Workplace Segregation in the United States: Race, Ethnicity, and Skill[*]

Judith Hellerstein
Department of Economics and MPRC,
University of Maryland,
and NBER

David Neumark
University of California at Irvine,
NBER, and IZA

Revised: October 2006

[*]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 useful 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, as well as anonymous referees, for helpful comments. 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 or the Russell Sage Foundation.

Workplace Segregation in the United States: Race, Ethnicity, and Skill

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, and by race and ethnicity, and – since skill is often correlated with race and ethnicity – we assess the role of education- and language-related skill differentials in generating workplace segregation by race and ethnicity. We define segregation based on the extent to which workers are more or less likely to be in workplaces with members of the same group, and we measure segregation as the observed percentage relative to maximum segregation.

Our results indicate that there is considerable segregation by education and language in the workplace. Among whites, for example, observed segregation by education is 17% (of the maximum), and for Hispanics, observed segregation by language ability is 29%. Racial (black-white) segregation in the workplace is of a similar magnitude to education segregation (14%), and ethnic (Hispanic-white) segregation is somewhat higher (20%). Only a tiny portion (3%) 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 (32%) can be attributed to differences in language proficiency. Finally, additional evidence suggests that segregation by language likely reflects complementarity among workers speaking the same language.

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 observed 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), but the causes of the remaining gaps are more widely disputed, and many researchers attribute at least part of these wage gaps to discrimination that results in equally productive workers who belong to different groups being paid differently (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 – that is, the extent to which members of different groups tend to work with co-workers who are more like themselves than would be predicted by random allocation of workers to establishments. The evidence that does exist points to the existence of segregation in the labor market, at least along the dimensions of sex, race, and ethnicity. 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), and 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). 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.

Finally, there is almost no evidence on the extent of segregation by skill (one exception is the very limited evidence reported in Kremer and Maskin, 1996). The paucity of research on workplace segregation is presumably a function of the lack of data linking workers to the establishments in which they work.

Workplace segregation by skill 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. Alternatively, race and ethnic segregation in the workplace may be a function of varying forms of discrimination in the labor market,[1] residential segregation coupled with constraints in commuting to work (spatial mismatch), or labor market networks that exist along racial or ethnic lines.

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 role of (observable) skill differences in generating race and ethnic segregation. Our contribution is empirical in that we focus on the measurement of segregation along these dimensions, as well as exploring the extent to which segregation by skill can account for segregation by race and ethnicity. We do not explicitly test theories as to why there is segregation by skill, or why there is segregation by race and ethnicity after accounting for skill. These are important behavioral questions left to future research.

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. The matching yields data on multiple workers matched to establishments, providing the means to

[1] See, e.g., Arrow (1972).

measure workplace segregation in the United States based on a large, fairly representative data set.[2] The use of the Decennial Census of Population as the source of information on workers allows us to measure segregation along multiple dimensions and to condition our segregation measures on various characteristics of workers.

Our empirical analysis proceeds in three steps that exploit these strengths 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.[3] 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.[4]

Simple calculations of workplace segregation 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

---

[2] 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.

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

[4] 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).

conditions."[5] In contrast, workplaces – specifically establishments – are units of observation that are generated by economic forces and in which people clearly do interact in a variety of ways, including work, social activity, labor market networks, etc.[6] 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.[7] 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). In our measurement of racial and ethnic segregation, we focus on results that condition on the distribution of workers across MSA's. This helps to remove the influence of geographic segregation broadly defined, which is especially pronounced with respect to the distribution of Hispanic workers across the United States.

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.[8] Saint-Paul

---

[5] 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.

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

[7] 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.

[8] 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. Based on this model, they suggest that increased differences between more- and less-skilled workers may have led to increased segregation by skill. They also provide some very limited cross-sectional evidence on this relationship, based on the evidence on segregation by education and the distribution of education across states for U.S. manufacturing plants. Hirsch and Macpherson (1999) do not posit a formal model of sorting by skill, but assume that employers tend to hire workers of similar

(2001) generates skill segregation across firms by assuming that there are productivity-related spillovers among workers within an establishment.[9] Cabrales and Calvó-Armengol (2002) show that when workers' utility depends on interpersonal comparisons with nearby workers (such as those in the same firm), segregation by skill results.[10] 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 be generated wholly or partially as a by-product of 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 focus primarily on the extent to which segregation by English language ability can explain Hispanic-white workplace segregation, treating language ability as another important dimension of skill.[11] We also

---

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.

