NBER WORKING PAPER SERIES

WORKPLACE SEGREGATION IN THE UNITED STATES: RACE, ETHNICITY, AND SKILL

Judith Hellerstein
David Neumark

Working Paper 11599
http://www.nber.org/papers/w11599

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2005

Neumark is also a Research Fellow at IZA. This is a substantially revised version of an earlier version of this paper by the same name. This research was funded by the Russell Sage Foundation and NIH. We are grateful to Megan Brooks, Joel Elvery, Gigi Foster, and especially Melissa McInerney for outstanding research assistance, to Stephen Raphael and Seth Sanders for helpful discussions, and to seminar participants at the Public Policy Institute of California, the BLS, the Census Bureau, the Federal Reserve Board, the NBER Summer Institute, ITAM, The University of California-Berkeley, and the Color Lines Conference at Harvard University. This paper reports the results of research and analysis undertaken while the authors were research affiliates at the Center for Economic Studies at the U.S. Census Bureau. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. It has been screened to ensure that no confidential information is revealed. Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the Census Bureau, the Public Policy Institute of California, or the Russell Sage Foundation. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Workplace Segregation in the United States: Race, Ethnicity, and Skill
Judith Hellerstein and David Neumark
NBER Working Paper No. 11599
August 2005
JEL No.

ABSTRACT

We study workplace segregation in the United States using a unique matched employer-employee data set that we have created. We present measures of workplace segregation by education and language–as skilled workers may be more complementary with other skilled workers than with unskilled workers–and by race and ethnicity, using simulation methods to measure segregation beyond what would occur randomly as workers are distributed across establishments. We also assess the role of education- and language-related skill differentials in generating workplace segregation by race and ethnicity, as skill is often correlated with race and ethnicity. Finally, we attempt to distinguish between segregation by skill based on general crowding of unskilled poor English speakers into a narrow set of jobs, and segregation based on common language for reasons such as complementarity among workers speaking the same language.

Our results indicate that there is considerable segregation by education and language in the workplace. Racial segregation in the workplace is of the same order of magnitude as education segregation, and segregation between Hispanics and whites is larger yet. Only a tiny portion of racial segregation in the workplace is driven by education differences between blacks and whites, but a substantial fraction of ethnic segregation in the workplace can be attributed to differences in language proficiency.

Judith Hellerstein
Department of Economics
University of Maryland
College Park, MD 20742
and NBER
hellerst@econ.umd.edu

David Neumark
Public Policy Institute of California
500 Washington Street, Suite 800
San Francisco, CA 94111
and NBER
neumark@ppic.org

I. Introduction

Wage differentials by education, race, and ethnicity in the United States have been extensively documented. When it comes to wage differentials by education, the past two decades have generally been marked by increased returns to education, the extent and sources of which have been the subject of much discussion (see, e.g., Katz and Murphy, 1992; Juhn, et al., 1992; Card and DiNardo, 2002; Autor, et al., 2004). As for wage differences by race and ethnicity (as documented in, e.g., Donohue and Heckman, 1991; Cain, 1986; Altonji and Blank, 1999; Welch, 1990; and Ihlanfeldt and Sjoquist, 1990), there has been extensive research trying to uncover their sources. Most researchers agree that skill differences such as education (including its quality) and language account for sizable shares of wage gaps by race and ethnicity (e.g., O'Neill, 1990; Trejo, 1997), with the sharper dispute whether gaps in these and other skills (such as those captured in test scores) fully explain these wage gaps or whether discrimination contributes as well (e.g., Darity and Mason, 1998; Neal and Johnson, 1996).

In contrast to this vast literature on wage differences, much less is known about the extent and sources of segregation in the labor market. There has been speculation that one source of increased wage inequality by education is increased segregation by skill (e.g., Kremer and Maskin, 1996), but there is little evidence of the extent of labor market segregation by education in the first place with which to test this hypothesis. Moreover, while there is widespread agreement that there is labor market segregation by race and ethnicity, and that this segregation accounts–at least in a statistical sense–for a sizable share of wage gaps between white males and other demographic groups (e.g., Carrington and Troske, 1998a; Bayard, et al., 1999; King, 1992; Watts, 1995; Higgs, 1977), there has been very little work trying to uncover whether this segregation is due to discrimination or other sources.[1]

[1] This segregation may occur along industry and occupation lines, as well as at the more detailed level of the establishment or job cell (occupations within establishments). For example, Bayard, et al. (1999) found that, for men, job cell segregation by race accounts for about half of the black-white wage gap and a larger share of the Hispanic-white wage gap. Carrington and Troske (1998a, 1998b) use data sets much more limited in scope than the one we use here to examine workplace segregation by race and sex. In general, the paucity of research on workplace segregation is presumably a function of the lack of data linking workers to establishments.

Workplace segregation by education, or by skill more generally, and workplace segregation by race and ethnicity have the potential to be intimately connected. There are numerous models suggesting that employers may segregate workers across workplaces by skill, most likely because of complementarities among workers with more similar skills. Because in U.S. labor markets skill is often correlated with race and ethnicity, an unintended effect of profit-maximizing skill segregation in the workplace may be segregation along racial and ethnic lines.[2] Alternatively, race and ethnic segregation in the workplace may be a function of discrimination in the labor market. Perhaps the most convincing evidence of discrimination in employment comes from audit studies of hiring (Cross, et al., 1990; Turner, et al., 1991), although this work does not speak to segregation per se.[3]

This paper has two goals: to use a new matched employer-employee data set to provide the best available measurements of workplace segregation by education, language, race, and ethnicity in the United States; and to present evidence that helps in understanding the sources of this segregation, in particular the role of skill in generating race and ethnic segregation. We pursue these goals using the 1990 Decennial Employer-Employee Database (DEED), a unique data set that we have created. The 1990 DEED is based on matching records in the 1990 Decennial Census of Population to a Census Bureau list of most business establishments in the United States.[4] The matching yields data on multiple workers matched to establishments, providing the means to measure workplace segregation in the United States based on a large, fairly representative data set.[5] In addition, the reliance on the Decennial Census

[2] On the supply side, labor market networks can also generate workplace segregation; we do not focus on labor market networks in this paper.

[3] Heckman (1998) notes that even if there is hiring discrimination–as audit studies suggest–whether or not a wage differential arises depends on the discriminatory behavior of the marginal rather than the average employer. Black (1995) shows that in a search model discriminatory tastes on the part of some employers can result in a wage gap, even when the discriminatory employers do not hire any minorities.

[4] The 1990 Census of Population is currently the only Decennial Census for which this match has been done.

[5] For example, Carrington and Troske (1998a) study workplace segregation using the Worker-Establishment Characteristics Database (WECD), which includes only manufacturing plants, and the Characteristics of Business Owners, which is restricted to small establishments. Bayard, et al. (1999) use the New Worker-Establishment Characteristics Database, which extends beyond manufacturing, but because of the method of matching used is nonetheless heavily biased toward manufacturing.

of Population as the source of information on workers creates the capacity to link information on workplace segregation to information on other characteristics of workers. This allows us to examine the extent of segregation in the workplace by skill, and to examine the impact of skill segregation in generating segregation by race and ethnicity. Thus, the DEED provides unparalleled opportunities to study workplace segregation by race, ethnicity, and skill.

Our empirical analysis proceeds in three steps that exploit these various characteristics of the DEED. First, we present measures of workplace segregation in the United States, focusing on segregation along the lines of education, language, race, and ethnicity.[6] Rather than considering all deviations from proportional representation across establishments as an "outcome" or "behavior" to be explained, we scale our measured segregation to reflect segregation above and beyond that which would occur by chance if workers were distributed randomly across establishments, using Monte Carlo simulations to generate measures of randomly occurring segregation.[7]

Simple calculations of workplace segregation across establishments are important in their own right, aside from the questions we consider concerning the sources of workplace segregation. Most research on segregation by race and ethnicity focuses on residential segregation (e.g., Massey and Denton, 1987; and Cutler, et al., 1999). But the boundaries used in studying residential segregation may not capture social interactions, and are to some extent explicitly drawn to accentuate segregation among different groups; for example, Census tract boundaries are often generated in order to ensure that the tracts are "as homogeneous as possible with respect to population characteristics, economic status, and living conditions."[8] In contrast, workplaces–specifically establishments–are units of observation that are

[6] In studying segregation by ethnicity, we focus exclusively on Hispanic ethnicity. We leave the measurement of workplace segregation by sex to other work, partly due to space constraints.

[7] This distinction between comparing measured segregation to a no-segregation ideal versus segregation that is generated by randomness is discussed in other work (see, e.g., Cortese, et al., 1976; Winship, 1977; Boisso, et al., 1994; and Carrington and Troske, 1997).

[8] U.S. Census Bureau, www.census.gov/geo/www/GARM/Ch10GARM.pdf (viewed April 27, 2005). Echenique and Fryer (2005) develop a segregation index that relies much less heavily on ad-hoc definitions of geographical boundaries.

generated by economic forces and in which people clearly do interact in a variety of ways, including work, social activity, labor market networks, etc.[9] Thus, while it is more difficult to study workplace segregation because of data constraints, measuring workplace segregation may be more useful than measuring residential segregation, as traditionally defined, for describing the interactions that arise in society between different groups in the population.[10] Of course similar arguments to those about workplaces could be made about other settings, such as schools, religious institutions, etc. (e.g., James and Taeuber, 1985).

The second step in our analysis probes the relationship between skill segregation on the one hand and racial and ethnic segregation on the other. Numerous models suggest that employers find it useful to group workers of similar skills together. For example, Kremer and Maskin (1996) develop a model in which employers have incentives to segregate workers by skill when workers of different skill levels are not perfect substitutes and different tasks within firms are differentially sensitive to skill.[11] Saint-Paul (2001) generates skill segregation across firms by assuming that there are productivity-related spillovers among workers within an establishment.[12] Cabrales and Calvó-Armengol (2002) show that when

[9] For a discussion of the importance of the workplace as a venue for social interaction between groups see Estlund (2003).

[10] Moreover, industry code, the closest proxy in public-use data to an establishment identifier, is a very crude measure to use to examine segregation. For example, we calculate that racial and ethnic segregation at the three-digit industry level in the DEED is typically on the order of one third as large as the establishment-level segregation we document below.

