.165	18.274	18.164
	(10.691)	(10.882)	(10.443)	(11.065)	(11.508)	(12.665)
$(Age - 18)^2/100$	5.810	4.261	3.902	4.897	4.664	4.903
	(5.113)	(4.757)	(4.461)	(5.035)	(5.109)	(5.616)
Married	0.756	0.685	0.663	0.754	0.694	0.640
	(0.429)	(0.464)	(0.473)	(0.431)	(0.461)	(0.480)
High school degree	0.278	0.251	0.298	0.226	0.137	0.081
	(0.448)	(0.433)	(0.458)	(0.418)	(0.344)	(0.273)
Some college	0.226	0.189	0.247	0.141	0.059	0.030
	(0.418)	(0.391)	(0.431)	(0.348)	(0.236)	(0.169)
Associates degree	0.086	0.058	0.075	0.044	0.019	
	(0.281)	(0.233)	(0.263)	(0.204)	(0.137)	
B.A.	0.223	0.073	0.104	0.038	0.014	
	(0.416)	(0.261)	(0.305)	(0.191)	(0.120)	
Advanced degree	0.105	0.031	0.042	0.018	0.009	
	(0.306)	(0.173)	(0.201)	(0.132)	(0.095)	
Live in MSA	0.859	0.919	0.905	0.932	0.947	0.952
	(0.348)	(0.273)	(0.293)	(0.251)	(0.225)	(0.213)
Hispanic × don't speak English	—	0.043	—	—	—	—
		(0.202)				
Hispanic × speak English poorly	—	0.146	—	—	—	—
		(0.353)				
Hispanic × speak English well	—	0.206	—	—	—	—
		(0.405)				
Share of co-workers who are Hispanic	0.100	0.438	0.367	0.479	0.594	0.701
	(0.135)	(0.308)	(0.288)	(0.305)	(0.295)	(0.270)
Share of Hispanic co-workers who:						
Don't speak English	0.007	0.046	0.013	0.033	0.072	0.481
	(0.061)	(0.142)	(0.063)	(0.099)	(0.144)	(0.296)
Speak English poorly	0.047	0.146	0.050	0.114	0.568	0.221
	(0.163)	(0.252)	(0.125)	(0.184)	(0.298)	(0.219)
Speak English well	0.130	0.198	0.089	0.570	0.147	0.125
	(0.255)	(0.279)	(0.157)	(0.314)	(0.188)	(0.168)

			Hispanic and speak English:			
	White (1)	Hispanic (2)	Very well (3)	Well (4)	Poorly (5)	Not at all (6)
Speak English very well	0.816	0.610	0.848	0.283	0.212	0.173
	(0.308)	(0.383)	(0.229)	(0.287)	(0.244)	(0.212)
Share of co-workers with:						
High school degree	0.304	0.272	0.292	0.267	0.220	0.183
	(0.178)	(0.219)	(0.224)	(0.212)	(0.202)	(0.188)
Some college	0.233	0.222	0.250	0.202	0.161	0.131
	(0.121)	(0.198)	(0.203)	(0.186)	(0.174)	(0.163)
Associates degree	0.087	0.066	0.075	0.060	0.046	0.038
	(0.077)	(0.110)	(0.115)	(0.106)	(0.098)	(0.089)
B.A.	0.185	0.121	0.142	0.104	0.077	0.060
	(0.147)	(0.163)	(0.174)	(0.147)	(0.127)	(0.117)
Advanced degree	0.076	0.041	0.050	0.033	0.023	0.016
	(0.102)	(0.096)	(0.107)	(0.083)	(0.069)	(0.056)
Average age of workers	38.849	37.254	37.192	37.808	37.047	36.153
	(4.577)	(6.627)	(6.348)	(6.817)	(7.082)	(7.667)
Share married	0.693	0.654	0.650	0.671	0.651	0.627
	(0.147)	(0.235)	(0.232)	(0.231)	(0.246)	(0.259)
Observations	327,190	69,103	41,812	14,256	10,092	2,943

Note: Standard deviations are reported in parentheses. The higher education categories for Hispanics who speak English poorly or not at all had to be merged because of small sizes and Census Bureau confidentiality rules. The co-worker characteristics are always computed excluding the reference worker.

69,000 Hispanic workers, among whom 61% self-report speaking English very well, 21% report speaking English well, 15% report speaking English poorly, and 4% report not speaking English at all.[20]

Not surprisingly, Table 8 shows that white men in the sample are older, better-educated, more likely to be married, and less likely to live in an MSA than Hispanic men. Among Hispanic men, the only real difference across the groups defined by English language proficiency is that there is a strong positive association between English proficiency and education. Indeed, the sample sizes of Hispanic men who do not speak English and have more than a high school education are too small by Census Bureau standards to even be able to report these means separately in the table.[21]

20. There are two ways for non-English speakers to fill out the Census. First, the Census form is available in Spanish. Second, a Census employee can be sent to help a respondent fill out the form. However, it appears that data are available only on whether the form that was used was in Spanish. Because many of the Hispanic men whose English language proficiency is reported to be either poor or "not at all" are immigrants, we generally include controls for immigration cohort when looking at Hispanics.

21. For confidentiality reasons, means and coefficients can only be reported for groups larger than 75.

The extent of segregation by Hispanic ethnicity and English language proficiency is reported in the next five rows of Table 8.[22] It should be noted that these results are unique – one needs to know the demographic composition of workers within an establishment in order to construct these segregation numbers, and this kind of information is only available in the United States in matched employer-employee data, where information on characteristics such as English proficiency comes from the individual data on employees. The average number of workers per establishment in the observations that make up Table 8 is only 7.1 (the sample of 396,293 divided by 55,793 establishments). But because all matched workers are used to construct the co-worker share variables for the worker observations in Table 8, and because workers from establishments with more matches appear more frequently in the data set, the numbers of co-workers on which the share variables are based are typically much higher – for example, for the observations in Table 8, the median number of co-workers exceeds 45 (and the mean is 129).

Table 8 shows that there is good deal of segregation by Hispanic ethnicity and by English proficiency.[23] A Hispanic man is much more likely than a white man to work with other Hispanics; for Hispanics, the average share of co-workers who are Hispanic is 43.8%, whereas for white men, the average share of co-workers who are Hispanic is 10.0%.[24]

Segregation by English language proficiency is also pronounced, and there are a number of ways to see this. First, among Hispanics, as English language proficiency drops the average share of Hispanics rises, so that for a Hispanic who speaks English very well, 36.7% of his co-workers are Hispanic, whereas for one who does not speak English at all, 70.1% of his co-workers are

---

22. The share variables are constructed excluding the contribution of the individual. For the share Hispanic, for whites we simply compute the proportion of Hispanic workers among the matched workers at an establishment. The sample used to calculate this variable includes all workers in the DEED (*e.g.*, part-time as well as full-time, women as well as men, etc.). For Hispanics, this proportion is multiplied by total employment, rounding to the nearest integer to get the imputed number of Hispanics at the establishment. We then subtract one off of the imputed share and off of total employment (to exclude the individual Hispanic from the calculation), and then compute the ratio. (Thus, this exclusion has a much larger impact on small than large establishments.)

    Calculating the share of Hispanics in a particular language proficiency category is a bit more complicated. For whites, we simply compute the ratio of Hispanics in that language category among all Hispanics matched to the establishment. For Hispanics, we use this ratio to impute the number of Hispanics in that language category. We then subtract one from that number only if the individual is in that language category, and subtract one from the total imputed number of Hispanics, after which we form the ratio. The only difficulty is when the individual is the only Hispanic matched to the plant and all workers are matched (which occurs in 149 cases in the analysis sample). In this case, this ratio would be zero divided by zero, after subtracting one from the numerator and denominator. We therefore forego the subtraction in this case, assigning a value of one. One interpretation of this is that while our data indicate we have all workers matched, this may be an error, and we are assuming that a non-matched Hispanic worker would be in the same language category as the individual, which seems the best guess given the extent of language segregation. However, we also experimented with defining this share as zero and simply dropping these cases; not surprisingly given the small number of observations involved, these alternatives had no detectable impact on the estimates.

23. In results not reported here, we confirmed that the segregation we find is not a result just of "randomness" due to measurement error or the small unit problem as described in CARRINGTON and TROSKE [1997].

24. Recall that we have restricted the sample to white men who work in an establishment where we have matched at least one Hispanic worker. This reported figure of 10.0% is therefore conditional on a white man working with at least one Hispanic.

Hispanic. Second, while on average almost 82% of a white male's Hispanic co-workers speak English very well, on average only 61% of a Hispanic's co-workers speak English very well. There is segregation by English language proficiency among Hispanics as well. For a Hispanic who speaks English very well, on average 84.8% of his co-workers speak English very well, whereas at the other end of the spectrum, for a Hispanic who does not speak English, on average only 17.3% of his co-workers speak English very well. Conversely, for a Hispanic who speaks English very well, on average 1.3% of his Hispanic co-workers do not speak English,[25] while for a Hispanic who does not speak English, 48.1% of his Hispanic co-workers do not speak English at all.

Finally, the table reports descriptive statistics for other characteristics of co-workers, which figure in some of the ensuing analyses. These figures reveal that Hispanics tend to work alongside less-educated co-workers (not surprisingly, given segregation by Hispanic ethnicity and the lower education of Hispanics), and that this is more true for those with poorer language skills.

Information on segregation is also summarized in Figure 2, which reports the distributions of shares of workers Hispanic by ethnicity and language proficiency. The first two histograms indicate the degree of segregation by Hispanic ethnicity, revealing the far greater prevalence of high shares Hispanic among Hispanic workers. The final four histograms indicate the far greater prevalence of high shares Hispanic among those Hispanics whose English is less proficient.