[9] 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 returns to skill.

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

[11] 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. Given that Hispanics have

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 skill complementarities in language are the driving factor behind the segregation of Hispanics and whites that is explained by English language proficiency differences, there should also be workplace segregation among those whose English proficiency is poor, but whose native (and spoken) languages differ. We examine this explicitly by comparing the extent of workplace segregation between Hispanics with differing levels of English proficiency with workplace segregation among Hispanics and non-Hispanics who all speak English poorly (and who presumably do not all speak Spanish).

Our analysis focuses on larger establishments; the first quartile of (employment-weighted) establishment size in our analysis is approximately 40 workers. By comparison, the first quartile of the employment-weighted size distribution of all establishments in the universe from which our establishments are drawn 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 Census Long Form respondents – 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. To the extent that workplace segregation may be generated by hiring discrimination, larger employers are an important subset in which to study workplace segregation because most legislation aimed at combating discrimination is directed at larger employers; EEOC laws cover employers with 15 or more workers and affirmative action rules for federal contractors cover employers with 50 or more workers.

---

lower education than whites, we also report on some analyses taking account of language ability as well as education.

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, and education also accounts for a non-negligible part of 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 1990 DEED, which we have created at the Center for Economic Studies at the U.S. Bureau of the Census. The 1990 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 Business Register, 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, which is not captured in the SEDF, is referred to as the "Write-In" file. We use employer names and addresses for each worker in the Write-In file to match the Write-In file to the Business Register. 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 employers' names and addresses are not necessarily recorded identically on the two files. To match workers and establishments based on the Write-In file, we use MatchWare – a specialized record linkage software program that has been used previously to link various Census Bureau data sets (Foster, et al., 1998). The first step in the matching process is to standardize employer names and addresses across the Write-In file and the Business Register, and the second step is to select and implement the matching specifications. The software uses a probabilistic matching algorithm that accounts for missing information, misspellings, and even inaccurate information. It also permits users to control which matching variables to use, how heavily to weight each matching variable, how similar two addresses must be in order to constitute a match, and how many attempts ("passes") to make in trying to find a match.

It is clear that different criteria for matching may produce different sets of matches. Matching criteria need to be broad enough to cover as many potential matches as possible, but narrow enough to ensure that only high probability matches are linked. Our general strategy was to impose the most stringent criteria in the earliest passes of the matching algorithm, 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. 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% of workers in the SEDF and 19% of establishments in the Business Register.[12] In Table 1 we provide descriptive statistics for the matched workers from the DEED

[12] 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

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% of the SEDF and 47% of the full DEED. Nonetheless, there are a few discrepancies. Perhaps most salient for this analysis is discrepancies in race and ethnicity. In the SEDF, white, Hispanic, and black workers account for 82, 7, and 8% of the total, respectively.[13] The comparable figures for the full DEED are 86, 5, and 5%. 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.[14]

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 (7) 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

---

in an industry that was considered "out-of-scope" in the Business Register. ("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 Business Register data for businesses in out-of-scope industries.)

[13] Both blacks and whites can also be classified as Hispanic, and a very small share of Hispanics (fewer than 1%) are black. However, we define black Hispanics as black, and only non-black Hispanics as Hispanic. In addition, in the analysis of Hispanic-white segregation, we drop black Hispanics.

[14] For example, approximately 4% of workers in the SEDF do not provide any business address information at all.

dummies substantially, so that conditional on including address information, blacks (Hispanics) are only 6 (5) 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 employers in the SEDF,[15] 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.[16] Second, our analysis generally focuses on differences between whites and blacks and whites and Hispanics. We therefore

---

[15] 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 Business Register census tract of the matched establishment in more than half the cases.

[16] 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 MSA's and PMSA's 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 (discussed below).

exclude individuals who do not fall into those categories, with one exception. Because one of our analyses below compares Hispanics who speak English poorly to others who speak English poorly, 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 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 Business Register and the DEED. Table 2 shows descriptive statistics for establishments in each data set. 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 Business Register have 18 employees, while the average in the DEED is 53 workers. The distributions of establishments across industries in the DEED relative to the Business Register are similar to those for workers in the worker sample. For example, manufacturing establishments are somewhat over-represented in the DEED, constituting 13% of establishments, relative to 6% in the Business Register. 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 a measure of segregation that is based on measures of the percentages of workers in an individual's establishment, or workplace, in different demographic groups. Consider a dichotomous classification of workers (e.g., whites and Hispanics). For each worker in our sample, we compute the percentage of Hispanic workers in the establishment in which that worker works, excluding the worker him or herself. Because we exclude an individual's own ethnicity in this calculation, our analysis of segregation is conducted on establishments where we observe at least two workers.