[11] For example, let the production function be $f(L_1, L_2) = L_1^c L_2^d$, with $d > c$. Assume that there are two types of workers: unskilled workers (L_1) with labor input equal to one efficiency unit, and skilled workers (L_2) with efficiency units of $q > 1$. Kremer and Maskin show that for low q, it is optimal for unskilled and skilled workers to work together, but above a certain threshold of q (that is, a certain amount of skill inequality), the equilibrium will reverse, and workers will be sorted across firms according to skill. Hirsch and Macpherson (1999) do not posit a formal model of sorting by skill, but assume that employers tend to hire workers of similar skills, and use this assumption–coupled with an assumption that blacks are on average less skilled than whites in terms of both observed and unobserved (to the researcher) skills–to suggest that the wage penalty associated with working in establishments with a large minority share in the workforce in part reflects lower unobserved skills of workers in such establishments.

[12] For example, positive spillovers may be reflected in each worker's productivity being the product of his productivity and an increasing function of the establishment's average skill level. Negative spillovers may arise because of fixed factors of production. All that is required for segregation in Saint-Paul's model is that over some range of average skill levels of an establishment's workforce there are increasing

workers' utility depends on interpersonal comparisons with nearby workers (such as those in the same firm), segregation by skill results.[13] And, of course, there are potential benefits to employers from grouping together workers who speak the same language.

Because race and ethnicity are correlated with skill (for example, blacks have less education than whites and Hispanics have lower English proficiency), racial and ethnic segregation may not reflect discrimination, but may be generated by segregation along skill lines. We begin by calculating the extent of segregation in the workplace by education. We calculate education segregation measures focusing only on whites, assuming implicitly that segregation by education for whites is generated by employers solely for reasons of economic efficiency. We then measure the extent of segregation between blacks and whites, and calculate how much of this segregation can be explained by differences in educational attainment between blacks and whites. We contrast these results with the extent to which wage differences between blacks and whites in our sample can be explained by education.

We repeat the analysis for the extent of segregation between Hispanics and whites. In considering the impact of skill in generating workplace segregation by Hispanic ethnicity, we measure the extent to which segregation by English language ability can explain Hispanic-white workplace segregation, treating language ability as another important dimension of skill.[14] We also compare these results to those from wage regressions where we measure how much of the Hispanic-white wage gap is driven by English language ability.

Finally, language is associated not only with skill, but also with country of origin, immigrant status, and assimilation. Consequently, if discriminatory forces lead to the segregation of blacks or

returns to skill.

[13] These authors also discuss evidence consistent with sorting by skill across employers, including Brown and Medoff (1989) and Davis, et al. (1991).

[14] We first documented segregation by language ability and explored its consequences for wages in Hellerstein and Neumark (2003). Because language may reflect things other than skill, there may be additional influences on hiring by language, including customer discrimination or the need for workers to speak the same language as customers, which, coupled with residential patterns, lead to this form of workplace segregation.

Hispanics from whites, they can also operate to segregate workers with poor English skills (immigrants, most likely) from other workers, in which case segregation by language would not reflect skill complementarities. We probe this question by exploring segregation among those whose English proficiency is poor, but whose native (and spoken) languages differ.

Our analysis focuses on larger establishments–the first quartile of establishment size for workers distribution in our analysis is approximately 40 workers. By comparison, the first quartile of the employment-weighted size distribution of all establishments in the SSEL is 20. The focus on larger establishments arises for two reasons. First, there are important methodological advantages to examining segregation in establishments where we observe at least two workers, which occurs infrequently for small establishments. Second, we match respondents to the Census Long Form–who are a randomly chosen one-sixth of the population–to establishments, and there is always a greater likelihood that any given number of workers will be sampled from a large establishment than a small establishment. Although we acknowledge that it would be nice to be able to measure segregation in all establishments, this is not the data set with which to do that convincingly. On the other hand, most legislation aimed at combating discrimination is directed at larger establishments; EEOC laws cover employers with 15 or more workers and affirmative action rules for federal contractors cover employers with 50 or more workers. Since policy has been directed at larger establishments, examining the extent of workplace segregation by race and ethnicity (in particular) in larger establishments is particularly salient.

Our results point to workplace segregation by education and race, and more so by ethnicity and language (at least for Hispanics). We find, however, that education plays very little role in generating workplace segregation by race. In contrast, segregation by language ability can explain approximately one third of overall Hispanic-white segregation. Finally, the evidence from poor English speakers points to segregation of Hispanics from others, suggesting that the role of language segregation among Hispanics is driven by complementarity in language skills.

II. Data

The analysis in this paper is based on the DEED, which we have created at the Center for Economic Studies at the U.S. Bureau of the Census. The DEED is formed by matching workers to establishments. The workers are drawn from the 1990 Sample Edited Detail File (SEDF), which contains all individual responses to the 1990 Decennial Census of Population one-in-six Long Form. The establishments are drawn from the Standard Statistical Establishment Listing (SSEL), an administrative database containing information for all business establishments operating in the United States in 1990. Here we provide a brief overview of the construction of the DEED; more details regarding the matching of the data are provided in Hellerstein and Neumark (2003).

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. The file containing this employer name and address information is referred to as the "Write-In" file, and had previously been used only for internal Census Bureau purposes. The Write-In file contains the information written on the questionnaires by Long-Form respondents, but not actually captured in the 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 to other Census Bureau establishment- or firm-level data sets that contain more detailed employer characteristics. 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. Because the name and address information on the Write-In file 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. Finally, 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. Thus, this procedure yields a very large data set with workers matched to their establishments, along with all of the information on workers from the SEDF.

Matching workers and establishments is a difficult task, because we would not expect employers' names and addresses to be recorded identically on the two files. 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). The value of parsing the addresses into multiple pieces is that we can match on various combinations of these components.

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

The second step of the matching process is to select and implement the matching specifications. The AutoMatch software uses a probabilistic matching algorithm 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 be in order to constitute 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 matches that are correct with a high probability are linked. Because the AutoMatch algorithm is not exact there is always a range of quality of matches, and we were therefore cautious in accepting 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, while always maintaining criteria that erred on the side of avoiding false matches. 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, obtaining most of our matches in the earliest passes. Finally, we engaged in a number of procedures to fine-tune the matching process, involving hand-checking of thousands of matches and subsequent revision of the matching procedures.

The final result is an extremely large data set of workers matched to their establishment of employment. The DEED consists of information on 3.3 million workers matched to nearly one million establishments, which accounts for 27 percent of workers in the SEDF and 19 percent of establishments in the SSEL.[15] In Table 1 we provide descriptive statistics for the matched workers from the DEED as compared to the SEDF. Column (1) reports summary statistics for the SEDF for the sample of workers who were eligible to be matched to their establishments. Column (2) reports summary statistics for the full DEED sample. The means of the demographic variables in the full DEED are quite close to the means in the SEDF across many dimensions. For example, female workers comprise 46 percent of the SEDF and 47 percent of the full DEED. Nonetheless, there are a few discrepancies. Perhaps most

[15] For both the DEED and SEDF we have excluded individuals as follows: with missing wages; who did not work in the year prior to the survey year (1989) or in the reference week for the Long Form of the Census; who did not report positive hourly wages; who did not work in one of the fifty states or the District of Columbia (even if the place of work was imputed); who were self-employed; who were not classified in a state of residence; or who were employed in an industry that was considered “out-of-scope” in the SSEL. (“Out-of-scope” industries do not fall under the purview of Census Bureau surveys. They 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.)

salient for this analysis is discrepancies in race and ethnicity. In the SEDF, white, Hispanic, and black workers account for 82, 7, and 8 percent of the total, respectively. The comparable figures for the full DEED are 86, 5, and 5 percent. While these differences are not huge, given that we are examining race and ethnic segregation, it is worth considering why they exist. In particular, there are many individuals who meet our sample inclusion criteria but for whom the quality of the business address information in the Write-In file is poor.[16]

In Appendix Table 1 we report a series of linear probability models where we examine the probability a worker who appears in the SEDF is successfully matched to an employer and appears in the DEED, as a function of observable characteristics. For this analysis we further limit the SEDF sample of column (1) of Table 1 to whites, blacks, and Hispanics. As shown in Appendix Table 1, column (1), blacks (Hispanics) are 11 (seven) percentage points less likely than whites to appear in the DEED. In column (2) we add a series of controls for whether an SEDF worker included business address information that appears in the Write-In file. Not surprisingly, a worker who included an employer name on the Write-In file is 23 percentage points more likely to be matched to an employer than a worker who did not. More important, including this set of controls reduces the coefficients on the black and Hispanic dummies substantially, so that conditional on including address information, blacks (Hispanics) are only six (five) percentage points less likely to appear in the DEED. In column (3) we include a full set of demographic characteristics as well, further reducing somewhat the estimated coefficient on the black and Hispanic dummy variables. In sum, these basic controls explain at least half of the racial and ethnic discrepancies in the probability that a worker is matched to the DEED. Many, if not all, of these controls likely are associated with attachment to the labor force and even with attachment to a specific employer. This leads to two conclusions. First, it is not a good idea to try to impute non-matched workers to

[16] For example, approximately four percent of workers in the SEDF do not provide any business address information at all.

employers in the SEDF,[17] or to re-weight the segregation measures we obtain to try to account for non-matched workers, given that non-matched workers differ substantially in observable and unobservable ways from matched workers. Second, one might therefore interpret the segregation results we obtain below as measuring of the extent of segregation among workers who have relatively high labor force attachment and high attachment to their employers. For measuring workplace segregation, this is a reasonable sample of workers to use, but another dimension along which it is not fully representative.

Returning to Table 1, column (3) reports summary statistics for the workers in the DEED who comprise the sample from which we calculate segregation measures and conduct inference. The sample size reduction between columns (2) and (3) arises for three reasons. First, we exclude workers who do not live and work in the same Metropolitan Statistical Area/Primary Metropolitan Statistical Area (MSA/PMSA). We use this U.S. Census Bureau measure of metropolitan areas because it is defined to some extent based on areas within which substantial commuting to work occurs.[18] Second, our analysis generally focuses on differences between whites and blacks and whites and Hispanics. We therefore exclude individuals who do not fall into those categories, with one exception. Because one of our analyses below compares Hispanics who do not speak English well to others who do not speak English well, we include in column (3) all workers, regardless of race and ethnicity, who self-reported speaking English “not well” or “not at all”. Third, we exclude workers who are the only workers matched to their establishments. The latter restriction effectively causes us to restrict the sample to workers in larger establishments, which is the main reason why some of the descriptive statistics are slightly different

[17] Even imputing place of work at the level of the census tract does not appear to be easy. For example, there are workers in the SEDF that we are able to match to an employer in the DEED using name and address information whose place of work code actually is allocated in the SEDF. For these workers, the allocated census tract in the SEDF disagrees with the SSEL census tract of the matched establishment in more than half the cases.