There is thus clear evidence of rather strong establishment-level segregation by Hispanic ethnicity and by English language proficiency. This segregation is consistent with transaction costs generated by combining workers who are heterogeneous with respect to language or ethnicity. It can also be generated by the standard crowding hypothesis, and by residential segregation along the same dimensions. To gain more information regarding the impact of these types of segregation on workers, the next section explores the possibly complex relationships between wages and a worker's own ethnicity and language skills, as well as these characteristics of his co-workers.

4 The Impact on Wages of Segregation by Hispanic Ethnicity and English Proficiency

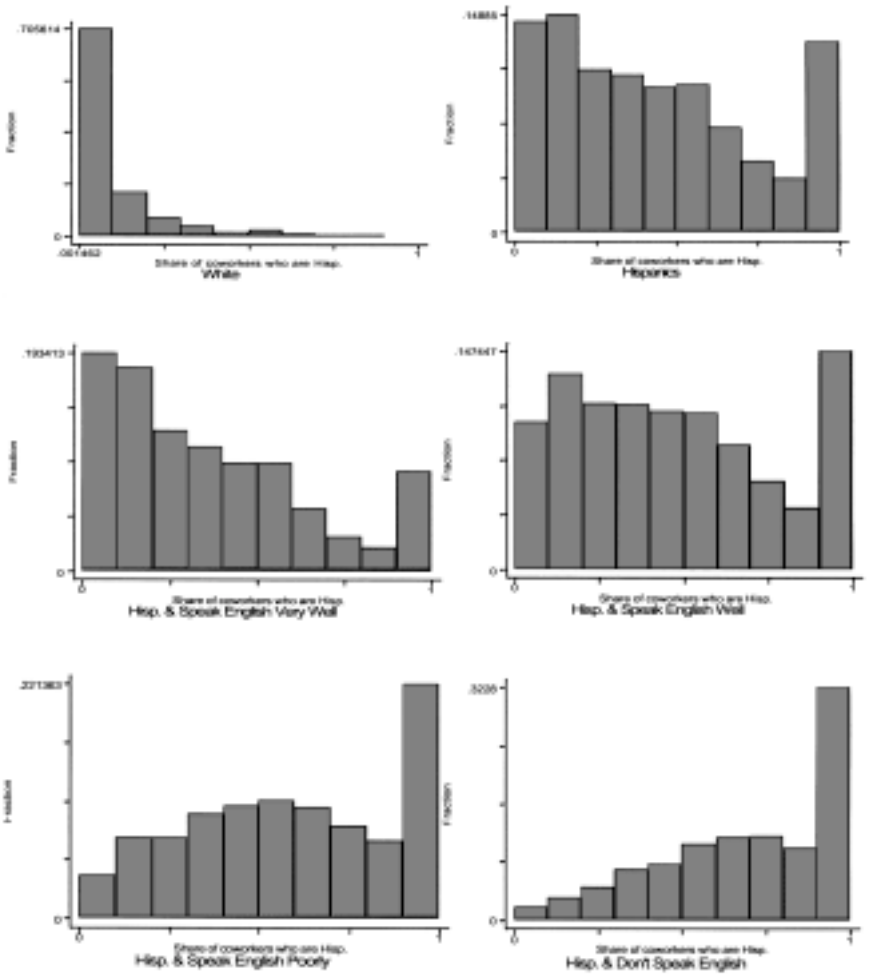
A. Wages and Segregation by Hispanic Ethnicity

We first focus on the impact on wages of segregation just by Hispanic ethnicity. Table 9 reports results from various versions of human capital wage regressions using the sample of men from the DEED in establishments with at least one matched Hispanic worker (the same sample as in Table 8). Column (1) contains the baseline results from a simple human capital regression estimated for whites and Hispanics, where we include a dummy variable for

25. Indeed, for the vast majority of Hispanics who speak English very well (93%), none of their Hispanic co-workers do not speak English.

FIGURE 2

Histograms for Segregation by Ethnicity and Language Proficiency



Hispanic, age and its square, a dummy variable for whether a worker is married, education group dummy variables, a dummy variable for whether an individual works in an MSA, and dummy variables for state and 1-digit industry. In addition, because many Hispanics in the labor force were not born in this country, it is possible that any wage penalties we estimate for Hispanics are in fact due to the immigrant status of the worker; we therefore also include dummy variables for the cohort in which a Hispanic immigrant came to the United States.[26] The estimated coefficients on the human capital characteristics are consistent with those from other U.S. data sets; there is a quadratic age profile, a positive marriage premium of 14%, positive returns to education, and a positive wage associated with living in an MSA. It should be noted that most of these coefficients are identified off of white males since they constitute 83% of the sample. In column (1), the coefficient on the Hispanic dummy is – 0.14 and statistically significant, indicating that there is a wage penalty of 14 log points associated with being Hispanic.[27]

In column (1') we replicate the specification in column (1), but add establishment fixed effects. The most important reason for doing this is that the inclusion of fixed effects isolates the effect on wages of being Hispanic from any other factors that are associated with differences in the establishments in which white men and Hispanic men work, since the coefficient on Hispanic in the fixed-effects specification is identified off of within-establishment variation between whites and Hispanics. In addition, in the specification that follows, in column (2), we add the share of co-workers who are Hispanic and who are in various demographic categories to the column (1) specification. As the establishment fixed effects subsume these share variables, the fixed-effects specification in column (1') can be interpreted as a non-parametric version of the specification in column (2), where, in particular, we do not restrict the functional form of the effect of segregation to be linear in the percent Hispanic in the establishment. Finally, the use of fixed effects to capture the share Hispanic (and everything else about the establishment) avoids the measurement error inherent in our share variables,[28] while also, of course, precluding the estimation of the effects of various workforce shares. In column (1'), the estimated coefficient on the Hispanic dummy falls somewhat, to – 0.09, while still remaining sizable and statistically significant.

---

26. The 1990 SEDF has 10 different categories for immigrant status, ranging from first immigrating to the United States before 1950 to immigrating between 1987 and 1990. These categories do not all cover the same number of years. One caveat to note is Lubotsky's (2000) finding – based on a comparison with social security administrative data – that self-reported year of first immigration may be incorrect, and may instead reflect most recent year of immigration (since many immigrants emigrate back to their home countries for some period of time).

27. The wage penalty for Hispanic ethnicity is about 50% larger excluding the immigration controls.

28. As should be clear from Section II, we only have a sample of workers matched to each plant. Thus, variables measuring the share of workers in any category in an establishment are sample estimates, and hence the regression estimates are subject to measurement error bias. This of course is a problem that in principle affects all matched employer-employee data sets that do not obtain data on the full populations of each plant's workforce. Because the error variance differs across establishments depending on the true proportion in each category, correcting for the measurement error based solely on sampling variation (which would otherwise be simple) is not trivial; see COCKBURN and GRILICHES [1987] and MAIRESSE and GREENAN [1999] for a related application. While we plan to address this issue more fully in future research, below we consider some evidence assessing the impact of measurement error that arises from sampling workers in establishments.

TABLE 9

Log Wage Regressions Including Hispanic Ethnicity
Men from DEED in Establishments with At Least One Matched Hispanic Worker

	Pooled (1)	Pooled (1')	Pooled (2)	Whites (2')	Hispanics (2")	Pooled (3)	Whites (3')	Hispanics (3")	Pooled (4)
Hispanic	– 0.135	– 0.090	– 0.108	–	–	– 0.104	–	–	– 0.043
	(0.003)	(0.003)	(0.003)			(0.003)			(0.004)
Share of co-workers	–	–	– 0.104	– 0.037	– 0.168	– 0.034	0.080	– 0.099	–
who are Hispanic			(0.008)	(0.012)	(0.009)	(0.008)	(0.012)	(0.009)	
× Hispanic	–	–	–	–	–	–	–	–	– 0.216
									(0.011)
Age	0.044	0.039	0.043	0.046	0.035	0.041	0.044	0.034	0.039
	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)
$Age^2/100$	– 0.001	– 0.001	– 0.001	– 0.001	– 0.001	– 0.001	– 0.001	– 0.001	– 0.054
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
Married	0.143	0.127	0.131	0.134	0.117	0.131	0.134	0.117	0.127
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.002)	(0.002)	(0.004)	(0.002)
High school degree	0.148	0.107	0.115	0.112	0.104	0.113	0.110	0.098	0.104
	(0.003)	(0.002)	(0.002)	(0.003)	(0.005)	(0.002)	(0.003)	(0.005)	(0.002)
Some college	0.241	0.177	0.182	0.178	0.176	0.183	0.178	0.164	0.174
	(0.003)	(0.002)	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.005)	(0.002)
Associates degree	0.289	0.221	0.227	0.223	0.225	0.224	0.220	0.214	0.218
	(0.004)	(0.003)	(0.003)	(0.004)	(0.008)	(0.003)	(0.004)	(0.008)	(0.003)
B.A.	0.519	0.415	0.418	0.412	0.407	0.420	0.414	0.390	0.411
	(0.004)	(0.003)	(0.004)	(0.004)	(0.008)	(0.004)	(0.004)	(0.008)	(0.003)

	Pooled (1)	Pooled (1')	Pooled (2)	Whites (2')	Hispanics (2")	Pooled (3)	Whites (3')	Hispanics (3")	Pooled (4)
Advanced degree	0.727	0.586	0.596	0.587	0.575	0.596	0.587	0.558	0.583
	(0.006)	(0.003)	(0.005)	(0.005)	(0.014)	(0.005)	(0.005)	(0.014)	(0.003)
Live in MSA	0.147	0.008	0.114	0.109	0.087	0.095	0.090	0.085	0.007
	(0.007)	(0.016)	(0.007)	(0.007)	(0.010)	(0.005)	(0.005)	(0.008)	(0.016)
Co-worker characteristics	no	no	yes	yes	yes	yes	yes	yes	no
3-digit industry and establishment size	no	no	no	no	no	yes	yes	yes	no
Establishment fixed effects	no	yes	no	no	no	no	no	no	yes
Observations	396,293	396,293	396,293	327,190	69,103	396,293	327,190	69,103	396,293
R^2	0.47	0.32	0.49	0.45	0.42	0.51	0.47	0.44	0.43

Notes: Robust standard errors are reported in parentheses, based on clustering by establishment. Regressions include state and 1-digit industry fixed effects, and dummy variables for immigration cohorts for Hispanics. Age is adjusted by subtracting off 18, to better capture years in the labor market. The employment size categories are 1-24, 25-49, 50-99, 100-249, 250-499, 500-999, and 1000 and up. When the 3-digit industry controls are added the 1-digit controls drop out. 55,793 establishments are represented in the sample. The R^2 values reported in columns (1') and (4) are for the within-establishment variation. In column (4), the share Hispanic effect for whites is subsumed into the establishment fixed effects.