We then average these percentages separately for white workers in our sample and for Hispanic workers. These averages are segregation measures commonly used in the sociology literature. The average percentage of co-workers in Hispanic workers' establishments who are Hispanic, denoted H_H, is called the "isolation index," and the average percentage of co-workers in white workers' establishments who are Hispanic, denoted W_H, is called the "exposure index." We focus more on a third measure, the difference between these,

$$CW = H_H - W_H,$$

as a measure of "co-worker segregation." CW 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.[17]

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 – e.g., CW^O. One important point that is often overlooked in research on segregation, however, is that some segregation occurs even if workers are assigned randomly to establishments, 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 (Carrington and Troske, 1997). Rather than considering all deviations from proportional representation across establishments as an "outcome" or "behavior" to be

[17] 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 $CW' = W_W - H_W$.

explained, we subtract from our measured segregation the segregation that would occur by chance if workers were distributed randomly across establishments, using Monte Carlo simulations to generate measures of randomly occurring segregation. We denote this random segregation CW^R, and then focus on the difference $\{CW^O - CW^R\}$, which measures segregation above and beyond that which occurs randomly.[16] (Although theoretically one can have $CW^O < CW^R$ – that is, there is *less* segregation than would be generated randomly – in our data we always have $CW^O > CW^R$.) Again following Carrington and Troske, we scale this difference by the maximum segregation that can occur, or $\{100 - CW^R\}$, and refer to this ratio as "effective segregation." Thus, the effective segregation measure is:

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

which measures the share of the maximum possible segregation that is actually observed.

There are two reasons that we exclude the worker's own ethnicity when computing the fraction of Hispanics with which he or she works, First, this ensures that in large samples of workers, if workers are randomly allocated across establishments, H_H and W_H both equal the share Hispanic in the population. That is, in the case of random allocation we expect to have CW^R equal to 0. This is a natural scaling to use, and stands in contrast to what happens when the worker is included in the calculations, in which case CW^R will exceed 0 because Hispanic workers are treated as working with "themselves." Second, and perhaps more important, when the own worker is excluded our segregation measures are invariant to the sizes of establishments studied. To see this in a couple of simple examples, first consider a simple case of an economy with equal numbers of Hispanics and whites all working in two-person establishments. Establishments can therefore be represented as HH (for two Hispanic workers), HW, or WW. With random allocation, 1/4 of establishments are HH, ½ are WH, and 1/4 are WW. Thus, excluding the own worker, $H_H^R = (1/2)\cdot 1 + (1/2)\cdot 0 = 1/2$, $W_H^R = (1/2)\cdot 1 + (1/2)\cdot 0 = 1/2$, and $CW^R = 0$.[17]

[16] Of course to build up CW^R we also compute the isolation and exposure indexes that would be generated in the case of random allocation of workers, and report these as well.

[17] For the first calculation, for example, ½ of Hispanic workers are in HH establishments, for which the share Hispanic is 1, and ½ are in WH establishments, for which the share Hispanic (excluding the worker) is 0.

If we count the individual, then $H_H^R = (1/2)\cdot 1 + (1/2)\cdot(1/2) = 3/4$, $W_H^R = (1/2)\cdot(1/2) + (1/2)\cdot 0 = 1/4$, and $CW^R = 1/2$. With three-worker establishments and random allocation, 1/8 of establishments are HHH (employing 1/4 of Hispanic workers), 1/8 are WWW (employing 1/4 of white workers), 3/8 are HWW (employing 1/4 of Hispanic and ½ of white workers), and 3/8 are HHW (employing ½ of Hispanic and 1/4 of white workers). Going through the same type of calculation as above, if we include the worker, then $H_H^R = (1/4)\cdot 1 + (1/4)\cdot(1/3) + (1/2)\cdot(2/3) = 2/3$, $W_H^R = (1/4)\cdot 0 + (1/4)\cdot(2/3) + (1/2)\cdot(1/3) = 1/3$ and $CW^R = 1/3$, whereas if we exclude the worker we again get $H_H^R = 1/2$, $W_H^R = 1/2$, and $CW^R = 0$.