[18] See U.S. Census Bureau, http://www.census.gov/geo/lv4help/cengeoglos.html (viewed April 18, 2005). This is not to say that residential segregation at a level below that of MSAs and PMSAs may not influence workplace segregation. However, an analysis of this question requires somewhat different methods. For example, in conducting the simulations it is not obvious how one should limit the set of establishments within a metropolitan area in which a worker could be employed.

between the second and third columns. Finally, in columns (4) and (5) we report results for the subsample of workers who are used to construct two of our main segregation results, segregation by race and segregation by Hispanic ethnicity.

In addition to comparing worker-based means, it is useful to examine the similarities across establishments in the SSEL and the DEED. Table 2 shows descriptive statistics for establishments in each data set. As column (1) indicates, there are 5,237,592 establishments in the SSEL; of these, 972,436 (19 percent) also appear in the full DEED as reported in column (2). Because only one in six workers are sent Decennial Census Long Forms, as noted earlier, it is more likely that large establishments will be included in the DEED. One can see evidence of the bias toward larger employers by comparing the means across data sets for total employment. (No doubt this also influences the distribution of workers and establishments across industries.) On average, establishments in the SSEL have 18 employees, while the average in the DEED is 53 workers. The distributions of establishments across industries in the DEED relative to the SSEL are similar to those for workers in the worker sample. For example, manufacturing establishments are somewhat over-represented in the DEED, constituting 13 percent of establishments, relative to six percent in the SSEL. In column (3) we report descriptive statistics for establishments in the restricted DEED, corresponding to the sample of workers in column (3) of Table 1. In general, the summary statistics are quite similar between columns (2) and (3), with a small and unsurprising right shift in the size distribution of establishments. Overall, analyses reported in Hellerstein and Neumark (2003) indicate that the DEED sample is far more representative than previous detailed matched data sets for the United States.

III. Methods

We focus our analysis on one measure of segregation which is based on measures of the percentages of workers in an individual's establishment, or workplace, in different demographic groups. For a dichotomous classification of workers (e.g., whites and Hispanics), we define two segregation variables: the average percentage of Hispanic workers with which Hispanic workers work, denoted H_H;

and the average percentage of Hispanic workers with which white workers work, denoted W_H. The difference between these,

$$CW = H_H - W_H$$

is our measure of observed "co-worker segregation," and measures the extent to which Hispanics are more likely than are whites to work with other Hispanics. For example, if Hispanics and whites are perfectly segregated, then H_H equals 100, W_H is zero, and CW equals 100.[19]

To be precise, we exclude an individual's own ethnicity in calculating H_H and W_H. In the sociology literature on segregation, the percentage of Hispanics in a Hispanic's workplace is often called the "isolation index" and the percentage of whites in a Hispanic's workplace is the "exposure index." In that literature, the isolation and exposure indexes always include the own worker's ethnicity. However, when the own worker's ethnicity is included, the co-worker segregation measure is sensitive to the number of matched workers in the establishment. The problem is particularly severe in cases where there are only a few workers matched to each establishment, which happens frequently in the DEED. To see this, consider, as a simple example, the index of isolation, which in this case would be the average fraction of Hispanics among the workers in a Hispanic's establishment (including the worker himself). For a Hispanic worker in an establishment with two workers observed, the index is either 1 or 0.5, and most importantly never less than 0.5. But for a Hispanic worker in an establishment with 200 workers observed, the index can range from 0.005 to 1. In contrast, if we exclude the worker, the index can range from 0 to 1 for either type of establishment. If workers are randomly allocated across establishments, in a large sample we should expect the index of isolation to equal the share Hispanic. But when the worker is himself included this cannot happen in this example because the index is constrained to be between 0.5 and 1. In contrast, when the own worker's ethnicity is excluded from the segregation measure, as we do, this problem does not arise.

[19] We could equivalently define the percentages of white workers with which Hispanic or white workers work, H_W and W_W, which would simply be 100 minus these percentages, and $CW' = W_W - H_W$.

One limitation this poses is that the co-worker index can only be computed for establishments where we observe at least two workers. Coupled with the matching of Long-Form respondents, this contributes to yielding a sample in which small establishments are under-represented. Nonetheless, as we argued above, larger establishments may be a particularly important subset to examine given that policies to combat discrimination in the workplace are aimed at larger establishments.

There are, of course, other possible segregation measures, such as the traditional Duncan index (Duncan and Duncan, 1955) or the Gini coefficient. We prefer the co-worker segregation measure (CW) to these other measures for two reasons. First, the Duncan and Gini measures are insensitive to the proportions of each group in the workforce. For example, if the number of Hispanics doubles, but they are allocated to establishments in the same proportion as the original distribution, the Duncan index is unchanged but CW will rise, and in our view this doubling of the number of Hispanics implies more segregation. Second, these alternative segregation measures have the same basic problem outlined above with respect to sensitivity to the number of matched workers, and because they are measures that are calculated at only the establishment-level–unlike the co-worker segregation measure we use–there is no conceptual parallel to excluding the own worker from the calculation.[20]

We first report observed segregation, which is simply the sample mean of the segregation measure across workers. We denote this measure by appending an ‘O’ superscript to the segregation measures–i.e., CW^O. One important point that is often overlooked in research on segregation, however, is that some segregation occurs even with random assignment, and we are presumably most interested in the segregation that occurs systematically–i.e., that which is greater than would be expected to result from randomness. In the case of an infinitely large sample of workers, random allocation across establishments would imply that the co-worker segregation measure as we have defined it would be equal to zero, since, for example, H_H and W_H would be equal to the population Hispanic share.

[20] We believe this explains why, in Carrington and Troske (1998a, Table 3), the estimated Gini indexes are often extremely high due to small samples of workers within establishments.

There are two reasons why in our analysis the segregation measure with randomly assigned workers is not necessarily expected to be zero. The first is that some of our segregation measures are calculated conditional on geography and skill. So, for example, when we condition on geography, we calculate the extent of segregation that would be expected if workers were randomly allocated across establishments within a geographic area. If Hispanics and whites are not evenly distributed across geographic borders, random allocation of workers within geography will yield the result that Hispanics are more likely to have Hispanic co-workers than are white workers, because for example, more Hispanics will come from the areas where both whites and Hispanics work with a high share of Hispanic workers. Second, although the baseline sample size in our data is large, the actual samples that we use to calculate segregation below are not always large when we condition on geography and skill, or at least not necessarily large enough to approximate well this asymptotic result. For that reason, in order to determine how much segregation would occur randomly, we conduct a Monte Carlo simulation of the extent of segregation with random allocation of workers. We denote the mean measure of "random segregation" across the simulations CW^R.

Following Carrington and Troske (1997), to measure segregation beyond that which would occur randomly, we compute the difference between observed segregation and the mean level of random segregation, and scale the difference by the maximum segregation that can occur. We refer to this as "effective segregation." For $CW^O > CW^R$, the effective segregation measure is:

$$[\{CW^O - CW^R\}/\{100 - CW^R\}]\times 100 .$$

The denominator, $100 - CW^R$, is the maximum by which observed segregation can exceed random segregation, and so the scaling converts the difference $CW^O - CW^R$ into the share of this maximum possible segregation that is actually observed.[21]

For the Monte Carlo simulations that generate measures of random segregation, we first define the unit within which we are considering workers to be randomly allocated–which for most of the

[21] In principle, CW^O can be lower than CW^R. This never happens in our application.

analysis is metropolitan areas. We then calculate for each metropolitan area the numbers of workers in each category for which we are doing the simulation–for example, blacks and whites–as well as the number of establishments and the size distribution of establishments (in terms of sampled workers). Within a metropolitan area, we then randomly assign workers to establishments, ensuring that we generate the same size distribution of establishments within a metropolitan area as we have in the sample. We do this simulation 100 times, and compute the random segregation measures as the means over these 100 simulations. Not surprisingly, the random segregation measures are very precise; in all cases the standard deviations were trivially small.

With our measures of observed and random segregation, we then construct the effective segregation measures, which capture the level of effective segregation to which the average worker is subjected, conditional on the distribution of workers across metropolitan areas.

For descriptive purposes we also present some "unconditional" nationwide segregation measures where we do not first condition on metropolitan area, and where in the simulations we randomly assign workers to establishments anywhere in the country.[22] For comparability, when we construct these unconditional segregation measures we use only the workers included in the MSA/PMSA sample.[23] We emphasize the conditional measures much more in the paper. These can be interpreted as taking residential segregation by city as given. In contrast, because the "counterfactual" for the unconditional measure is that workers are randomly distributed across establishments nationwide, one could interpret the unconditional measures as estimating the degree of workplace segregation attributable to the distribution of workers both across cities as well as across establishments within cities.[24] As a result, for

[22] Not surprisingly, all the simulations we report where we randomly assign workers to establishments anywhere in the country lead to random segregation measures that are zero or virtually indistinguishable from zero.

[23] The results in this paper are robust to measuring segregation at the level of the MSA/CMSA metropolitan area, as well as measuring unconditional segregation by including all workers in the United States whether or not they live or work in a metropolitan area.

[24] Carrington and Troske (1998b) go further and characterize the interpretation of the national unconditional measure as "employment segregation causes residential segregation" (p. 243). We prefer to simply characterize it as capturing the joint effects of workplace segregation and the distribution of

the observed segregation measures the conditional and unconditional measures yield identical results; only the simulations differ.

Finally, in addition to constructing estimates of effective segregation in the workplace along various dimensions, we are interested in comparisons of measures of effective segregation across different samples. Given also that we are sometimes comparing estimates across samples that have some overlap,[25] we assess statistical significance of measures of effective segregation or differences between them using bootstrap methods. Briefly, the evidence indicates that our estimates are quite precise, and that the differences between the effective segregation indexes discussed in detail in the next section are generally strongly statistically significant.

IV. Basic Wage Regressions

Prior to presenting the results from the analysis of segregation, we first report the results of some basic wage regressions, illustrating in our DEED subsample the black-white wage gap, the Hispanic-white wage gap, and the importance of measured education and English language proficiency in explaining these gaps. These results parallel those commonly reported elsewhere, and serve to provide a benchmark for our later analysis of the link between workplace segregation, race, and ethnicity. As with most studies, we treat education and English language proficiency as measured without error and exogenous to wages (and similarly, for our segregation measures, place of employment).

First, in Table 3 we report results for the sample of black and white workers (corresponding to Table 1, column (4)). In columns (1) and (2) we report the educational distributions among whites and blacks. Only ten percent of whites in the sample have less than a high school degree, whereas 18 percent of blacks do. In contrast, at the top end of the education distribution, 25 percent of whites have at least a college degree but only 14 percent of blacks do. In column (3) we report that the coefficient on the black

workers across metropolitan areas by demographic group, without literally attributing this distribution across metropolitan areas to the forces that might generate workplace segregation.