Having established these baseline estimates, the remaining columns of the table present the more substantive results of our analysis of the impact of segregation by Hispanic ethnicity. Column (2) adds to the basic specification the share of the establishment's workforce that is Hispanic, as well as other demographic characteristics of the establishment's workforce (share in each education category, share married, and the average age of the workforce). The inclusion of the shares in other demographic groups is meant to capture human capital characteristics of co-workers that may affect productivity, and thus wages, and that may be correlated with the share Hispanic variable.[29] The inclusion of the share Hispanic variable parallels the standard specification in the literature on crowding and segregation (*e.g.*, GROSHEN [1991]), and is meant to capture the extent to which crowding of some groups of workers affects wages.[30] The estimated impact of the share Hispanic is – 0.104 and is statistically significant. One way to understand the economic significance of the – 0.104 estimated coefficient on the share Hispanic variable is to note that if a worker moves from an establishment with the average share Hispanic for white workers (0.10) to an establishment with the average share Hispanic for Hispanic workers (0.44), he will incur a wage loss of approximately 3.5 percent.

One qualification that we emphasize at the outset is that our estimates of the effects of the share Hispanic (and, below, the shares with various language skills) are based on across-establishment variation. We do not have changes over time from which we might more reliably identify a causal effect – if we had changing workforce composition but were willing to assumed fixed establishment-level unobservables – nor a more compelling quasi-experiment with a source of exogenous variation in these shares.[31] While we extend the literature in a number of ways that we regard as significant, this limitation of the existing work on segregation remains. We discuss this further below.

In column (2) the estimate of the Hispanic dummy coefficient is – 0.108. This is somewhat lower than the estimate in column (1) where the share Hispanic variable is omitted, but is very close to the estimate in column (1'). These results confirm that the Hispanic-white wage gap is at least partially driven by establishment-level segregation, but that most of the individual-level Hispanic wage penalty is a within-establishment phenomenon. The fact that the coefficient on the Hispanic dummy in column (2) is so close to that in column (1') indicates that the inclusion of the share Hispanic variable in column (2) does a good job of capturing the way in which wages between Hispanics and whites are affected by the establishments in which they work. The coefficients on the individual human capital characteristics in columns (1') and (2) are also very similar. In results not presented, we confirmed that these similarities are driven by the inclusion of the human capital share variables in the

29. Because these variables are meant to capture unobservable human capital characteristics of workers, we do not include standard segregation measures such as share female and share black. We did confirm that including these variables does not markedly affect any of the results for Hispanic or language segregation.

30. Most segregation studies do not contain data at the establishment level, but instead consider segregation at the industry or occupation level. GROSHEN [1991] is one exception.

31. Nonetheless, to avoid cumbersome language we refer to "effects" and "impacts" of our share variables, trusting the reader to keep this qualification in mind. LENGERMANN [2001] uses longitudinal data (from the LEHD) on workers and establishments to try to estimate a human capital co-worker (or peer) effect, conditioning on individual and firm fixed effects.

regression in column (2), again indicating that these share variables do a good job of capturing the way in which co-worker or establishment unobservables are correlated with a workers' own characteristics and wages. Nonetheless, since our focus is on Hispanic-white wage differences, what we most care about is the fact that the effect of the inclusion of the establishment fixed effects on the Hispanic-white wage gap mimics so closely the inclusion of the share Hispanic variable.

In columns (2') and (2") we estimate this specification separately for whites and Hispanics, respectively, dropping the restriction that the effect of segregation by Hispanic ethnicity is the same for whites and Hispanics. The estimated coefficient on the share Hispanic for whites in column (2') is – 0.037, much smaller than that for the pooled sample in column (2); not surprisingly, then, for Hispanic workers the estimated coefficient on the share Hispanic variable is much larger (– 0.168) than for the pooled sample.

The fact that the impact for Hispanic workers of working with Hispanics is negative while for white workers the impact is so close to zero is consistent with a simple model of Hispanic segregation. In this type of model, since only Hispanics are segregated, only Hispanics incur a wage penalty; whites are mobile and wages for whites equalize across the economy so that there is no wage penalty for whites when they work with Hispanics. Moreover, the fact that the effect of the share Hispanic is negative for Hispanic workers is inconsistent with the hypothesis that segregation by ethnicity is a means of reducing transaction costs between dissimilar workers (in this case white and Hispanic), since if the mechanism for ethnic segregation were one based on minimizing transaction costs across ethnic groups, the interaction between the Hispanic dummy and the share Hispanic would be positive, consistent with the idea that the wage penalty for being Hispanic is mitigated if one works with Hispanics.[32]

Of course, the theoretical model does not necessarily imply this parameterization of the effect of co-workers' ethnicity. For example, it is conceivable that only at higher levels of concentrations of Hispanic workers does the effective "language" of communication shift, reducing or eliminating the penalty to being Hispanic. However, when we explored more flexible specifications of the effect of the share Hispanic, the evidence did not point to a diminution of the negative effect of the share Hispanic for Hispanic workers as the share Hispanic rose. Indeed, the estimates indicated that for whites the marginal negative effect of the share Hispanic weakens at higher shares Hispanic, while for Hispanics this marginal negative effect strengthens.[33] Thus, the results indeed are more consistent with models of discrimination-based segregation, where – aside from lower wages within establishments – Hispanics work in disproportionately lower-paid establishments.

It is conceivable that the share Hispanic effects estimated in Table 9, columns (2) and (2'), reflect solely unobservables about the types of establishments in which Hispanics are concentrated. To some extent, of course, the segregation

---

32. Indeed, if this were all that was going on, one might expect the sum of the coefficient on the share Hispanic variable and its interaction with the Hispanic dummy to be positive, suggesting that, among Hispanics, those who work with other Hispanics earn a higher wage.

33. These estimates are based on replacing the simple linear share Hispanic variable with dummy variables for each quartile of the share Hispanic distribution based on Hispanic workers.

argument is about the concentration of Hispanics in low-wage establishments, suggesting that including detailed establishment controls runs the risk of "over-controlling" for exactly the characteristics of the establishments along which segregation occurs. However, in the standard story it is the segregation, per se, that reduces wages, rather than establishment characteristics. Moreover, models of segregation that stem from transaction costs between workers are really models of within-establishment interactions among workers, so a specification that gets as close as possible to isolating within-establishment interactions provides the best test of these types of models. We cannot include the share Hispanic variable and establishment fixed effects simultaneously in our regression specifications, but in the remainder of Table 9 we do two things to assess the robustness of the results to establishment-level differences.

First, we check how much detailed establishment controls known to be strongly associated with wages reduce the negative relationship between wages and the share Hispanic (for Hispanics). In columns (3)-(3") of Table 9 we add detailed controls for industry (3-digit controls) and size (seven categories based on employment) to the specifications reported in columns (2)-(2") (which themselves already contain some important establishment controls such as co-worker human capital characteristics and 1-digit industry controls). The inclusion of these establishment controls further reduces the estimated effect of the share Hispanic in the pooled regression, to a rather small – 0.034. This, however, masks rather sharp differences between whites and Hispanics.

For whites, the estimated impact of the share Hispanic actually becomes positive. Overall, because the estimate of this coefficient in column (2'), excluding these extra establishment controls, is small and negative, and because the controls in column (3') may actually sweep out some of the relevant segregation, we interpret the evidence for whites as suggesting that the effects of the share Hispanic for white workers are best viewed as near zero – or at least there is no clear evidence of a positive or negative effect.

In contrast, for Hispanic workers, although the estimate of share Hispanic falls somewhat (to − 0.099) as we move from column (2") to (3"), evidence of a sizable negative effect of a high concentration of Hispanic workers remains. For the same calculation as above, the estimate implies that if an Hispanic worker moves from an establishment with the average share Hispanic for white workers to an establishment with the average share Hispanic for Hispanic workers, he would incur a wage loss of approximately 3.4%, which is large relative to the typical unexplained Hispanic-white wage differential of approximately 10%. Thus, for Hispanic workers in particular, even within very similar types of establishments there is a substantial wage penalty associated with working with a large share of Hispanics.