Although we just argued that in the case of random allocation Hispanics and whites should work with equal percentages of Hispanic co-workers on average (so that CW^R is zero), this result may not hold in parts of our analysis for two reasons. First, this is a large-sample result, and although the baseline sample size in our data set is large, the actual samples that we use to calculate some of our segregation measures are not always large, or at least not necessarily large enough to generate this asymptotic result. Second, some of our segregation measures are calculated conditional on geography (in particular, MSA/PMSA of residence), for reasons explained below. 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 still 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. For that reason, in all cases, in order to determine how much segregation would occur randomly, we conduct Monte Carlo simulations of the extent of segregation that would occur with random allocation of workers.

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 scale invariant,

meaning that they 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 and Gini indexes are unchanged. This is not true for CW. Except for those establishments that are perfectly segregated, the doubling of Hispanics leads each Hispanic worker in the sample to work with a larger percentage of Hispanic co-workers, and also each white worker to work with more Hispanics. In general, this implies that both the isolation and exposure indexes (H_H and W_H, respectively), will increase. But the isolation index will increase by more, since establishments with more Hispanics to begin with will have larger increases in the number of Hispanic workers, and hence CW will increase.[18] In our view, and we recognize that it is a subjective one, this kind of increase in the number of Hispanic workers *should* be characterized as an increase in segregation. Second, these alternative segregation measures are also sensitive to the number of matched workers in an establishment (the same issue outlined above), 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.[19]

We present some "unconditional" nationwide segregation measures, as well as "conditional" measures that first condition on metropolitan area (MSA/PMSA) of residence. In the first, the simulations randomly assign workers to establishments anywhere in the country; not surprisingly, in these simulations the random segregation measures are zero or virtually indistinguishable from zero. For comparability, when we construct these unconditional segregation measures we use only the workers

[18] More generally, W_H will also increase, but not by as much as H_H, and CW will therefore rise. For perhaps the simplest such case, start with four establishments as follows: one HHH, one HHW, one HWW, and one WWW. In this case $H_H^R = 2/3$, $W_H^R = 1/3$, and $CW^R = 1/3$. Doubling the number of Hispanics and allocating them proportionally, we get the following four establishments: HHHHHH, HHHHHW, WWHH, and WWW: In this case H_H rises to 29/36 (increasing by 5/36), W_H rises to 14/36 (increasing by 2/36), and CW rises to 15/36 (increasing by 3/36).

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

included in the MSA/PMSA sample used for the conditional analysis.[20] The unconditional estimates provide the simplest measures of the extent of segregation by skill, race, or ethnicity in the workplace. However, they reflect the distribution of workers both across cities and across establishments within cities. As such, the unconditional measures may tell us less about forces operating in the labor market to create segregation, whereas the conditional measures – which can be interpreted as taking residential segregation by city as given – may tell us more about this. Because we use the same samples for the conditional and unconditional analyses, for these analyses the observed segregation measures are identical. Only the simulations differ, but these differences of course imply differences in the effective segregation measures.

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. We use U.S. Census Bureau MSA/PMSA designations, because these are defined to some extent based on areas within which substantial commuting to work occurs.[21] 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 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, and we then compute our co-worker segregation measure for this randomly allocated sample. We do this simulation 100 times, and define our random co-worker segregation measure (CW^R) as the mean of the segregation measures across the 100 simulations. Not surprisingly, all of the random segregation measures we obtain are very precise; in all cases the standard deviations were trivially small.

---

[20] The results in this paper are generally robust to measuring segregation at the level of the MSA/CMSA metropolitan area (rather than the MSA/PMSA level), 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.

[21] 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 MSA's and PMSA's 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.

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,[22] we assessed statistical significance of measures of effective segregation or differences between them using bootstrap methods. (See the Appendix.) 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. 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.[23] Table 3 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 (PMSA's and MSA's) 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% of matched white co-workers are also low education. In contrast, high education workers

[22] 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.