[25] For example, we compare effective segregation between Hispanics who speak English poorly and Hispanics who speak English well, to effective segregation between Hispanics who speak English poorly and non-Hispanics who speak English poorly.

dummy in a log wage regression with only a control for race is −0.204. In column (4), we report results from a log wage regression where we include a dummy variable for black as well as dummy variables for educational attainment. The coefficients on the education dummies illustrate the usual monotonically increasing return to education. More important, the coefficient on the black dummy falls to −0.127, a reduction of 38 percent, indicating that education accounts for a large share of the black-white wage gap.[26]

In Table 4 we report results of a similar exercise where we examine the wage gap between Hispanic and white workers and the impact of English language proficiency on the Hispanic-white wage gap. In columns (1) and (2) we report the distributions of self-reported English language proficiency for whites and Hispanics, respectively. In the sample, almost 99 percent of (a very large sample of) whites report speaking English very well, whereas only 63 percent of Hispanic workers do. Many more Hispanics report speaking English not well or not at all. The raw Hispanic-white log wage gap, as reported in column (3) is −0.277. In column (4) we include controls for English language proficiency. The coefficients on the language dummies themselves show that the return to language proficiency is monotonic and increasing, and causes the coefficient on the Hispanic dummy to fall to −0.204, a 26 percent drop.[27] Similar results have been found in other work on the Hispanic-white wage gap (and in our previous work with the DEED, in Hellerstein and Neumark, 2003). Like for the black-white wage gap and education, skill therefore accounts for a sizable share of the Hispanic-white wage gap.[28]

[26] For a small fraction of the sample used in Tables 2 and 6 (less than one percent), hourly wages are less than one dollar per hour, rendering the log wage negative. Excluding workers whose measured hourly wage is less than $2 does not markedly affect either the wage regression results, or the measurement of black-white segregation reported later.

[27] The result is larger (a 42 percent drop) if we control for a quadratic in age and a sex dummy in the regression, but is very robust to trimming the sample to exclude workers who earn hourly wages computed to be below $2 per hour.

[28] Education also helps explain the Hispanic-white wage gap, and indeed explains more of the wage gap than language. However, because language differences are larger than education differences between Hispanics and whites, and because we find that much more of the Hispanic-white workplace segregation we document below can be explained by language differences than by education differences, we limit our focus to language.

With these results in mind we turn to the key contribution of the paper, measuring and explaining workplace segregation by skill, race, and ethnicity.

V. Segregation Results

Workplace Segregation by Education

The segregation analysis begins with measures of workplace segregation by education for whites. We focus first on whites so as not to confound our measures of segregation by education with segregation that is driven by other factors, such as race, which are correlated with education. Because it is easiest to characterize segregation with a binary measure of education, we define workers as low education if they have a high school degree or less, and high education if they have at least some college.[29] Table 5 reports results for education segregation, using the sample of establishments with two or more matched workers. To provide a sense of overall segregation by education for whites, column (1) provides the various segregation measures at the unconditional national level, looking at all urban areas (PMSAs and MSAs) as a whole. Column (2) presents the conditional national segregation indexes that are constructed by weighting up to the national level each individual PMSA/MSA segregation measure.

In column (1), looking first at the observed co-worker segregation by education for whites , we see extensive segregation. In particular, low educated white workers on average work in establishments in which 53.0 percent of matched white co-workers are also low education. In contrast, high education workers work in establishments with white co-workers who are only 33.1 percent low education on average. Below these figures we present the calculations from the simulations. Given that we randomize workers in this sample across the whole United States in conducting this simulation, it is not surprising that the results of the simulation imply that, on average, both low and high educated white workers work with co-workers who are 41.3 percent low education–the sample average. That is, for this particular exercise, the random co-worker segregation measure is zero, so that the effective co-worker simulation measure, 20.0, is simply the observed co-worker simulation measure (CW^O). One useful way to interpret

[29] We further disaggregate workers by education below when we consider how much of segregation by race is attributable to segregation by education.

this number is that 20 percent of the maximum amount of segregation that could arise due to non-random factors is actually observed in the data. While it is not clear to what one should compare this result, it suggests that there is substantial segregation by education.

Column (2) looks at segregation within urban areas defined as PMSAs/MSAs. As noted earlier, observed co-worker segregation is the same within and across urban areas; only the random segregation measure differs. The random segregation measure is 4.2 (no longer zero for reasons explained above, because workers are reallocated for the simulation only within urban areas); the pattern of random segregation has low education workers working, on average, with co-workers who are 43.7 percent low educated, while for high education workers the corresponding figure is 39.6 percent. As a result, the effective segregation measure in column (2) falls to 16.5. That is, about 17 percent of the maximum amount of segregation that could arise due to non-random factors is observed in the data.

Column (3) of Table 5 calculates segregation by education for blacks in the sample, conditional on the metropolitan area in which they live. There are more low education blacks in the sample than whites, but observed and random segregation (CW^O and CW^R) across the two columns are very similar, so that the effective segregation measure for education segregation for blacks is 15.0, similar to the 16.5 estimate for whites. This is suggestive evidence that the factors driving skill segregation, as defined here by education, are the same for whites as for blacks, as would be expected if skill segregation is arising due to profit-maximizing behavior.

Workplace Segregation by Race

Table 6 reports results for overall black-white segregation which can be compared to segregation by education (Table 5). In column (1) of Table 6, we report the extent of segregation by race (black versus white) in the whole United States where random segregation is defined by allowing workers to work anywhere. In column (2), random segregation by race is calculated by conditioning on the MSA/PMSA in which a worker lives. On average, black workers work with co-workers who are 23.7 percent black, while white workers work with co-workers who are 5.8 percent black. Based on the sample average of blacks in the population, random allocation across the United States would imply that blacks and whites should each work with co-workers who are 7.1 percent black, so that the overall level of effective segregation as reported in column (1) is 17.8. Because there is some racial segregation across urban areas, when we simulate random segregation within urban areas, in column (2), there is some segregation that arises randomly. In particular, random assignment would lead blacks to work in establishments with co-workers who are on average 11.2 percent black, versus an average percent black of 6.8 percent for whites. Based on the comparison between observed and random segregation, the effective segregation measure is 14.0, meaning that 14 percent of the maximum amount of racial segregation that could arise due to non-random factors is actually observed in the data.

A comparison of Tables 5 and 6 shows that the extent of segregation by race is very similar to that of segregation by education. Although the overall fraction of black workers is much lower than the fraction of low educated workers in the sample, the observed and random co-worker segregation measures are remarkably similar when comparing racial segregation to education segregation. As a result, the overall extent of racial segregation in the United States (14.0) is very similar to the extent of education segregation for whites (16.5) or blacks (15.0).

Workplace Segregation by Race, Conditional on Education

Next, we measure the extent to which racial segregation in the workplace can be explained by education differences between blacks and whites. We do this by constructing new "conditional" random segregation measures, where we simulate segregation holding the distribution of education fixed across all workplaces. So, for example, if an establishment in our sample is observed to have three workers with a high school degree, three workers with a high school degree will be randomly allocated to that establishment. We again compute the average (across the simulations) simulated fraction of co-workers who are black for blacks, denoting this $B_B{}^C$, and the average (across the simulations) simulated fraction of co-workers who are black for whites, denoting this $W_B{}^C$. The difference between these two is denoted CW^C, and we define the extent of "effective conditional segregation" to be:

$[\{CW^O - CW^C\}/\{100 - CW^R\}]\times 100$,

where CW^R is the measure of random segregation obtained when not conditioning on education. A conditional effective segregation measure of zero would imply that all of the effective segregation between blacks and whites can be attributed to education segregation that is coupled with differences in the education distribution between blacks and whites. Conversely, a conditional effective segregation measure equal to that of the (unconditional) effective segregation measure would imply that none of the effective segregation between blacks and whites can be attributed to education segregation across workplaces. We first do this calculation with the same two-way classification of education used in Table 5, and then expand to four educational levels; we also use an occupational classification with six groupings that we consider to be skill-related.

Column (1) of Table 7 reports the results for the two-way education classification. Observed segregation between blacks and whites is unaffected by this conditioning, of course, and so the top part of column (1) of Table 7, which reports the observed segregation between blacks and whites, repeats the results from Table 6. We report the conditional random segregation measures starting in the middle of the rows of Table 7. On average, random allocation of workers, conditional on randomization within the two education categories and within MSA/PMSA results in black workers working, on average, with co-workers who are 11.4 percent black, and white workers working, on average, with co-workers who are 6.8 percent black. These numbers are very close to the (unconditional) simulated numbers reported in Table 6, column (2). As a result, the conditional effective segregation measure is 13.9, very close to the unconditional segregation measure of 14.0. In other words, segregation by the binary education distinction (which we measure to be extensive) can explain only a tiny fraction (0.9 percent) of overall black-white segregation.

We repeat this analysis in column (2) of Table 7, this time conditioning on four education groupings when randomizing workers to workplaces: less than high school; high school degree; some college or associates degree; and bachelors degree or above. The results of the conditional random segregation are very similar to that obtained with two education groupings, so that our conditional effective segregation measure falls only to 13.6.

Education is, of course, only one dimension of skill across which employers may sort workers and which may be correlated with race. Another possible mechanism by which workers may be sorted is by occupation. Sorting by occupation may represent skill sorting, or it may be a proxy for a sorting mechanism in which employers engage for other reasons (such as alleviating employee discrimination). We explore the role of occupation sorting by computing random segregation conditional on the six one-digit occupation categories in column (3) of Table 7 (listed in the notes to the table). While this conditioning has slightly more effect than conditioning on education, the effective conditional segregation measure is still 12.9, accounting for only eight percent of overall black-white segregation.

As we reported in Table 3, it is not the case that education differences between blacks and whites are too small in this sample to have meaningful consequences for workplace segregation by race. There are large differences in education between blacks and whites, particularly at the upper and lower ends of the spectrum, and these differences can explain a large fraction of black-white wage differences. This implies that while education differences between blacks and whites go a reasonably long way toward explaining wage differences, they do not explain differences in where black and white workers in our sample work.