Second, although we cannot include establishment fixed effects and still meaningfully estimate the effect of share Hispanic on wages, we can estimate a version of the fixed effects equation reported in column (1') where we interact the share Hispanic with the Hispanic indicator for an individual. This does not allow us to estimate separately the effects of share Hispanic on white workers and Hispanic workers, which are key parameters in models of segregation, but it allows us to estimate the *differential effect* of share Hispanic on Hispanics relative to white workers, holding fixed across Hispanics and whites all other (observed and unobserved) establishment-level characteristics, and

assuming that they impact Hispanics and whites equally. We report the results of this regression in column (4). The reported coefficient of this interaction implies that, conditional on establishment characteristics, Hispanics incur a 21.6% larger wage penalty for working with other Hispanics than do white workers. This result is slightly larger than the difference in penalties of 17.9% one can calculate by comparing columns (3') and (3"), and further demonstrates that it is not unobserved establishment characteristics that generate an Hispanic co-worker wage penalty that is substantially larger for Hispanics than for white workers.

B. **Wages and Segregation by Language Proficiency**

Up to this point, we have not incorporated English language proficiency into the empirical analysis. We begin to do so in Table 10, which follows the key specifications from Table 9, but with the addition of English language proficiency dummy variables. Column (1) of Table 10 reports results for Hispanics from the same specification as column (2") of Table 9, augmented by the dummy variables for English language proficiency; the omitted proficiency category is speaking English very well.[34] The estimates show clearly that, for Hispanics, there are large wage premiums associated with English language proficiency – the wage penalty for speaking English well relative to very well is 5.7%, the wage penalty for speaking English poorly is 16.5%, and the wage penalty for not speaking English at all is 24.9%.[35] The coefficient on the share Hispanic variable in column (1) is – 0.146; this is 2.2 percentage points smaller than its counterpart from Table 9 without the English proficiency variables.

In columns (2) and (2') we report estimates of the specification (now for whites and Hispanics) after adding to the regressions not only the variable representing the share of Hispanic co-workers in an establishment, but also the shares of Hispanic co-workers who do not speak English, who speak English poorly, and who speak English well. The omitted category is the share of Hispanic co-workers who speak English very well. It should be noted that these English language shares add to one as they represent the shares of Hispanics in each category *among Hispanics in the establishment*. The coefficients on these variables can thus be interpreted as the impact of working with varying kinds of Hispanic workers defined by their English language proficiencies, conditional on working with a given share Hispanic.

The estimated coefficients on the share Hispanic variable are similar to those in column (2') and (2") of Table 9 (– 0.041 for whites and – 0.141 for Hispanics), indicating a substantial wage penalty for working alongside Hispanics only for Hispanics. More importantly, once we condition on the Hispanic composition of the establishment and the worker's own language proficiency, there is also evidence of wage penalties for Hispanics who work with Hispanics with worse English language proficiency, with the penalty rising monotonically as the language skills of co-workers fall. For example, the

34. There is no corresponding new specification for whites because only those who speak English very well are in the sample.

35. These are larger still if the immigration cohort variables are excluded. They are a shade smaller when establishment fixed effects are included.

TABLE 10

Log Wage Regressions Including Language Skills
Men from DEED in Establishments with At Least One Matched Hispanic Worker

	Hispanics (1)	Whites (2)	Hispanics (2')	Whites (3)	Hispanics (3')
Share of co-workers who are Hispanic	– 0.146	– 0.041	– 0.141	0.070	– 0.073
	(0.009)	(0.012)	(0.009)	(0.012)	(0.009)
Hispanic × don't speak English	– 0.249	–	– 0.206	–	– 0.214
	(0.009)		(0.011)		(0.011)
Hispanic × speak English poorly	– 0.165	–	– 0.137	–	– 0.144
	(0.006)		(0.007)		(0.007)
Hispanic × speak English well	– 0.057	–	– 0.039	–	– 0.041
	(0.005)		(0.005)		(0.005)
Share of Hispanic co-workers who:					
Don't speak English	–	0.067	– 0.094	0.077	– 0.077
		(0.020)	(0.017)	(0.019)	(0.017)
Speak English poorly	–	0.008	– 0.056	0.020	– 0.031
		(0.009)	(0.010)	(0.008)	(0.010)
Speak English well	–	– 0.007	– 0.039	0.003	– 0.019
		(0.006)	(0.008)	(0.005)	(0.008)
Co-worker characteristics	yes	yes	yes	yes	yes
3-digit industry and establishment size	no	no	no	yes	yes
Observations	69,103	327,190	69,103	327,190	69,103
R^2	0.42	0.45	0.43	0.47	0.45

Notes: Robust standard errors are reported in parentheses, based on clustering by establishment. Regressions include state and 1-digit industry fixed effects, and dummy variables for immigration cohorts for Hispanics. The employment size categories are 1-24, 25-49, 50-99, 100-249, 250-499, 500-999, and 1000 and up. When the 3-digit industry controls are added the 1-digit controls drop out. 55,793 establishments are represented in the sample. The other controls listed in Table 9 are also included, but not reported here.

estimated effect of the share of Hispanic co-workers who do not speak English is – 0.094, implying that an Hispanic worker moving from an establishment with the average share in this category for those who speak English well (0.013) to the average share for those who do not speak English (0.481) would suffer a wage decline of 4.4%. The estimated effect of a higher share who speak English poorly is just above half of this, and the estimated effect of a higher share who speak English well (again, relative to very well) is about one-third as large. At the same time, the substantial wage penalties associated with a worker's own poor English proficiency and the overall share Hispanic persist, although falling a bit, indicating that these latter wage penalties are for the most part not attributable to language segregation. Interestingly, for whites there are no such penalties to working with Hispanics who have poor English proficiency; in fact, there is a positive premium for whites who work with Hispanics who do not speak English. This result could stem from co-worker discrimination that is strongest for minorities who do not speak English, or it may arise if whites tend to be in supervisory positions when working alongside many Hispanics with poor English skills.

In the remaining columns ((3) and (3')) we add the detailed establishment controls. Once again, the share Hispanic coefficient turns positive for whites, but remains negative for Hispanics, and the difference in the coefficients between whites and Hispanics is larger (0.143) between columns (3) and (3') than between (2) and (2') (0.100). The co-worker language share variables for whites (Hispanics) are slightly more positive (less negative) with the inclusion of the detailed establishment controls, but once again the differences in the coefficients between whites and Hispanics remains large between the columns. Thus, again, the evidence points to Hispanics having wages lowered via a higher share Hispanic, and even more so via a higher share with poor language skills.

Finally, since the specifications in Table 10 are in some respects our key estimates of the impact of segregation by ethnicity and language, we use these specifications to return to the issue of measurement error that arises from estimating the shares of workers in ethnic and language skill categories. Thus far, the estimates have been based on all establishments with at least one matched Hispanic worker. However, in small plants or plants with few matched workers, the estimated share Hispanic, and even more so the estimated shares of Hispanic workers in each language category, may be estimated quite imprecisely. Offsetting this to some extent, though, is the implicit higher weight given to establishments with more matched workers, since these establishments contribute more workers to the sample. To assess whether the findings are influenced by measurement error from sampling, we re-estimated the final specifications in Table 10 using, first, establishments with 250 or more workers, and second, establishments with at least 20 matched workers. It is the latter number that determines how precisely the various shares are reported, but obviously the number of matched workers is closely related to the size of the establishment. We focus on the specifications with the full set of controls to isolate the effect of sampling, and avoid the confounding influence of differences in other characteristics of larger establishments.

The results are reported in Table 11. We note a couple of features of the results prior to turning to the coefficient estimates. First, the sample size for Hispanics falls by nearly two-thirds, while that for whites falls by less than one-third, reflecting the employment of Hispanic workers in smaller establishments. Second, the R^2 values for the Hispanic sample, in particular, rise considerably, which we would expect from more accurate measurement of the share variables, but which could also reflect greater homogeneity of the samples, and perhaps less unexplained variation in larger establishments. The estimated coefficients for the share Hispanic for white workers are larger than the comparable estimates in column (3) of Table 10, although the estimates of this coefficient have not been robust throughout the analyses we report; the difference may reflect more sizable supervisory wage gaps in larger establishments.

But the key results for Hispanics regarding the share Hispanic, and more importantly the effects of the shares of Hispanics by language skill, are quite robust. These results generally indicate that Hispanic men earn less when they work alongside other Hispanics, and in particular alongside Hispanics with poor language skills. Furthermore, the estimated magnitudes are similar to those in Table 10. It is noteworthy that the estimated effects of these shares on Hispanic workers are not uniformly larger when we restrict the observations to workers in larger establishments or those with many matched workers,

TABLE 11

Log Wage Regressions Including Language Skills
Men from DEED in Establishments with At Least One Matched Hispanic Worker, Alternative Establishment Size Cutoffs

	Establishments with at least 250 workers		Establishments with at least 20 matched workers	
	Whites (2)	Hispanics (2')	Whites (3)	Hispanics (3')
Share of co-workers who are Hispanic	0.198	– 0.055	0.191	– 0.080
	(0.024)	(0.025)	(0.026)	(0.033)
Hispanic × don't speak English	–	– 0.202	–	– 0.169
		(0.022)		(0.025)
Hispanic × speak English poorly	–	– 0.155	–	– 0.141
		(0.012)		(0.013)
Hispanic × speak English well	–	– 0.051	–	– 0.052
		(0.008)		(0.008)
Share of Hispanic co-workers who:				
Don't speak English	0.095	– 0.114	0.017	– 0.132
	(0.037)	(0.047)	(0.039)	(0.067)
Speak English poorly	(0.020)	0.012	0.030	0.023
	(0.013)	(0.026)	(0.014)	(0.033)
Speak English well	0.006	– 0.022	0.007	– 0.002
	(0.008)	(0.018)	(0.008)	(0.019)
Co-worker characteristics	yes	yes	yes	yes
3-digit industry and establishment size	yes	yes	yes	yes
Observations	230,852	21,217	241,232	19,110
R^2	0.48	0.50	0.48	0.50

Notes: See notes to Table 10. Specifications are comparable to those in columns (3) and (3') of Table 10.

which we would expect from simple measurement error. While this could stem from differences in the underlying coefficients in establishments of different sizes, another interpretation is that some of the other differences, such as for white workers, stem from real differences in wage outcomes rather than measurement error bias. Overall, while we believe the issue of measurement error in matched employer-employee data sets using samples of workers merits serious attention, it does not appear to have much impact on the results we report in this paper.