[23] Below we further disaggregate workers by education when we consider how much of segregation by race is attributable to segregation by education.

work in establishments with white co-workers who are only 33.1% 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% 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 segregation measure, 20.0, is simply the observed co-worker segregation measure (i.e., $CW^R = CW^O$). This number can be interpreted as implying that 20% of the maximum amount of segregation that could arise due to non-random factors is actually observed in the data. Since there is so little evidence on workplace segregation to date, it is impossible to compare the extent of this segregation relative to any given benchmark. To us, however, this result suggests that there is substantial segregation by education.

Column (2) looks at segregation within urban areas defined as PMSA's/MSA's. 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 because, as explained above, for this simulation workers are reallocated only within urban areas); the pattern of random segregation has low education workers working, on average, with co-workers who are 43.7% low educated, while for high education workers the corresponding figure is 39.6%. As a result, the effective segregation measure in column (2) falls to 16.5. That is, about 17% of the maximum amount of segregation by education that could arise due to non-random factors is observed in the data.

Column (3) of Table 3 calculates segregation by education for blacks in the sample, conditional on the metropolitan areas 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.

Workplace Segregation by Race

Table 4 reports results for black-white segregation. In column (1) of Table 4, 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% black, while white workers work with co-workers who are 5.8% 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% black, on average, 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% black, versus an average percent black of 6.8% for whites. Based on the comparison between observed and random segregation, the effective segregation measure is 14.0, meaning that 14% of the maximum amount of racial segregation that could arise due to non-random factors is actually observed in the data. 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 3, 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 5 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 5, which reports the observed segregation between blacks and whites, repeats the results from Table 4. We report the conditional random segregation measures starting in the middle of the rows of Table 5. 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% black, and white workers working, on average, with co-workers who are 6.8% black. These numbers are very close to the (unconditional) simulated numbers reported in Table 4, 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%) of overall black-white segregation.

We repeat this analysis in column (2) of Table 5, 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); after all, occupation is not an exogenous worker characteristic, but an outcome of the hiring process. We explore the role of occupation sorting by computing random segregation conditional on six one-digit occupation categories (listed in the notes to the table) in column (3) of Table 5. While this conditioning has slightly more effect than conditioning on education, the effective conditional segregation measure is still 12.9, accounting for only 8% of overall black-white segregation.

While education (and occupation) only account for a small fraction of workplace segregation by race, it is not the case that education differences between blacks and whites are too small in this sample to have potentially 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. Moreover, these differences can explain a large fraction of black-white wage differences.

To show this explicitly, in Table 6 we report the education distributions of workers by race, and we report estimates of black-white wage gaps with and without accounting for educational differences. In columns (1) and (2) we report the educational distributions among whites and blacks. Only 10% of whites in the sample have less than a high school degree, whereas 18% of blacks do. In contrast, at the

top end of the education distribution, 25% of whites have at least a college degree but only 14% of blacks do. In column (3) we report that the coefficient on the black 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%, indicating that education accounts for a large share of the black-white wage. Column (5) replicates the specification in column (3), but includes establishment fixed effects. The coefficient on the black dummy actually becomes more negative when we include establishment fixed effects, implying that blacks work in slightly higher-wage establishments, rather than lower-wage establishments.[24] When we add the education controls to this specification, in column (6), the coefficient on the black dummy again falls by about one-third.

The fact that when education controls are added the coefficient on the black dummy falls by the same amount with or without the establishment fixed effects 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. At the same time, including the establishment fixed effects does substantially reduce the estimated returns to education, indicating that there is sorting by education across establishments, with more-educated workers in higher-wage establishments. But the sorting of workers by education across establishments (that we established directly in Table 3) is largely independent of the sorting of workers by race.

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.

---

[24] Including one-digit industry dummy variables in the regression leaves the coefficient on the black dummy almost unchanged and has very little effect on the coefficients on the education variables.

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 7, and the basic conclusion is that there is extensive workplace segregation by Hispanic ethnicity. The segregation figures for the unconditional national indexes indicate more segregation by ethnicity than their counterparts for race as reported in Table 4. Specifically, in column (1) of Table 7 the average share of the establishment workforce that is Hispanic for Hispanic workers is 39.4%, versus a comparable figure of 23.7% 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 7, for example, the randomly allocated share Hispanic for Hispanic workers is 24.4%, compared with a parallel share Hispanic for white workers of 5.6%. This difference arises because Hispanics are much less evenly dispersed across metropolitan areas than are blacks, with some metropolitan areas having 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 7, 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% speak English poorly. In contrast, for white workers in the good English category, only 0.4% 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 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% also speak English poorly. In stark contrast, Hispanics in the "good English" category are likely to have Hispanic co-workers of whom, on average, only 15.4% 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% of Hispanic co-workers speaking English poorly, while workers in the "good English" category would have 21.7% of co-workers speaking English poorly. 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 8, 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 levels of 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 7, 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% Hispanic, whereas whites work with co-workers who are 5.5% 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% = (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 conditional on 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.