To show this explicitly, in Table 8 we report wage regressions where we compare the black-white wage gaps as estimated with and without including establishment fixed effects in the regressions. In columns (1) and (2) we repeat the results from Table 3, where we estimate the black-white wage gap without controlling for establishment fixed effects. Column (3) replicates the specification in column (1), but includes establishment fixed effects. The coefficient on the black dummy actually becomes more negative when we include a dummy variable for race and establishment fixed effects, implying that blacks work in slightly higher-wage establishments, rather than lower-wage establishments.[30] When we

[30] Including one-digit industry dummy variables in the regression leaves the coefficient on the black dummy almost unchanged from columns (1) and (2) and has very little effect on the coefficients on the

add the education controls to this specification, in column (4), the coefficient on the black dummy again falls by about one-third. The fact that the difference in the coefficient on the black dummy falls by as much with the fixed effects as without them indicates that the role of education in explaining the black-white wage gap does not arise through sorting of blacks and whites across establishments based on education. This is consistent with our evidence that education contributes minimally to black-white workplace segregation.

These results do not indicate that sorting across establishments has nothing to do with skill. On the contrary, the results from columns (2) and (4) of Table 8 show that including the establishment fixed effects in the regression reduces the estimated returns to education substantially, suggesting that there is, in fact, sorting by skill across establishments so that education differences of workers within a given establishment play a reduced role in explaining wage differences between workers. But what these results do suggest is that the sorting of workers by education across establishments (that we established in Table 5) is not related to the sorting of workers by race that leads to wage gaps between blacks and whites.

Given that education essentially plays no role in generating what we consider to be the rather substantial amount of racial segregation in the workplace, it is difficult to imagine that unobservable skill differences between blacks and whites could explain a sizable fraction of workplace segregation by race. The mechanism(s) behind workplace segregation by race therefore appear not to be skill related. Alternative mechanisms such as labor market discrimination, residential segregation/spatial mismatch within urban areas, or labor market networks are all possibilities worthy of future exploration.

Workplace Segregation by Ethnicity

We now turn to an examination of the extent and causes of workplace segregation by Hispanic ethnicity. The baseline estimates for the extent of Hispanic-white segregation are reported in columns (1) and (2) of Table 9, and the basic conclusion is that there is extensive workplace segregation by

education variables.

Hispanic ethnicity. The first specific thing to note is that the segregation figures for the unconditional national indexes indicate more segregation by ethnicity than their counterparts for race as reported in Table 6. Specifically, in column (1) of Table 9 the average percentage of Hispanics with whom Hispanics work is 39.4 percent, versus a comparable figure of 23.7 percent for blacks. The effective segregation measures are similarly different: 34.9 for Hispanic-white segregation versus 17.8 for black-white segregation.

The results are not as starkly different when we condition on metropolitan areas. This occurs because, for Hispanics, randomly-generated segregation is quite far from zero, conditional on metropolitan areas. In column (2) of Table 9, for example, the randomly allocated share Hispanic for Hispanic workers is 24.4 percent, compared with a parallel share Hispanic for white workers of 5.6 percent. This difference mainly arises because Hispanics are not as evenly dispersed across metropolitan areas as are blacks, some of which have few Hispanics. The net result is that, conditional on metropolitan area, the effective co-worker segregation measure is only somewhat higher for Hispanics (19.8) than for blacks (14.0).

In columns (3) and (4) of Table 9, we explore the extent of workplace segregation by English language proficiency for whites and Hispanics separately. As for education, employers may find it efficient to segregate workers by English language proficiency. Indeed, it is possible that the motives for segregation by language are even stronger than for segregation by education since workers who cannot communicate with each other impose clearly impose costs on employers relative to the alternative. We divide language proficiency into two categories. The first, "poor English," consists of workers who report speaking English not well or not at all. The second, "good English," consists of workers who report speaking English well or very well.

In column (3) we report the extent of workplace segregation by language for whites. Less than one half of one percent of the white sample are in the poor English category, yet a worker in this category works, on average, with co-workers for whom 6.9 percent speak English poorly. In contrast, for white workers in the good English category, only 0.4 percent of their co-workers speak English poorly. Random co-worker segregation for this sample, while not zero, is small (0.6). As a result, effective segregation for whites by language proficiency is 6.0. While the scale of this is smaller than for the other effective segregation measures computed thus far, we think it is notable given the very small percentage of poor English speakers among the whites.

The results on language segregation for Hispanics, in column (4), illustrate more starkly that there is extensive workplace segregation by language proficiency. Hispanics who speak English not well or not at all are likely to have Hispanic co-workers among whom, on average, 48.1 percent also speak English poorly or not at all. In stark contrast, Hispanics in the "good English" category are likely to have Hispanic co-workers of whom, on average, only 15.4 percent are in the "poor English" group. The random segregation measures indicate that some segregation arises randomly, conditional on geographic area. Under random allocation Hispanics in the "poor English" category would have 26.8 percent of Hispanic co-workers speaking English poorly or not at all, while workers in the "good English" category would have 21.7 percent of co-workers speaking English poorly or not at all. All together, this implies that the effective segregation measure for language segregation for Hispanics is 29.1, much larger than any other (within MSA/PMSA) segregation measure thus far.

In Table 10, we explore the extent to which the very pronounced language segregation for Hispanics may be driving Hispanic-white workplace segregation, since Hispanics have so much lower English language proficiency, on average, than whites. In the top panel of column (1) we repeat the figures for observed Hispanic-white segregation from Table 9, column (2); as reported earlier, the difference between co-worker segregation for Hispanics and whites is 34.9. We then report conditional random segregation for Hispanics and whites, conditional on the two language groupings used in the previous table (in addition to MSA/PMSA). With random allocation within the two language groups, Hispanics on average work with co-workers who are 26.8 percent Hispanic, whereas whites work with co-workers who are 5.5 percent Hispanic. That is, the simulated difference between the co-worker segregation measures is 21.3. Together these numbers lead to an effective segregation measure of 16.7. When we repeat this exercise in column (2), this time randomizing workers within the four language groups for which workers self-report English language proficiency (not at all, not well, well, very well), the effective segregation measure is 13.5. This figure can be interpreted as saying that of the Hispanic-white unconditional effective segregation measure of 19.8, nearly a third (32 percent = (19.8-13.5)/19.8) can be explained by language segregation.

Paralleling the analysis for black-white segregation, in column (3) we explore the extent to which Hispanic-white segregation can be explained by segregation across 1-digit occupation. The results indicate that segregation by 1-digit occupation is 16.6 and therefore explains about the same amount of Hispanic-white segregation as can segregation by language proficiency when defined as a dichotomous variable as in column (1). This is not surprising, given the large overlap in the distributions of occupation and English language proficiency among Hispanics. For example, among Hispanic managers, 97% report speaking English well or very well, as compared to only 66% for Hispanic laborers. Indeed, in unreported results, the effective segregation measure conditional on both 1-digit occupation and the two English language proficiency categories is 14.0, not much below that of conditioning only on English language proficiency.

The result that English language proficiency can explain a large fraction of Hispanic-white segregation is starkly different from the result we obtained for black-white workplace segregation, which could not be explained by the large differences in educational attainment between blacks and whites. It is useful, then, to examine whether the impact of language on ethnic workplace segregation manifests itself in wage gaps between Hispanics and whites. We do this in Table 11. Columns (1) and (2) report the basic wage regression results from Table 4, where we noted that controlling for English language proficiency caused a 26 percent drop in the Hispanic-white wage gap. Columns (3) and (4) report results including establishment fixed effects. Including fixed effects causes the "raw" (unconditional on language) Hispanic-white wage gap to fall from −0.277 to −0.255, indicating that Hispanics work in somewhat lower-paying establishments than whites. With fixed effects included, however, including English language proficiency only causes the Hispanic-white wage gap to fall to −0.221, as shown in column (4). This is only a 13 percent drop from column (3). Moreover, each of the coefficients on the dummy variables for English language proficiency falls with the inclusion of the fixed effects. This indicates that the role of English language proficiency in explaining wage gaps between Hispanics and whites is partially manifested in the role of language in sorting workers across establishments. This is consistent with our workplace segregation finding that language differences between Hispanics and whites can explain a large fraction of workplace segregation by Hispanic ethnicity.

In sum, skill differences between Hispanics and whites, at least as defined by language proficiency, explain approximately the same share of Hispanic-white workplace segregation as of the Hispanic-white wage gap, and are consistent with the role of sorting across establishments in explaining the Hispanic-white wage gap. This contrasts with the finding that skill differences between blacks and whites, as defined by education differentials, explain virtually none of racial segregation in the workplace.

Understanding Workplace Segregation by Language Proficiency

For Hispanic workers we have documented that substantial workplace segregation is generated by skill differences, at least as defined by language proficiency. One interpretation of this evidence is that employers have good reasons to pursue such segregation, and because language proficiency is correlated with ethnicity, segregation by language arising for non-discriminatory reasons generates segregation by ethnicity. Another possibility, though, is that language is associated with other dimensions along which employers discriminate–such as national origin or socioeconomic factors–and on the basis of which employers crowd workers into a subset of jobs (typically jobs that pay less). It can be difficult to distinguish between these competing hypotheses.[31] In the case of language skills, however, we believe some progress can be made on this question.

In particular, to test whether there are efficiency reasons for segregation by language skill, as opposed to simple segregation of those with poor English into a subset of jobs, we can consider employment patterns for workers who speak poor English but who also speak different languages. If Hispanic poor English speakers (who generally speak Spanish) are not segregated from non-Hispanic poor English speakers (who speak a language other than Spanish), then this would suggest that those with low skills are clustered in the same workplaces for reasons other than efficiency gains from grouping workers who speak the same language; such segregation would be more consistent with simple segregation of "less desirable" workers into a subset of jobs. In contrast, if Hispanic poor English speakers are segregated from those who have poor English skills but speak languages other than Spanish, then segregation by language skills may be arising for reasons of greater complementarity between workers who speak the same language (or a related economic incentive to segregate workplaces by common language). Alternatively, such segregation by language may be a function of residential segregation and/or hiring networks where workers who speak the same language have access to the same subset of employers. Network relationships can themselves be efficiency enhancing if they make it easier for workers to find jobs or for employers to find workers.

The results of this analysis are reported in Table 12. Column (1) repeats the calculations from Table 9 for segregation between Hispanic workers with poor English skills and Hispanic workers with good English skills. In contrast, column (2) reports calculations for segregation between Hispanics with poor English skills and non-Hispanics (including non-whites) with poor English skills. These figures indicate much more extensive segregation than in column (1): 49.5 versus 29.1. Note that in column (2)

[31] This is potentially true in many contexts, even though it is often ignored. For example, Bertrand and Mullainathan (2004) provide evidence from an audit study that employers are less likely to interview job candidates with "black-sounding" names. This may be because of race discrimination per se, or because of discrimination against workers whose names suggest a certain cultural and socioeconomic upbringing (or the intersection of the two), but the paper has been interpreted as providing evidence of discrimination on the basis of race. (See also Fryer and Levitt, 2003.)

random segregation is far from zero, much of this resulting from sorting across MSA/PMSAs. Thus, this evidence suggests that much of the segregation of Hispanics with poor English skills arises because of factors other than the general crowding of low-skilled workers with poor English skills into the same set of low-paying workplaces.