C. **Interactions between Own and Co-worker Language Proficiency**

The specifications in Table 10, while variants of common specifications in the literature on segregation, do not adequately capture possible effects of own and co-workers' language that may arise if language segregation serves to lower transaction costs by allowing one to communicate with those who speak the same language. Such an hypothesis can be captured in our data by allow-

ing for interactions between one's own English language proficiency and the English language proficiencies of one's co-workers. A specification like this may reveal a wage advantage (at least to some extent offsetting the negative impacts of one's own ethnicity and language deficiencies) associated with having co-workers whose language ability is similar to one's own or who speak the same language. Rather than incorporating and interpreting a large series of interaction terms in the specifications in Table 10, we instead turn to examining the impact of ethnic and language segregation separately by English language proficiency. This is equivalent to incorporating into the regressions full sets of interactions between one's own ethnicity and language with not just the characteristics of one's co-workers, but also with all the human capital characteristics (and other controls) included in the wage equations.

As a preliminary specification, in Table 12, we examine the effect on wages of establishment-level segregation by Hispanic ethnicity for Hispanics, now separating Hispanics by English language proficiency. In column (1) we report results for the sample of Hispanics who speak English very well, including (as in Table 10) the share Hispanic variable. Panel A includes the baseline controls and co-worker human capital characteristics, whereas Panel B also adds the detailed establishment controls. In columns (2)-(4) we report the results for the three other language categories.

Before discussing the results for the effects of segregation by Hispanic ethnicity, we note that as one moves down the English proficiency scale the returns to human capital characteristics fall almost monotonically, and often quite dramatically. While the returns to these characteristics for Hispanics who speak English very well are quite similar to those for white men, the returns are much smaller for Hispanics who speak English poorly or not at all. This basic pattern has two important implications. First, from a practical standpoint, it makes the construction of standard Oaxaca-Blinder-type decompositions of the wage differences between white men and Hispanic men of varying English language proficiencies extremely sensitive to the arbitrary choice of whose returns are considered the "base" group of returns. That is, one may get a very different sense of the importance of establishment-level segregation when one asks the hypothetical question, "how much would an Hispanic who does not speak English well get paid if he had his human capital and establishment characteristics but were paid for those like a white man?" versus the hypothetical question, "how much would a white man get paid if he had his human capital and establishment characteristics but were paid for those like an Hispanic who does not speak English?" As a result, we have chosen not to present these type of decompositions. Second, from a research standpoint, we find the differences in returns to human capital characteristics by English language proficiency fascinating, and we have found virtually no references to this phenomenon elsewhere.[36] Some of these differences may be due to the fact that the majority of poor English speakers in our sample are immigrants, whose human capital may have been attained in foreign countries. However,

36. TREJO [1997] reports lower returns to education and experience for immigrant whites and Mexicans, which in the literature on immigration is interpreted as indicating that the human capital immigrants acquired in their home countries does not transfer perfectly to the U.S. labor market.

TABLE 12

Log Wage Regressions by Ethnicity and Language Category Hispanic Men from DEED in Establishments with At Least One Matched Hispanic Worker

	Speak English:			
	Very well (1)	Well (2)	Poorly (3)	Not at all (4)
A. With basic controls and co-worker characteristics				
Share of co-workers who are Hispanic	– 0.146	– 0.149	– 0.089	– 0.092
	(0.011)	(0.017)	(0.020)	(0.039)
Age	0.042	0.028	0.017	0.008
	(0.001)	(0.001)	(0.001)	(0.002)
$Age^2/100$	– 0.001	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Married	0.128	0.109	0.085	0.062
	(0.005)	(0.009)	(0.010)	(0.018)
High school degree	0.108	0.070	0.020	0.059
	(0.006)	(0.009)	(0.013)	(0.028)
Some college	0.169	0.138	0.059	0.130
	(0.007)	(0.011)	(0.018)	(0.038)
Associates degree	0.233	0.133	0.071	**
	(0.010)	(0.018)	(0.029)	**
B.A.	0.402	0.320	0.128	**
	(0.009)	(0.023)	(0.041)	**
Advanced degree	0.595	0.354	0.118	**
	(0.015)	(0.038)	(0.050)	**
Live in MSA	0.100	0.061	0.110	0.009
	(0.010)	(0.017)	(0.025)	(0.047)
R^2	0.42	0.32	0.30	0.23
B. Add 3-digit industry and establishment size controls				
Share of co-workers who are Hispanic	– 0.064	– 0.098	– 0.050	– 0.067
	(0.011)	(0.018)	(0.021)	(0.040)
R^2	0.45	0.36	0.33	0.30
Observations	41,812	14,256	10,092	2,943

Note: Robust standard errors are reported in parentheses, based on clustering by establishment. Basic controls include state and 1-digit industry fixed effects, and dummy variables for immigration cohorts. Panel B includes the controls reported in Panel A. Age is adjusted by subtracting off 18, to better capture years in the labor market. The employment size categories are 1-24, 25-49, 50-99, 100-249, 250-499, 500-999, and 1000 and up. When the 3-digit industry controls are added the 1-digit controls drop out. 55,793 establishments are represented in the sample. ** indicates too few observations in cell to report statistics.

in unreported results from the full sample of the SEDF, where the sample sizes are large enough to allow estimation of human capital regressions by English language proficiency for U.S. natives alone, we find similar results, at least for returns to age and marriage. We plan to explore this phenomenon more fully in future research.

In Panel A, column (1), of Table 12, for Hispanic workers who speak English very well, the coefficient on the share Hispanic is – 0.146; these Hispanic workers make up more than half of the sample of Hispanics. The coefficient on the share Hispanic for workers who speak English well is – 0.149 as reported in column (2); it falls to – 0.089 in column (3) for those who speak English poorly, and it is – 0.092 for those who do not speak English. The fact that the magnitude of this coefficient declines as one moves from those who speak English very well or well to those who speak English poorly or not at all, combined with the decreases in returns to human capital, suggest to us that perhaps Hispanic workers with poor English language proficiency simply have poor job opportunities, regardless of their own human capital and the ethnic composition of their co-workers, so that there is much less room for variation in these factors to affect their wages. There is some evidence supporting this in Panel B, where the inclusion of detailed establishment controls eliminates systematic differences in the effect of the share Hispanic between the groups. However, regardless of the specification, the negative effect of share Hispanic persists across all language categories, indicating that, regardless of English language proficiency, there is a wage penalty associated with working with other Hispanics.

We examine the impact of segregation by English language proficiency along with Hispanic segregation separately by language proficiency groups in Table 13. The specifications are the same as in Table 12 except that we now add the shares of Hispanic co-workers in each language category. Looking first at Panel A, the coefficients on the Hispanic English language proficiency shares do not display a clear pattern as one moves across columns (1)-(4). All but two of the coefficients are negative, and most are statistically significant. The impact of working with a large share of Hispanic workers who do not speak English is negative for all four groups, and a higher share of workers in this lowest language category is always associated with the largest wage penalties. However, this effect is not particularly weaker (and certainly not positive) for workers who do not speak English. Similarly, the negative impact of working with a large share who speak English poorly is strongly negative for three of the four groups. Again, though, this wage penalty is not weaker for those who speak English poorly or not at all. If anything, it appears that there is some evidence that the penalties for working with Hispanics with poor English proficiency are somewhat stronger for those whose own language proficiency is weak. The same qualitative results hold in Panel B, although paralleling the earlier results the wage penalties overall are a bit smaller.

These results do not provide evidence that, for Hispanic workers, there are wage advantages (which may offset some of the wage penalties associated with own Hispanic ethnicity and weak own language proficiency) to working with Hispanics of similar language ability. For example, if there are transaction costs associated with language and these impact wages, one would expect to find a positive interaction for Hispanic non-English speakers when they work with other Hispanics who do not speak English, all else

TABLE 13

Log Wage Regressions Including Language Skills
Hispanic Men from DEED in Establishments with At Least One Matched Hispanic Worker

	Speak English:			
	Very well (1)	Well (2)	Poorly (3)	Not at all (4)
A. With basic controls and co– worker characteristics				
Share of co-workers who are Hispanic	– 0.146	– 0.162	– 0.110	– 0.137
	(0.011)	(0.018)	(0.020)	(0.040)
Share of Hispanic co-workers who don't	– 0.068	– 0.156	– 0.186	– 0.168
speak English	(0.037)	(0.039)	(0.033)	(0.037)
Share of Hispanic co-workers who	0.006	– 0.097	– 0.136	– 0.107
speak English poorly	(0.020)	(0.024)	(0.019)	(0.044)
Share of Hispanic co-workers who	0.028	– 0.119	– 0.028	– 0.085
speak English well	(0.014)	(0.014)	(0.026)	(0.051)
R^2	0.42	0.32	0.31	0.24
B. Add 3-digit industry and establishment size controls				
Share of co-workers who are Hispanic	– 0.058	– 0.110	– 0.074	– 0.125
	(0.011)	(0.019)	(0.022)	(0.043)
Share of Hispanic co-workers who don't	– 0.087	– 0.115	– 0.169	– 0.165
speak English	(0.037)	(0.039)	(0.033)	(0.041)
Share of Hispanic co-workers who speak	– 0.010	– 0.056	– 0.098	– 0.080
English poorly	(0.020)	(0.024)	(0.020)	(0.045)
Share of Hispanic co-workers who speak	– 0.016	– 0.058	– 0.012	– 0.070
English well	(0.014)	(0.016)	(0.026)	(0.053)
R^2	0.45	0.36	0.33	0.30
Observations	41,812	14,256	10,092	2,943

Note: Robust standard errors are reported in parentheses, based on clustering by establishment. Basic controls include state and 1-digit industry fixed effects, dummy variables for immigration cohorts, and the other controls listed in Table 9. The employment size categories are 1-24, 25-49, 50-99, 100-249, 250-499, 500-999, and 1000 and up. When the 3-digit industry controls are added the 1-digit controls drop out. 55,793 establishments are represented in the sample.

equal. This is inconsistent with the finding in column (4) that, for Hispanic workers who do not speak English, the coefficient on the share Hispanic who do not speak English is negative and significant, and indeed larger than the estimated coefficients of the other language share variables.