Finally, because Hispanics also have lower education than whites, and education can independently contribution to segregation (and lower education is associated with worse language

ability),[25] in the final column of Table 8 we look at skill along two dimensions, asking how much segregation by both language and education accounts for Hispanic-white segregation. We find that the remaining Hispanic-white segregation falls somewhat further compared with the estimates in column (2), which uses the same language skill breakdown but ignores education, with effective conditional segregation falling to 11.1. This implies that skill segregation based on language and schooling accounts for 44% of Hispanic-white segregation, up from 32% when we conditioned only on language, reinforcing the conclusion that segregation by skill contributes substantially to ethnic segregation.[26]

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. This difference in results is also reflected in wage equation estimates. Columns (1) and (2) of Table 9 report the distributions of self-reported English language proficiency for whites and Hispanics, respectively. In the sample, almost 99% of (a very large sample of) whites report speaking English very well, whereas only 63% 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

[25] For example, 38% of Hispanics have less than a high school education, versus 10% of whites, and only 10% have a college degree or more, versus 25% for whites. And of Hispanics who speak no English or speak English poorly, 77% have less than a high school education, while of those who speak English very well, only 22% have less than a high school education.

[26] Although not shown in the table, when we conditioned only on the four education categories, effective conditional segregation was 16.2, compared with 13.5 when we condition only on the four language categories. Because language ability and education are closely related, the results conditional on only one or only the other capture the effects of both.

coefficient on the Hispanic dummy to fall to −0.204, a 26% drop.[27] Like for the black-white wage gap and education, skill therefore accounts for a sizable share of the Hispanic-white wage gap.[28]

Columns (5) and (6) 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, adding English language proficiency only causes the Hispanic-white wage gap to fall to −0.221, a 13% drop, accounting for less of the Hispanic-white wage gap than when establishment fixed effects were not included. In contrast to the results for blacks and whites, then, this smaller role for language within than across establishments implies that the role of language in explaining the Hispanic-white wage gap arises through sorting of Hispanics and whites across establishments based on language. This, too, is consistent with our evidence showing that language contributes substantially to Hispanic-white workplace segregation.

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 language is an important skill, and that language segregation arises as employers seek to exploit complementarities among workers who speak the same language; because language proficiency is correlated with ethnicity, segregation by language generates segregation by ethnicity. Another possibility, though, is that language skills per se are not driving the segregation, but rather that language is associated with other dimensions along which employers make hiring decisions that reflect their

---

[27] The result is larger (a 42% 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. 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).

[28] For the sake of brevity we limit our focus in Table 9 to language differences because language differences are larger than education differences between Hispanics and whites, and because we find that more of the Hispanic-white workplace segregation we document in Table 8 can be explained by language differences than by education differences. Education also helps explain the Hispanic-white wage gap, however, and interestingly it actually explains more of the wage gap than language.

discriminatory tastes, and on the basis of which employers crowd workers into a subset of jobs (typically jobs that pay less). Alternatively, poor English skills can reflect low levels of other unobserved skills, so that the language segregation just reflects skill segregation along other dimensions. It can be difficult to distinguish between these competing hypotheses.[29] 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 more likely arises because of complementarity between workers who speak the same language (or a related economic incentive to segregate workplaces by common language). And conversely, if poor language skill was simply a proxy for low unobservable skill, we would expect less segregation between Spanish and non-Spanish speakers with poor language skills than between Hispanics with poor language skills and Hispanics with better language skills. Of course segregation by language could also 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

[29] 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.)

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 10. Column (1) repeats the calculations from Table 7, column (4), 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) random segregation is far from zero, much of this resulting from sorting across MSA/PMSA's. 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. In particular, this evidence is consistent with a skill-based explanation for the large role that English language proficiency plays in explaining Hispanic-white segregation.