Differences in Workplace Segregation by Establishment Size

Finally, in Table 13 we report the effective segregation measure for various dimensions of segregation by establishment size, for approximately the four quartiles of the establishment size distribution in our sample. This is of interest for a few reasons. First, we might expect to find less segregation in larger establishments simply because employers may be able to achieve the goal of segregation–whether it is separating workers by race or ethnicity, taking advantage of skill complementarity, or something else–by segregating workers within establishments.[32] Second, as noted earlier, EEO and affirmative action target larger employers, which may tend to discourage segregation in large establishments.[33]

The estimates are consistent with these expectations. In the first two rows, Hispanic-white and black-white segregation effective segregation range from 24-27 in the smallest establishments to 9-12 in the largest establishments, and in the third row skill segregation among whites falls from 18.0 to 12.7. The differences are sharper still when we compare segregation between Hispanics and non-Hispanics who speak English poorly and Hispanics who speak English well and those who speak it poorly. Segregation of Hispanics by language ability follows a roughly similar pattern to the other forms of segregation documented in the preceding rows in the table. But segregation of Hispanics from non-

[32] As an anecdotal example, an article in the *New York Times* describes a Texas factory that nearly completely segregates its Hispanic and Vietnamese workers into two different departments in the factory (with the Hispanics working in the lower-paying department). This article also points to the role of language complementarities between workers and supervisors, as one of the company's defenses of this practice is that the supervisor of the higher-paying department speaks Vietnamese but not Spanish (Greenhouse, 2003).

[33] Other research has documented a pattern of lower hiring of blacks in small establishments, and has argued that this reflects weaker or non-existing anti-discrimination policies at those establishments (Chay, 1998; Holzer, 1998; Carrington, et al., 2000).

Hispanics when both groups have poor English skills is very high in the small establishments (77.8), and falls by nearly 50 percentage points in going from the smallest to the largest establishments. The very high segregation by language in small establishments, coupled with the sharp drop as we move to larger establishments, reinforces the idea that language complementarities contribute to workplace segregation by language among those who speak poor English. Nonetheless, if residential location is less important in determining employment at large establishments than small establishments, which would be the case if those working at large establishments tend to be drawn from a wider geographic area, these results may again be consistent with residential segregation between Hispanics and other groups with poor English skills driving the workplace segregation results.

V. Conclusions

We use a unique data set of employees matched to establishments to study workplace segregation in the United States. We document that there is rather extensive segregation by education for white workers, consistent with models where employers find it efficient to segregate workers by skill. Similarly, among Hispanics we document extensive segregation by language, which is perhaps even stronger evidence that skill complementarities in the workplace generate segregation. We also document that there is segregation by race in the workplace of the same magnitude as education segregation, and segregation by Hispanic ethnicity that is slightly larger.

After documenting these different dimensions of segregation, our analysis focuses on whether racial and ethnic workplace segregation reflects race or ethnicity per se, likely stemming from discrimination, or instead is attributable to skills that differ across race and ethnic groups and along which employers might find it useful to segregate workers. For racial segregation, we find that virtually none of it is attributable to skill differences, at least as these are manifested in education (or occupation) differences between blacks and whites. In contrast, we show that approximately one-third of ethnic segregation in the workplace is attributable to language proficiency. These results are reflected in wage regressions, where sorting across establishments does not decrease (and even increases) black-white wage gaps while it decreases the impact of education on wages, whereas sorting across establishments by Hispanic ethnicity decreases the ethnic wage gap and decreases the importance of language proficiency in explaining Hispanic-white wage gaps.

Finally, in order to further probe the role of skill in generating ethnic (and language) segregation, we ask whether segregation by skill likely arises due to the consignment of less-skilled workers to the same subset of workplaces, perhaps because of discrimination against workers on the basis of numerous characteristics associated with low skills–such as immigrant status–or whether other factors such as skill-based complementarities lead certain types of workers to work together. Providing evidence inconsistent with the first hypothesis, we find that Hispanics with poor English skills are considerably more segregated from workers with poor English skills who speak other languages than they are from Hispanics with good English skills. It therefore appears that the process by which Hispanic and white workers are sorted into workplaces is not simply one whereby low-skilled workers are relegated to the same set of (low-paying) workplaces, but rather is driven in part by sorting on language skills.

In addition to finding that there is extensive segregation by skill in the workplace, our results document the reality of racial and ethnic segregation in U.S. workplaces. For blacks, the fact that education differences between blacks and whites explain virtually none of racial workplace segregation means that further research must be conducted to uncover the sources of racial segregation in the workplace, and that this research necessarily must examine explanations that are not skill-based: discrimination, residential segregation, and labor market networks are the most obvious possibilities. While language proficiency can explain a large fraction of ethnic segregation in the workplace, these alternative explanations must also be considered with regard to the remaining ethnic segregation. Finally, understanding the mechanisms that lead segregation across workplaces to decrease with establishment size may help in understanding the sources of workplace segregation more generally, while for larger establishments it may be important to examine whether workers remain segregated within the workplace.

Appendix

From the point of view of drawing statistical inferences, we need to be able to assess the statistical significance of our effective segregation measures and of differences between them. Given the precision of the simulated segregation measures as discussed in Section III, the effective segregation measures are also likely relatively precise. To assess this more formally, we explore bootstrapped distributions for the effective segregation measures.

We use as our base sample the "Restricted DEED" as in Table 1, column (1). The data generating process for that sample can be approximated to a first order as a random sample of workers who are then matched to establishments, where workers have a constant probability of being matched to their establishment. For our bootstrap exercise we therefore draw a sample of workers with replacement of the original size of the Restricted DEED sample, maintaining for that worker the original sample estimate of the fraction of workers in the categories of interest (e.g. percent black, percent Hispanic). We then calculate the observed segregation measures in the entire paper for that bootstrap sample, making sample restrictions for each table in the paper as necessary from that bootstrap sample. We do not recalculate random segregation, but instead treat it as a population parameter from the Restricted DEED. Finally, we collect the information on the empirical distributions of the observed and effective segregation measures.

We do not report full results from the bootstrap replications. Observed segregation is measured very precisely in each case so that observed segregation is always statistically significantly different from random segregation. For example, consider Table 6, column (2). Observed co-worker segregation is 17.8 and random segregation is 4.4. From the bootstraps, we find that the standard error of the estimate of observed segregation is 0.08.

Finally, in order to assess whether the differences in estimated effective segregation between any two columns in the tables are statistically significant, we pair each of the 100 bootstraps across the two results, calculate the difference in the segregation measures across the samples for each bootstrap, and calculate the standard deviation of the difference in the segregation measures across columns. The differences in effective segregation across columns of the tables are virtually always highly significant.

References

Altonji, Joseph G., and Rebecca M. Blank. 1999. "Race and Gender in the Labor Market." In Handbook of Labor Economics, Vol. 3, eds. Ashenfelter and Card (Amsterdam: Elsevier), pp. 3143-259.

Autor, David H., Lawrence F. Katz, and Melissa S. Kearney. 2004. "Trends in U.S. Wage Inequality: Re-Assessing the Revisionists." Unpublished manuscript, MIT.

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

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, Vol. 94, No. 4, September, pp. 991-1013.

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

Boisso, Dale, Kathy Hayes, Joseph Hirschberg, and Jacques Silber. 1994. "Occupational Segregation in the Multidimensional Case." *Journal of Econometrics*, Vol. 61, No. 1, March, pp. 161-71.

Brown, Charles, and James Medoff. 1989. "The Employer Size Wage Effect." *Journal of Political Economy*, Vol. 97, No. 5, October, pp. 1027-59.

Cabrales, Antonio, and Antoni Calvó-Armengol. 2002. "Social Preferences and Skill Segregation." Unpublished paper, Universitat Pompeu Fabra.

Cain, Glen. 1986. "The Economic Analysis of Labor Market Discrimination: A Survey." In Handbook of Labor Economics, Vol. 1, eds. Ashenfelter and Layard (Amsterdam: North-Holland), pp. 693-785.

Card, David, and John E. DiNardo. 2002. "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles." *Journal of Labor Economics*, Vol. 20, No. 4, October, pp. 733-83.

Carrington, William J., and Kenneth R. Troske. 1997. "On Measuring Segregation in Samples with Small Units." *Journal of Business & Economic Statistics*, Vol. 15, No. 4, October, pp. 402-9.

Carrington, William H., and Kenneth R. Troske. 1998a. "Interfirm Racial Segregation and the Black/White Wage Gap." *Journal of Labor Economics*, Vol 16, No. 2, April, pp. 231-60

Carrington, William J. And Kenneth Troske. 1998b. "Sex Segregation in U.S. Manufacturing." *Industrial and Labor Relations Review*, Vol. 51, April, pp. 445-464.

Carrington, William J., Kristin McCue, and Brooks Pierce. 2000. "Using Establishment Size to Measure the Impact of Title VII and Affirmative Action." *Journal of Human Resources*, Vol. 35, No. 3, Summer, pp. 503-23.

Chay, Kenneth Y. 1998. "The Impact of Federal Civil Rights Policy on Black Economic Progress: Evidence from the Equal Employment Opportunity Act of 1972." *Industrial and Labor Relations Review*, Vol. 51, No. 4, January, pp. 608-32.

Cortese, Charles, F., R. Frank Falk, and Jack K. Cohen. 1976. "Further Considerations on the Methodological Analysis of Segregation Indices." *American Sociological Review*, Vol. 51, No. 4, August, pp. 630-7.

Cross, Harry, Genevieve Kenney, Jane Mell, and Wendy Zimmerman. 1990. Employer Hiring Practices: Differential Treatment of Hispanic and Anglo Job Seekers (Washington, DC: Urban Institute Press).

Cutler, David M., Edward L. Glaeser, and Jacob L. Vigdor. 1999. "The Rise and Decline of the American Ghetto." *Journal of Political Economy*, Vol 107, No. 3, June, pp. 455-506.

Darity, William A., Jr., and Patrick L. Mason. 1998. "Evidence on Discrimination in Employment: Codes of Color, Codes of Gender." *Journal of Economic Perspectives*, Vol. 12, No. 2, Spring, pp. 63-92.

Davis, Steve J., John Haltiwanger, Lawrence F. Katz, and Robert Topel. 1991. "Wage Dispersion Between and Within U.S. Manufacturing Plants, 1963-1986." *Brookings Papers on Economic Activity: Microeconomics*, Vol. 1, pp. 115-200.