Instead, these results are entirely consistent with an alternative discrimination-type crowding model where employers crowd Hispanics with poor English ability into a restricted set of establishments, lowering the marginal products and hence wages of all Hispanics who work in those establish-

ments.[37] On the other hand, there may be transaction costs of the type associated with the language discrimination model, but enough labor market mobility and hence sensitivity to the outside market such that the characteristics of co-workers do not affect wages, yet poorer language skills are still associated with lower wages. But this is hard to reconcile with the negative effects on wages of the share of co-workers who are Hispanic or who have poor language skills.

D. **Possible Biases in Measuring Language Skills**

There are two potential sources of bias in our estimated effects of own language skills and co-worker language skills. First, the effects of own language may be upward biased because of underlying heterogeneity (ability?) that is correlated in the same direction with both language skills and wages (*e.g.*, BORJAS [1994]). On the other hand, in part because language skills are self-reported, they may be measured with error. To the extent that results from classical measurement error carry over, this would be expected to bias the estimated effects of own language skills toward zero (DUSTMANN and VAN SOEST, 2002). On net, the overall direction of bias is unclear, as these two biases are offsetting. Of more interest to us than the estimated effects of own language skills are possible biases in the estimated effects of the language skills of co-workers. But given the segregation by language skills that we document, own language skills and co-worker skills are positively correlated, so the bias in the estimated effect of co-worker language skills is in the opposite direction to the bias in the estimated effect of own language skills.

Using German panel data, Dustmann and van Soest propose some estimators to control for these sources of bias in own language skills. They conclude that downward bias from measurement error is quantitatively more important. If this were true in our data as well, then the estimated effects of co-worker language skills may be overstated. However, we believe there are some legitimate questions about the identification strategy used by Dustmann and van Soest,[38] and it is not clear that the results generalize to the United States.

We can also look at the estimated returns to schooling for Hispanics for evidence of downward bias in the estimated effects of own language skills. As Table 8 showed, education is strongly positively correlated with better language proficiency, so that measurement error bias in own language skills would also bias upward the estimated returns to schooling for Hispanics. Yet in Table 12 the estimated returns to schooling are a shade weaker for Hispanics than for whites, and the estimates decrease as English proficiency falls, suggesting that the strong effects of co-worker language skills for

37. Recall that within establishments the impact of crowding is greater on wages of Hispanics.

38. For example, their instrumental variables estimation relies on leads and lags of reported own language skills serving as valid instruments. But if the past time pattern of accumulation of language skills, or the expected future pattern, has anything to do with current wages, these instruments are not valid. This seems difficult to rule out. For example, a job that offered wage growth in the past may have offered stronger incentives for language acquisition, and a job that offers opportunities for language acquisition in the future may pay a lower wage currently.

Hispanics are real, rather than spuriously stemming from measurement error.[39]

Finally, to partially assess the influence of measurement error on the estimated effects of co-worker language skills, we re-estimated some of the basic models after collapsing the four categories for own language into two broader categories, on the assumption that there may well be misclassification error between the "not at all" and "poor" categories, and between the "well" and "very well" categories, but that misclassification across these two broader groups is far less likely. For the co-worker shares this measurement error should average out, so we do not collapse their language categories. If the effects of poor language proficiency of co-workers do not weaken upon doing this, we might conclude that measurement error in own language skills does not impart a quantitatively important bias to the estimated effects of co-worker language skills. In the resulting estimates, the effects of co-worker language skills on wages of Hispanics did not weaken and if anything generally became somewhat stronger. This is not a clean test, however, as the combination of categories for own language skills may entail specification error that, by aggregating individuals with different language skills, leads to overly strong estimated effects of co-worker language skills. But it at least provides some evidence that estimates of the effects of co-worker language skills are not extremely sensitive to measurement error in own language skills. While we cannot assert that we have unbiased estimates of the effects of co-worker language skills, we believe we have rather compelling evidence that wages of Hispanic workers are lowered, possibly substantially, when they work alongside other Hispanics with low English proficiency.

5 The Impact of Segregation in California and Florida

In this paper we examine the role of segregation by Hispanic ethnicity and by English language proficiency. Clearly, by defining segregation only along these dimensions we may be missing the impacts of segregation along other dimensions that define the identities of Hispanic workers and that may be linked to transaction costs of communication. Because Hispanics in the United States come from very divergent countries of origin, ranging from Mexico to Spain, treating Hispanics as a monolithic group obscures potentially important differences between them, such as in human capital and culture, that may affect how they communicate with each other and with whites in the workplace.

Examining the role of alternative definitions of identity for Hispanic workers is beyond the scope of this paper, so instead, in this section, we replicate the results we presented above separately using data from California and

39. Of course this is not decisive, as the true return to schooling for Hispanics may be lower than is indicated by the estimates.

Florida. This serves two purposes. First, it provides a robustness check for the full U.S. sample results. Second, it suggests important avenues for future research into the importance of language and ethnicity for Hispanics in the workplace, since as we show below, Hispanic workers in California and Florida are quite different along one important dimension.

In Table 14 we present means for the sample of men from California and Florida who are used in the full U.S. sample results. Column (1) presents means for white men from California, and the corresponding means for white men from Florida are in column (3). Column (2) presents means for Hispanics from California and its counterpart for Florida is column (4). California is a much bigger state than Florida and therefore represents a much bigger fraction of the U.S. sample that we use. Even given this, Californians make up a disproportionately large fraction of our Hispanic sample; over 38% (26,126) of the full sample of Hispanics that we use come from California, whereas 6% of the full sample of Hispanics come from Florida.

The first thing to note about the differences between California and Florida is the distributions of educational attainment by ethnicity. White men in California are more educated than white men in Florida,[40] 77.5% of the white men from California have more than a high school education as compared to 69.9% of the white men from Florida. In contrast, the Hispanic men in our sample from California are less educated, on average, than the Hispanic men from Florida. In California, only 32.6% of the Hispanic men have more than a high school degree, while in Florida 50.3% of the Hispanic men have more than a high school degree. The differences for Hispanics are particularly stark when one considers that 19% of Florida Hispanics have at least a B.A., while only 8.2% of California Hispanics have a B.A. or advanced degree. These differences in educational attainment may help explain the fact that average wages for white men in California are higher than for white men in Florida, whereas average wages for Hispanics in California are lower than for Hispanics in Florida. In addition to differences in educational attainment, Hispanic men in our sample in California are, on average, four years younger than Hispanic men in Florida (34.7 versus 38.9), and both white and Hispanic men are slightly less likely to be married in California than in Florida.

In addition to differences in observable characteristics of individuals between the states, there is evidence that California establishments are less segregated by Hispanic ethnicity than Florida establishments. For example, for white men in California, the average share of co-workers who are Hispanic is 20.4%; for white men in Florida the corresponding figure is 16.9%. Conversely, for Hispanic men in California the average share of co-workers who are Hispanic is 51.1%, whereas in Florida it is over 6 percentage points higher, at 57.5%. Differences in the patterns of segregation by English language proficiency across the states are much less pronounced, and the standard deviations of the Hispanic English language proficiency shares are large.

In Table 15 we present results from wage regressions for California and Florida for whites and Hispanics separately, using the specification where we include the establishment share Hispanic and the establishment shares of

40. Recall that the samples are limited to men aged 16-65 who were working full-time in the previous year, so retirees are excluded from the sample.