Donohue, John J., and James Heckman. 1991. "Continuous Versus Episodic Change: The Impact of Civil Rights Policy on the Economic Status of Blacks." *Journal of Economic Literature*, Vol. 29, No. 4, December, pp. 1603-43.

Duncan, Otis D., and Beverly Duncan. 1955. "A Methodological Analysis of Segregation Indices." *American Sociological Review*, Vol. 20, No. 2, April, pp. 210-7.

Echenique, Frederico, and Roland Fryer. 2005. "On the Measurement of Segregation." NBER Working Paper No. 11258.

Estlund, Cynthia. 2003. Working Together: How Workplace Bonds Strengthen a Diverse Democracy (New York: Oxford University Press).

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

Fryer, Roland G., and Steven D. Levitt. 2003. "The Causes and Consequences of Distinctively Black Names." NBER Working Paper No. 9938.

Greenhouse, Steven. 2003. "At a Factory in Houston, Hispanics Fight to Work in Coveted Department." *New York Times*, February 9, p. 14.

Heckman, James J. 1998. "Detecting Discrimination." *Journal of Economic Perspectives*, Vol. 12, No. 2, Spring, pp. 101-16.

Hellerstein, Judith, and David Neumark. 2003. "Ethnicity, Language, and Workplace Segregation: Evidence from a New Matched Employer-Employee Data Set." *Annales d'Economie et de Statistique*, Vol. 71-72, July-December, pp. 19-78.

Higgs, Robert. 1977. "Firm-Specific Evidence on Racial Wage Differentials and Workforce Segregation." *American Economic Review*, Vol. 67, No. 2, March, pp. 236-45.

Hirsch, Barry T., and David A. Macpherson. 2003. "Wages, Sorting on Skill, and the Racial Composition of Jobs." IZA Discussion Paper No. 741.

Holzer, Harry J. 1998. "Why Do Small Establishments Hire Fewer Blacks than Large Ones?" *Journal of Human Resources*, Vol. 33, No. 4, Fall, pp. 896-914.

Ihlanfeldt, Keith, and David Sjoquist. 1990. "Job Accessibility and Racial Differences in Youth Employment Rates." *American Economic Review*, Vol. 80, No. 1, March, pp. 267-76.

James, Daniel R., and Karl E. Taeuber. 1985. "Measures of Segregation." In Sociological Methodology, ed. Brandon Tuma (San Francisco: Jossey-Bass), pp. 1-32.

Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce. 1993. "Wage Inequality and the Rise in Returns to Skill." *Journal of Political Economy*, Vol. 101, No. 3, June, pp. 410-42.

Katz, Lawrence F., and Kevin M. Murphy. 1992. "Changes in Relative Wages, 1963-1987: Supply and Demand Factors." *Quarterly Journal of Economics*, Vol. 107, No. 1, February, pp. 35-78.

King, Mary C. 1992. "Occupational Segregation by Race and Sex, 1940-1988." *Monthly Labor Review*, April, pp. 30-7.

Kremer, Michael, and Eric Maskin. 1996. "Wage Inequality and Segregation by Skill." National Bureau of Economic Research Working Paper No. 5718.

Massey, Douglas, and Nancy Denton. 1987. "Trends in the Residential Segregation of Blacks, Hispanics, and Asians: 1970-1980." *American Sociological Review*, Vol. 52, No. 6, December, pp. 802-25.

Neal, Derek A., and William R. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy*, Vol. 104, No. 5, October, pp. 869-95.

O'Neill, June. 1990. "The Role of Human Capital in Earnings Differences between Black and White Men." *Journal of Economic Perspectives*, Vol. 4, No. 4, Fall, pp. 25-45.

Rivera, Elaine. 2003. "Area Bosses Try to Bridge Language Gaps." *Washington Post*, May 6, p. B1.

Saint-Paul, Gilles. 2001. "On the Distribution of Income and Worker Assignment under Intrafirm Spillovers, with an Application to Ideas and Networks." *Journal of Political Economy*, Vol. 109, No. 1, February, pp. 1-37.

Turner, Margery Austin, Michael Fix, and Raymond J. Struyk. 1991. Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring (Washington, DC: Urban Institute Press).

U.S. Census Bureau. "Census Geographic Glossary." http://www.census.gov/geo/lv4help/cengeoglos.html (viewed July 3, 2003).

U.S. Census Bureau, "Census Tracts and Block Numbering Areas." http://www.census.gov/geo/www/GARM/Ch10GARM.pdf (viewed May 10, 2004).

Watts, Martin J. 1995. "Trends in Occupational Segregation by Race and Gender in the U.S.A., 1983-92: A Multidimensional Approach." *Review of Radical Political Economics*, Vol. 27, No. 4, Fall, pp. 1-36.

Welch, Finis. 1990. “The Employment of Black Men.” *Journal of Labor Economics*, Vol. 8, No. 2, April, pp. S26-S75.

Winship, Christopher. 1977. “A Revaluation of Indexes of Residential Segregation.” *Social Forces*, Vol. 55, No. 4, June, pp. 1058-66.

Table 1: Means of Worker Characteristics

			Restricted	Black/White	Hispanic/White
	SEDF	Full DEED	DEED	sample	Sample
	(1)	(2)	(3)	(4)	(5)
Age	37.08	37.51	37.56	37.74	37.60
	(12.78)	(12.23)	(12.16)	(12.17)	(12.19)
Female	0.46	0.47	0.470	0.480	0.470
Married	0.60	0.65	0.630	0.630	0.640
White	0.82	0.86	0.870	0.93	0.930
Hispanic	0.07	0.05	0.060	---	0.070
Black	0.08	0.05	0.070	0.070	---
Full-time	0.77	0.83	0.840	0.840	0.840
Number of kids (if female)	1.57	1.53	1.46	1.44	1.43
	(1.62)	(1.55)	(1.53)	(1.51)	(1.51)
High school diploma	0.34	0.33	0.310	0.310	0.310
Some college	0.30	0.32	0.330	0.340	0.330
B.A.	0.13	0.16	0.170	0.180	0.180
Advanced degree	0.05	0.05	0.060	0.060	0.060
Ln(hourly wage)	2.21	2.30	2.37	2.39	2.39
	(0.70)	(0.65)	(0.64)	(0.64)	(0.64)
Hourly wage	12.10	12.89	13.67	13.91	13.86
	(82.19)	(37.07)	(27.72)	(28.36)	(28.43)
Hours worked in 1989	39.51	40.42	40.56	40.57	40.62
	(11.44)	(10.37)	(10.10)	(10.10)	(10.13)
Weeks worked in 1989	46.67	48.21	48.51	48.64	48.60
	(11.05)	(9.34)	(8.99)	(8.82)	(8.86)
Earnings in 1989	22,575	25,581	27,500	28,112	28,034
	(26,760)	(29,475)	(31,023)	(31,613)	(31,730)
Industry:					
Mining	0.01	0.01	0.010	0.010	0.010
Construction	0.07	0.04	0.030	0.030	0.040
Manufacturing	0.25	0.34	0.350	0.340	0.350
Transportation	0.08	0.05	0.060	0.060	0.050
Wholesale	0.05	0.07	0.080	0.080	0.080
Retail	0.20	0.17	0.150	0.150	0.150
FIRE	0.08	0.08	0.080	0.090	0.090
Services	0.26	0.24	0.240	0.250	0.240
Observations	12,143,183	3,291,213	1,755,825	1,618,876	1,625,953

Standard deviations of continuous variables are reported in parentheses. Column (3) is restricted to workers with at least one other worker matched to their establishment, and who work in the same metropolitan area (MSA/PMSA) in which they reside.

Table 2: Means for Establishments

	SSEL	Full DEED	Restricted DEED
Total employment	17.57	52.68	106.44
	(253.75)	(577.39)	(1011.57)
Establishment size:			
1 - 25	0.88	0.65	0.38
26 - 50	0.06	0.15	0.22
51 - 100	0.03	0.10	0.19
101 +	0.03	0.10	0.22
Industry:			
Mining	0.00	0.01	0.00
Construction	0.09	0.07	0.05
Manufacturing	0.06	0.13	0.19
Transportation	0.04	0.05	0.05
Wholesale	0.08	0.11	0.12
Retail	0.25	0.24	0.22
FIRE	0.09	0.10	0.10
Services	0.28	0.26	0.23
In MSA	0.81	0.82	1.00
Census Region:			
North East	0.06	0.06	0.05
Mid Atlantic	0.16	0.15	0.16
East North Central	0.16	0.20	0.22
West North Central	0.07	0.08	0.07
South Atlantic	0.18	0.16	0.16
East South Central	0.05	0.05	0.04
West South Central	0.10	0.10	0.09
Mountain	0.06	0.05	0.05
Pacific	0.16	0.15	0.15
Payroll ($1000)	397	1,358	2,963
	(5,064)	(10,329)	(16,818)
Payroll/total employment	21.02	24.24	26.73
	(1,385.12)	(111.79)	(184.25)
Share of employees matched	–	0.17	0.14
Multi-unit establishment	0.23	0.42	0.53
Observations	5,237,592	972,436	307,496

Standard deviations of continuous variables are reported in parentheses. 55 establishments in the Full DEED sample do not have valid county data from the SSEL. For these 55, the workers reported place of work was used to determine MSA status.

Table 3: The Distribution of Education by Race and the Impact of Education on Black-White Wage Gaps

	Sample means		Regression results	
	Whites	Blacks		
	(1)	(2)	(3)	(4)
Black	0	1	-0.204 (0.002)	-0.127 (0.002)
Less than a high school degree	0.10	0.18		
High school degree	0.31	0.32		0.196 (0.002)
Some college or Associates degree	0.33	0.37		0.331 (0.002)
College degree or above	0.25	0.14		0.744 (0.002)
Number of observations	1,503,640	115,236	1,618,876	1,618,876

The dependent variable in the regressions reported in columns (3) and (4) is the log of the hourly wage. There is a constant in the regressions.

Table 4: The Distribution of English Language Proficiency by Ethnicity and the Impact of Language Proficiency on Hispanic-White Wage Gaps

	Sample means		Regression results	
	Whites	Hispanics		
	(1)	(2)	(3)	(4)
Hispanic	0	1	-0.277 (0.002)	-0.204 (0.002)
Speak English “not at all”	0.0002	0.05		
Speak English “not well”	0.0036	0.14		0.210 (0.009)
Speak English well	0.0072	0.184		0.396 (0.009)
Speak English very well	0.989	0.626		0.471 (0.009)
Number of observations	1,513,277	112,676	1,625,953	1,625,953

The dependent variable in the regressions reported in columns (3) and (4) is the log of the hourly wage. There is a constant in the regressions.