TABLE 14

Means for Men from DEED in Establishments with At Least One Matched Hispanic Worker, California and Florida

	California		Florida	
	White (1)	Hispanic (2)	White (3)	Hispanic (4)
Log(wage)	2.886	2.305	2.656	2.303
	(0.550)	(0.530)	(0.573)	(0.559)
Age	39.816	34.719	39.142	38.903
	(10.984)	(10.674)	(10.905)	(12.401)
$Age^2/100$	17.059	13.193	16.510	16.672
	(9.221)	(8.309)	(9.054)	(10.295)
Married	0.685	0.657	0.716	0.717
	(0.465)	(0.475)	(0.451)	(0.451)
High school degree	0.171	0.216	0.230	0.208
	(0.377)	(0.411)	(0.421)	(0.406)
Some college	0.261	0.185	0.239	0.215
	(0.439)	(0.388)	(0.426)	(0.411)
Associates degree	0.098	0.059	0.107	0.098
	(0.297)	(0.236)	(0.310)	(0.297)
B.A.	0.273	0.058	0.247	0.128
	(0.445)	(0.234)	(0.431)	(0.335)
Advanced degree	0.143	0.024	0.106	0.062
	(0.350)	(0.152)	(0.308)	(0.241)
Live in MSA	0.980	0.984	0.966	0.992
	(0.138)	(0.127)	(0.182)	(0.088)
Hispanic × don't speak English		0.058		0.074
		(0.234)		(0.262)
Hispanic × speak English poorly		0.186		0.151
		(0.389)		(0.358)
Hispanic × speak English well		0.211		0.205
		(0.408)		(0.404)
Share of coworkers who are Hispanic	0.204	0.511	0.169	0.575
	(0.168)	(0.297)	(0.183)	(0.322)
Share of Hispanic coworkers who:				
Don't speak English	0.013	0.063	0.013	0.076
	(0.083)	(0.161)	(0.088)	(0.183)
Speak English poorly	0.061	0.184	0.049	0.152
	(0.171)	(0.265)	(0.169)	(0.247)
Speak English well	0.137	0.200	0.152	0.202
	(0.234)	(0.263)	(0.279)	(0.273)
Speak English very well	0.789	0.552	0.786	0.570
	(0.301)	(0.381)	(0.335)	(0.378)
Share of co-workers with:				
High school degree	0.200	0.224	0.263	0.242
	(0.161)	(0.203)	(0.172)	(0.223)

	California		Florida	
	White (1)	Hispanic (2)	White (3)	Hispanic (4)
Some college	0.266	0.225	0.250	0.227
	(0.151)	(0.202)	(0.145)	(0.215)
Associates degree	0.096	0.068	0.112	0.098
	(0.092)	(0.112)	(0.107)	(0.146)
B.A.	0.220	0.113	0.197	0.135
	(0.163)	(0.158)	(0.153)	(0.179)
Advanced degree	0.101	0.040	0.071	0.054
	(0.119)	(0.094)	(0.099)	(0.120)
Average age of worker	20.450	18.517	20.439	21.585
	(5.122)	(6.628)	(4.901)	(7.707)
Share of workers married	0.631	0.624	0.659	0.658
	(0.175)	(0.242)	(0.170)	(0.250)
Observations	49,170	26,126	8,554	4,145

Note: Standard deviations are reported in parentheses.

TABLE 15

Log Wage Regressions Including Language Skills
Men from DEED in Establishments with At Least One Matched Hispanic Worker

	California		Florida	
	Whites (1)	Hispanics (2)	Whites (3)	Hispanics (4)
Share of co-workers who are Hispanic	-0.046	-0.106	0.161	0.081
	(0.021)	(0.015)	(0.040)	(0.028)
Hispanic × don't speak English		-0.193		-0.228
		(0.015)		(0.039)
Hispanic × speak English poorly		-0.123		-0.210
		(0.010)		(0.031)
Hispanic × speak English well		-0.008		-0.109
		(0.009)		(0.026)
Share of Hispanic co-workers who:				
Don't speak English	0.067	-0.134	0.111	-0.072
	(0.037)	(0.024)	(0.052)	(0.060)
Speak English poorly	-0.004	-0.090	0.032	-0.080
	(0.017)	(0.016)	(0.044)	(0.043)
Speak English well	-0.001	-0.045	-0.070	-0.065
	(0.012)	(0.014)	(0.020)	(0.037)
Observations	49,170	26,126	8,554	4,145
R^2	0.39	0.45	0.38	0.37

Notes: Robust standard errors are reported in parentheses, based on clustering by establishment. Regressions include co-worker human capital characteristics, state and 1-digit industry fixed effects, dummy variables for immigration cohorts, and the other controls listed in Table 9.

Hispanics in each English language proficiency category (along with the other co-worker characteristics and 1-digit industry controls). These results correspond to the full sample results in Table 10 (columns (2') and (2")).[41] The key result in Table 15 is the differences in the coefficients on the share of co-workers who are Hispanic across the two states. The coefficient on the share Hispanic for whites in California is – 0.046, similar to that for the full United States. For Hispanics in California, the coefficient on share Hispanic (– 0.106) is smaller than that for the full United States, but is still large in magnitude. That is, just as in the full U.S. sample, these specifications point to a heavy wage penalty for Hispanics when they work in an establishment that is heavily Hispanic, but a much smaller penalty for whites.

The estimated coefficients for the share Hispanic in Florida are starkly different from those in California and the full sample. For both whites and Hispanics the coefficient on the share of co-workers who are Hispanic is positive and statistically significant (0.161 for whites, and 0.081 for Hispanics). In Florida, then, there is no evidence that segregation by Hispanic ethnicity has a negative impact on wages, and in fact for whites there is sizable a wage premium for working with Hispanics. In contrast to the results for the share Hispanic variable, the results on the coefficients of the English language proficiency shares are reasonably consistent between California and Florida.

The differences between California and Florida in the impact on wages of segregation by Hispanic ethnicity are consistent with the notion that Hispanic ethnicity is not a rich enough description of these workers to capture important features of labor market segregation or communication among workers. Indeed in our sample, as in the full SEDF, the composition of Hispanics in California and Florida is markedly different along the dimension of nativity, or ancestry. In Table 16 we present country or region of ancestry for the Hispanics in our sample from the two states. In California, 79% of the sample of Hispanics are of Mexican origin, whereas in Florida only 5% are. In contrast, 54% of the Florida sample of Hispanics are of Cuban origin whereas only 1% of the Hispanic workers in California are of Cuban origin. We leave it to future research to establish whether these differences in country of origin for Hispanics drive the differences in the impact of Hispanic segregation on wages, but no one would question that the political experiences and cultures of Cuba and Mexico, and of Cuban and Mexican immigrants, are quite different.

6 Conclusions and Future Research

In this paper we document the construction of a large, new matched employer-employee data set for the United States for 1990. We use this data set – the 1990 DEED – to establish a wide array of empirical facts regarding the importance of workplace segregation by Hispanic ethnicity and language proficiency. In particular, we document strong segregation along these dimensions. We

41. We do not present results for California and Florida separately by English language proficiency category because the sample sizes become small.

TABLE 16
Hispanic Ancestry Composition

	California (1) %	Florida (2) %
Central America	8.30	8.87
Cuba	1.26	53.61
Mexico	78.86	5.28
Puerto Rico	1.74	12.57
Other	9.84	19.66

then explore the influences of segregation by Hispanic ethnicity and English language proficiency on wages. The evidence points to substantial wage penalties for Hispanics when they are segregated into workplaces with high shares Hispanic, and particularly with high shares of Hispanics with poor English language proficiency.

In part to shed some light on the sources of segregation, we also consider whether the patterns of wage penalties appear to be consistent with transaction costs that might make it more efficient to group workers with poor English language proficiency together to minimize communication problems, and fail to find such evidence. Some might argue that we do not provide an explicit test of the transaction cost hypothesis, because in a labor market with enough mobility the wage penalties associated with transaction costs stemming from co-workers' language skills would equalize across establishments. However, such a view is difficult to reconcile with the estimated negative impact of the share Hispanic or the share with poor language skills on individual workers' wages. On the surface, then, the data are more consistent with crowding of Hispanics and those with poor language skills into a set of jobs, which results in low wages (and presumably low marginal products) for workers in those jobs.

Another perspective on the negative effects of poor language skills of Hispanic co-workers on Hispanics is that there are negative externalities associated with language skills. This is potentially quite interesting, because if poor language skills impose costs on others, the private incentives to invest in language skills will be too low, and efficient levels of language acquisition will not be achieved.

Aside from the particular results reported here, the richness of the data we have constructed and the empirical findings we present suggest many avenues for research that we plan to pursue. First, there are many ways to define language, and we will explore the impact of alternative definitions of language on the wages of Hispanics and other workers in future work. In particular, since the SEDF contains information on language spoken at home, we will define more finely the spoken languages of workers in the workplace, so that an establishment where no one speaks English but everyone speaks Spanish would not be treated similarly to an establishment where no one speaks English and workers speak a variety of languages. Moreover, we will examine whether cultural norms are a better way to define language differences in the

workplace for Hispanics, by disaggregating Hispanics by country of origin.

Second, there are natural extensions to the empirical example we present here. In Lang's model of language transaction costs, the cost is incurred when managers must speak with workers, and where managers speak the majority language but workers may not. In the empirical work we present here, we do not distinguish between managers and workers in defining the group with whom a worker must communicate. Because the DEED contains information on occupation, it may be possible to group workers into managerial and non-managerial roles within an establishment and more formally test the specifics of the model. Perhaps more importantly, it is not just with co-workers or managers that a worker may need to communicate. If firms in industries such as retail trade sell to workers who live near to these firms, the language of trade is likely to be the language of the local residents (see LAZEAR [2000] for a formal model of trade with multiple languages), and there can be transaction costs in communicating with customers. Because of this, wage penalties associated with poor English language ability are likely to be smaller when a worker speaks the (non-English) language of customers. Since we know the business addresses of establishments in the DEED, and we know the composition of nearby residents to these establishments from the SEDF, we can limit the sample to workers who work in industries that are likely to serve a local population (such as small retail establishments), and use measures of the language composition of nearby residents to these establishments to explore whether there is a positive wage premium associated with speaking the language of one's customer base, or whether this mitigates any negative wage effects associated with lower English language proficiency. Such evidence may also bear on hypotheses related to customer discrimination (BECKER [1971]).

In previous work (HELLERSTEIN and NEUMARK [1999]; HELLERSTEIN, *ET AL.* [1999]), we used earlier matched data on manufacturing establishments to test for wage discrimination against women by specifying and estimating a production function in order to recover estimates of marginal productivity differentials between workers, which we then compared to estimated wage equations. For manufacturing establishments in the DEED we can perform similar analyses examining whether wage differences by race, ethnicity, and language proficiency can be explained by productivity differences, and more specifically asking whether the transaction cost hypothesis receives empirical support from the production side.

Finally, from our standpoint, the most important piece of our future research agenda using the DEED is to explore the link between residential and workplace segregation. The empirical results we present here suggest the possibility that Hispanics with poor English language skills may have poor labor market opportunities, leading to lower returns to human capital and smaller effects of ethnic segregation. There are many possible reasons why these workers may have poor labor market opportunities, but one leading theory that has been proposed is that of "spatial mismatch" (*e.g.*, KAIN [1968]). In spatial mismatch models, residential segregation exists in a way that prevents some workers from having access to good jobs, lowering wages and employment levels for these workers. The DEED is uniquely suited to examining the link between residential and workplace segregation, not just for Hispanics but for blacks as well, because as a match between a household data set and an establishment data set, the DEED contains information on residential addresses and

workplace addresses 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 blacks and Hispanics.

We have laboriously and carefully constructed the 1990 DEED in order to be able to further our research agenda and assist the Census Bureau in meeting some of its objectives,[42] and we have plans to construct the corresponding 2000 version of the DEED when the Long-Form information from the 2000 Decennial Census is available at the Census Bureau. We are confident that the quality and scope of the data we have constructed will allow us to gain new insights into the mechanisms and importance of segregation in the labor market. ▼

42. The 1990 DEED is the property of the U.S. Census Bureau, and while proprietary, is available to other researchers who meet the Census Bureau criteria for restricted-use data.

• Références

ASHENFELTER O. C., CARD D., eds. (1999). – *Handbook of Labor Economics, Vols. 3A-3C* (Amsterdam: Elsevier Science Publishers).

ASHENFELTER O. C., LAYARD R., eds. (1986). – *Handbook of Labor Economics, Vols. 1-2* (Amsterdam: Elsevier Science Publishers).

ABOWD J. M., KRAMARZ F. (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), p. 2629-710.

BAYARD K. (2001). – « Measurement Error and Inter-Industry Wage Differentials. » Mimeograph, Board of Governors of the Federal Reserve System.

BAYARD K., HELLERSTEIN J., NEUMARK D., TROSKE K. (2000). – « The New Worker-Establishment Characteristics Database. » *Proceedings of the Second International Conference on Establishment Surveys* (American Statistical Association), p. 981-90.

BAYARD K., HELLERSTEIN J., NEUMARK D., TROSKE K. (2003). – « New Evidence on Sex Segregation and Sex Differences in Wages from Matched Employer-Employee Data. » *Journal of Labor Economics*, Vol. 21, No. 4, October, p. 887-922.

BAYARD K., HELLERSTEIN J., NEUMARK D., TROSKE K. (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.), p. 175-203.

BECKER G. S. (1971). – *The Economics of Discrimination*, Second Edition (Chicago: University of Chicago Press).

BERGMANN B. (1974). – « Occupational Segregation, Wages and Profits when Employers Discriminate by Race or Sex. » *Eastern Economic Journal*, Vol. 1, Nos. 1-2, p. 103-10.

BORJAS G. J. (1994). – « The Economics of Immigration. » *Journal of Economic Literature*, Vol. 32, p. 1667-717.

CARRINGTON W. J., TROSKE K. R. (1997). – « On Measuring Segregation in Samples with Small Units », *Journal of Business Economics and Statistics*, Vol 15, October, p. 402-9.

COCKBURN I., GRILICHES Z. (1987). – « Industry Effects and Appropriability Measures in the Stock Market's Valuation of R&D and Patents. » NBER Working Paper No. 2465.

DUSTMANN C., VAN SOEST A. – « Language and the Earnings of Immigrants. » *Industrial and Labor Relations Review*, Vol. 55, No. 3, April, p. 473-92.

FOSTER L., HALTIWANGER J., KRIZAN C.J. (1998). – « Aggregate Productivity Growth: Lessons from Microeconomic Evidence. » NBER Working Paper No. 6803.

GROSHEN E. L. (1991). – « The Structure of the Female/Male Wage Differential: Is it Who You Are, What You Do, or Where You Work? » *Journal of Human Resources*, Vol. 26, No. 3, Summer, p. 457-72.

HELLERSTEIN J., NEUMARK D. (1999). – « Sex, Wages, and Productivity: An Empirical Analysis of Israeli Firm-Level Data. » *International Economic Review*, Vol. 40, No. 1, February, p. 95-123.

HELLERSTEIN J., NEUMARK D. (1998). – « Wage Discrimination, Segregation, and Sex Differences in Wages and Productivity Within and Between Plants. » *Industrial Relations*, Vol. 37, No. 2, April, p. 232-60.

HELLERSTEIN J. K., NEUMARK D., TROSKE K. R. (1999). – « Wages, Productivity, and Worker Characteristics: Evidence from Plant-Level Production Functions and Wage Equations. » *Journal of Labor Economics*, Vol. 17, No. 3, July, p. 409-46.

HELLERSTEIN J. K., NEUMARK D., TROSKE K. R. (2002). – « Market Forces and Sex Discrimination. » *Journal of Human Resources*, Vol. 37, No. 2, Spring, p. 353-80.

KAIN J. (1968). – « Housing Segregation, Negro Employment, and Metropolitan Decentralization. » *Quarterly Journal of Economics*, Vol. 82, No. 2, May, 1968, p. 175-97.

LANG K. (1986). – « A Language Theory of Discrimination. » *Quarterly Journal of Economics*, Vol. 101, No. 2, May, p. 363-82.

LAZEAR E. P. (1999). – « Culture and Language. » *Journal of Political Economy*, Vol. 107, No. 6, Part 2, December, p. S95-126.

LENGERMANN P. A. (2001). – « Is it Who You Are, Where You Work, or With Whom You Work? Reassessing the Relationship Between Skill Segregation and Wage Inequality. » Mimeograph, University of Maryland.

LUBOTSKY D. (2000). – « Chutes or Ladders? A Longitudinal Analysis of Immigrant Earnings. » Mimeograph, Princeton University.

MAIRESSE J., GREENAN N. (1999). – « Using Employee Level Data in a Firm Level Econometric Study. » 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.), p. 489-514.

MATCHWARE TECHNOLOGIES, INC. (1997). – AutoMatch 4.2 User Manual (Burtonsville, MD).

ROSEN S. (1986). – « The Theory of Equalizing Differences. » In Orley C. Ashenfelter and Richard Layard, eds., *Handbook of Labor Economics*, Vol. 1 (Amsterdam: Elsevier Science Publishers), p. 641-92.

TREJO S. (1997). – « Why Do Mexican Americans Earn Low Wages? » *Journal of Political Economy*, Vol. 105, No. 6, December, p. 1235-68.

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), p. 371-404.

WILLIS R.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), p. 525-602.

APPENDIX A

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

TABLE A2

Hypothetical Matched Observations and Hand-Check Scores

	Worker-supplied information:	SSEL information:
Score1=1 & Score2=1:	Tiles 'R' Us 2440 Main St. Shelbyville, SW 11111 Industry=703	Tiles 'R' Us, Inc. 2440 S Main Shelbyville, SW 11111 Industry=703
Score1=2 & Score2=2:	Tiles 'R' Us 2240 E Main St. Gotham, SW 11111 Industry=703	Tiles 'R' Us, Inc. 2440 S Main Gotham, SW 11111 Industry=703
	or	
	Tiles 'R' Us Shopping Plaza Shelbyville, SW 11111 Industry=703	Tiles 'R' Us, Inc PO Box 222 Shelbyville, SW 11111 Industry=703
Score1=3 & Score2=3:	Grocery Store Chain Name Grocery Store Chain Name Shelbyville, SW 11111 Industry=601	Grocery Store Chain Name 2440 S Main Shelbyville, SW 11111 Industry=601
	or	
	Tiles 'R' Us 2440 S Main St Gotham, SW 11111 Industry=703	Tiles 'R' Us, Inc 2400 US Highway 10 Gotham, SW 11110 Industry=703
Score1=4 & Score2=4:	Shelbyville Hose Main Shelbyville, SW 11001 Industry=121	Shelbyville Manufacturing 2440 S Main Shelbyville, SW 11111 Industry=200
	or	
	Bank of Gotham 2440 Main St Gotham, SW 11111 Industry=700	Bank of Gotham 300 Fenwick R Gotham, SW 11111 Industry=700
Score1=5 & Score2=5:	Gotham Shop & Save 2440 Main St Shelbyville, SW 11111 Industry=603	Gotham Engine Repair Co. 2400 Peaceful St Shelbyville, SW 11111 Industry=751
	or	
	Reliable Car Repair 200 Main St Shelbyville, SW 11111 Industry=751	Reliable Dry Cleaners 2440 Main St Shelbyville, SW 11111 Industry=771

	Worker-supplied information:	SSEL information:
Score1=1 & Score2=2:	Shelbyville Hospital Main Shelbyville, SW 11101 Industry=831	Shelbyville Hospital 2440 Main St Shelbyville, SW 11111 Industry=831
Score1=1 & Score2=3:	Shelbyville Gas Works 2440 Main St. Shelbyville, SW 11111 Industry=201	Shelbyville Gas Works 2440 Main St. Shelbyville, SW 11111 Industry=641 (Eating Place industry code)
	or	
	Chuck & Dave's Bait Highway 10 Shelbyville, SW 11111 Industry=601	Chuck & Dave's 2440 Highway 10 Shelbyville, SW 11111 Industry=601
Score1=2 & Score2=5:	A1 Manufacturing 2440 Main St. Gotham, SW 11111 Industry=201	A1 Manufacturing Credit Union 2440 Main St. Gotham, SW 11111 Industry=702
Score1=2 & Score2=3:	Gotham Bank Gotham Bank Gotham, SW Industry=700	Gotham Bank 2440 Main St. Gotham, SW 11111 Industry=700

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