Table 5: Segregation by Education

	Segregation by education for whites:		Segregation by education for blacks:
	U.S., MSA/PMSA, sample	Within MSA/PMSA	Within MSA/PMSA
	%Low ed	%Low ed	%Low ed
	(1)	(2)	(3)
Co-worker segregation			
Observed segregation			
Low education workers (L_L^O)	53.0	53.0	58.9
High education workers (H_L^O)	33.1	33.1	41.0
Difference (CW^O)	19.9	19.9	17.8
Random segregation			
Low education workers (L_L^O)	41.3	43.7	51.6
High education workers (H_L^O)	41.3	39.6	48.3
Difference (CW^R)	0	4.2	3.3
Effective segregation, $[\{CW^O - CW^R\}/\{100 - CW^R\}]\times 100$	**20.0**	**16.5**	**15.0**
Number of workers	1,500,322	1,500,322	83,401
Number of establishments	273,084	273,084	19,062

Low education is defined as high school degree or less. High education is defined as more than high school. Calculations are for establishments with two or more matched workers, where, for example, for the sample of workers in the first two columns, the median number of workers matched to an establishment is 8, and the median share of the workforce matched is 7.7 percent. (The hypothetical maximum is 16.7 percent, given that only 1/6 of workers receive the Census long form.) All medians are reported as “fuzzy medians” to comply with confidentiality restrictions; but they are extremely close to actual medians.

Table 6: Black-White Segregation

	All workers	
	Black-white segregation in U.S.	Black-white segregation within MSA/PMSA
	%Black	%Black
	(1)	(2)
Co-worker segregation		
Observed segregation		
Black workers (B_B^O)	23.7	23.7
White workers (W_B^O)	5.8	5.8
Difference (CW^O)	17.8	17.8
Random segregation		
Black workers (B_B^R)	7.1	11.2
White workers (W_B^R)	7.1	6.8
Difference (CW^R)	0	4.4
Effective co-worker segregation	**17.8**	**14.0**
Number of workers	1,618,876	1,618,876
Number of establishments	285,988	285,988

See notes to Table 5.

Table 7: Black-White Segregation Conditional on Education or Occupation

	Black-white segregation conditional on 2 education groups	Black-white segregation conditional on 4 education groups	Black-white segregation conditional on 1-digit occupation (six categories)
	(1)	(2)	(3)
Co-worker segregation			
Observed segregation			
Black workers (B_B^O)	23.7	23.7	23.7
White workers (W_B^O)	5.8	5.8	5.8
Difference (CW^O)	17.8	17.8	17.8
Conditional random segregation			
Black workers (B_B^C)	11.4	11.6	12.2
White workers (W_B^C)	6.8	6.8	6.7
Difference (CW^C)	4.6	4.8	5.4
Effective conditional segregation, [{CW^O – CW^C}/ {100 – CW^R}]×100	**13.9**	**13.6**	**12.9**
Number of workers	1,618,876	1,618,876	1,618,876
Number of establishments	285,988	285,988	285,988

See notes to Table 5. In column (1) the education groups are: high school or less; more than high school. In column (2) the four education groups are: less than high school; high school degree; some college or associates degree; bachelors degree or higher. In column (3) the occupations are: managerial and professional specialty; technical, sales, and administrative support; service; farming, forestry, and fishery; precision production, craft, and repair; and operators, fabricators, and laborers.

Table 8: Black-White Wage Gaps without and with Establishment Fixed Effects

	Without establishment fixed effects		With establishment fixed effects	
	(1)	(2)	(3)	(4)
Black	-0.204 (0.002)	-0.127 (0.002)	-0.232 (0.002)	-0.164 (0.002)
High school degree		0.196 (0.002)		0.096 (0.002)
Some college or Associates degree		0.331 (0.002)		0.205 (0.002)
College degree or above		0.744 (0.002)		0.534 (0.002)
Number of observations	1,618,876	1,618,876	1,618,876	1,618,876

The dependent variable in the regressions is the log of the hourly wage. The category less than high school is omitted from the regressions in columns (2) and (4).

Table 9: Hispanic-White Segregation and Language Segregation by Ethnicity

	Establishment ethnic composition:			Establishment language composition:	
	Hispanic-white segregation in U.S. (MSA/PMSA sample)	Hispanic-white segregation within MSA/PMSA		Language segregation for whites	Language segregation for Hispanics
	%Hispanic	%Hispanic		%Poor English	%Poor English
	(1)	(2)		(3)	(4)
<u>Co-worker segregation</u>					
Observed segregation					
Hispanic workers (H_H^O)	39.4	39.4	Poor English workers (P_P^O)	6.9	48.1
White workers (W_H^O)	4.5	4.5	Good English workers (G_P^O)	0.4	15.4
Difference (CW^O)	34.9	34.9	Difference (CW^O)	6.6	32.7
Random segregation					
Hispanic workers (H_H^R)	6.9	24.4	Poor English workers (P_P^R)	0.9	26.8
White workers (W_H^R)	6.9	5.6	Good English workers (G_P^R)	0.4	21.7
Difference (CW^R)	0	18.8	Difference (CW^R)	0.6	5.1
Effective segregation, $[\{CW^O - CW^R\}/\{100 - CW^R\}]\times 100$	**34.9**	**19.8**		**6.0**	**29.1**
Number of workers	1,625,953	1,625,953		1,491,434	81,595
Number of establishments	293,989	293,989		271,101	21,933

See notes to Table 5. Results in columns (3) and (4) are derived within MSA/PMSA; poor English is defined as speaking English “not well” or “not at all”; good English is speaking English well or very well.

Table 10: Hispanic-White Segregation Conditional on Language and Occupation

	Hispanic-white segregation conditional on 2 language groups	Hispanic-white segregation conditional on 4 language groups	Hispanic-white segregation conditional on 1-digit occupation (six categories)
	%Hispanic	%Hispanic	%Hispanic
	(1)	(2)	(3)
Co-worker segregation			
Observed segregation			
Hispanic workers (H_H^O)	39.4	39.4	39.4
White workers (W_H^O)	4.5	4.5	4.5
Difference (CW^O)	34.9	34.9	34.9
Conditional random segregation			
Hispanic workers (H_H^O)	26.8	29.2	26.9
White workers (W_H^O)	5.5	5.3	5.4
Difference (CW^C)	21.3	23.9	21.4
Effective conditional segregation, $[\{CW^O - CW^C\}/\{100 - CW^R\}]\times100$	**16.7**	**13.5**	**16.6**
Number of workers	1,625,953	1,625,953	1,625,953
Number of establishments	293,989	293,989	293,989

See notes to Table 5. In column (1), the two language groups are: speak English “not well” or “not at all”; speak English well or very well. In column (2), the four language groups are: speak English not at all; speak English not well; speak English well; speak English very well. Occupations are listed in notes to Table 7.

Table 11: Hispanic-White Wage Gaps and the Importance of English Language Proficiency

	Without establishment fixed effects		With establishment fixed effects	
	(1)	(2)	(3)	(4)
Hispanic	-0.277 (0.002)	-0.204 (0.002)	-0.255 (0.002)	-0.221 (0.002)
Speak English not well		0.210 (0.009)		0.138 (0.009)
Speak English well		0.396 (0.009)		0.256 (0.009)
Speak English very well		0.471 (0.009)		0.330 (0.009)
Number of observations	1,625,953	1,625,953	1,625,953	1,625,953

The dependent variable is the log of the hourly wage. There is a constant in the regressions; the category speak English not at all is omitted from the regression in columns (2) and (4).

Table 12: Language Segregation, Within MSA/PMSA

	Establishment ethnic and skill composition:		
Hispanic workers, poor English-Hispanic workers, good English		Hispanic workers, poor English-non-Hispanic workers, poor English	
	%Hispanic, poor English		%Hispanic, poor English
	(1)		(2)
Co-worker segregation			
Observed segregation			
Hispanic workers, poor English	48.1	Hispanic workers, poor English	90.0
Hispanic workers, good English	15.4	Non-Hispanic workers, poor English	26.0
Difference	32.7		64.0
Random segregation			
Hispanic workers, poor English	26.8	Hispanic workers, poor English	80.1
Hispanic workers, good English	21.7	Non-Hispanic workers, poor English	51.5
Difference	5.1		28.6
Effective segregation, $\{CW^{O} - CW^{R}\}/\{100 - CW^{R}\}]\times 100$	**29.1**		**49.5**
Number of workers	81,595		19,926
Number of establishments	21,933		6,393

See notes to Table 5.

Table 13: Effective Segregation, Sensitivity to Establishment Size

	Employment ≤ 20	Employment > 20 and ≤ 80	Employment >80 and ≤ 380	Employment > 380
	(1)	(2)	(3)	(4)
Co-worker effective segregation				
Hispanic-white	26.6	23.0	19.6	11.9
Black-white	23.5	17.6	13.3	8.8
White, low education-white, high education	18.0	16.0	15.1	12.7
Hispanic workers, poor English-Hispanic workers, good English	34.0	28.9	25.7	23.7
Hispanic workers, poor English-non-Hispanic workers, poor English	77.8	61.3	46.2	28.4

The employment cutoffs chosen are approximately the 25th, 50th, and 75th percentiles of the employment-weighted establishment size distribution in the full SSEL. Effective segregation equals $\{CW^O - CW^R\}/\{100 - CW^R\}]\times 100$.

Appendix Table A1:Probability of an SEDF Worker Appearing in the DEED

	(1)	(2)	(3)
Intercept	0.300	-0.047	-0.084
Black	-0.110	-0.056	-0.047
Hispanic	-0.074	-0.048	-0.037
Information on Write-In File:			
Employer Name		0.232	0.229
Employer Address		0.026	0.022
Employer City		-0.014	-0.013
Employer State		-0.068	-0.068
Employer Zip Code		0.106	0.102
Street Number in Address		0.202	0.194
Age			0.000
Age squared			-0.001
Female			0.010
Less than High School			-0.018
Some College			0.005
Bachelors Degree			0.010
Advanced Degree			0.001
Working Full Time			0.038
Mining			0.017
Construction			-0.036
Manufacturing			0.128
Transportation			-0.037
Wholesale			0.100
Retail			0.002
FIRE			-0.004
Manager			0.009
Service			-0.061
Farming			-0.107
Production			-0.019
Laborer			-0.016
Sample Size	11,731,793	11,731,793	11,731,793

Estimated coefficients are reported. Standard errors in all cases but one are no larger than 0.001. The standard error for Farming in column (3) is 0.002.