Differences in Workplace Segregation by Establishment Size

In Table 11 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.[30] Second, as noted earlier, EEO and affirmative action target larger employers, which may tend to discourage segregation in large establishments.[31]

---

[30] 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).

[31] 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).

The estimates are consistent with these expectations. In the first two rows, Hispanic-white and black-white 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. 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-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.

Results with Duncan Index

Finally, we have presented all of our results thus far using the co-worker segregation measure. Although, as explained above, we have some preference for this measure compared to other establishment-based segregation measures such as the Duncan index, it is useful to know how robust our results are, at least qualitatively, to the choice of segregation measure. In Table 12, therefore, we first summarize the key segregation results reported in the previous tables (in the first column), and then give the same results based on the Duncan index. As before, we focus on effective segregation measures – both unconditional and conditional – which can be defined for the Duncan index just as we have done for the co-worker measure. We do not expect exactly the same results, of course, because the Duncan index has different properties than our co-worker measure. In particular, it is an establishment-based measure rather than a worker-based measure. For example, as noted earlier, it does not change if we double the

number of Hispanics; and even under random allocation of workers it is sensitive to the size distribution of establishments.

Nonetheless, as Table 12 indicates, the results are qualitatively very similar using the different segregation measures. Focusing on the conditional segregation measures, using the Duncan index education accounts for a bit more of race segregation, with estimates ranging from 4.1 to 11.8%, compared with 0.7 to 7.9% using the co-worker segregation measure. In either case, though, nearly all of the black-white segregation remains unaccounted for by education. Similarly, language skills explain a substantial amount of Hispanic-white segregation. In the first two rows of Table 8 the numbers are quite comparable for the different segregation measures. For example, using the four language categories, language explains 31.8% of Hispanic-white segregation using the co-worker measure, and 34.6% using the Duncan index. One difference is that other skill dimensions accounts for somewhat more of Hispanic-white segregation using the Duncan index, as reflected in the rows conditioning on 1-digit occupation, and conditioning on both language and education. And finally, the evidence pertaining to the sources of language segregation still suggests that a substantial part of language segregation likely reflects the need to group together workers who speak the same language, rather than other sources of segregation; in particular, segregation of poor-English-speaking Hispanics from poor-English-speaking non-Hispanics is considerably higher than segregation of poor-English-speaking Hispanics from Hispanics who speak English better, although with the Duncan index the difference is a bit smaller.

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 (17% by our measure, which is the percentage of observed segregation relative to the maximum our segregation measure could take on), consistent with models where employers find it efficient to segregate workers by skill. Similarly, among Hispanics we document extensive segregation by language (29%), 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 (14%), and segregation by Hispanic ethnicity that is somewhat higher (20%).

After documenting these different dimensions of segregation, our analysis focuses on whether racial and ethnic workplace segregation reflects race or ethnicity per se, 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 (3 to 8%) 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 (32%) 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 (3). The data generating process for that sample can be approximated to a first order as a random sample of workers who then are matched to their establishment, where all workers have the same probability of being matched. We then consider the individual-level characteristics of a worker and the characteristics of that worker's matched co-workers (e.g., percent black, percent Hispanic) as fixed for that worker, so that we effectively have a random sample of workers with data that describes characteristics of each of those workers. For our bootstrap exercise we draw with replacement a sample of workers from the Restricted DEED sample, with the sample size equal to that of the Restricted DEED itself. We then calculate all of the observed segregation measures reported in the paper for that bootstrap sample, making sample restrictions for each table in the paper as necessary from that bootstrap sample. We repeat this 100 times. 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 4, 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.

Arrow, Kenneth J. 1972. "Some Mathematical Models of Race Discrimination in the Labor Market." in Racial Discrimination in Economics Life, ed. A. H. Pascal (Lexington, MA: D.C. Heath), pp. 187-204.

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.

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.

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

	SEDF (1)	Full DEED (2)	Restricted DEED (3)	Black/white sample (4)	Hispanic/white sample (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
Bachelors degree	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

	Business Register	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 Business Register. For these 55, the workers reported place of work was used to determine MSA status.

Table 3: 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^R)	41.3	43.7	51.6
High education workers (H_L^R)	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%. (The hypothetical maximum is 16.7%, 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 4: 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 3.

Table 5: 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, $[\{CW^O - CW^C\}/\{100 - CW^R\}]\times 100$	**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 3. 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 6: Black-White Wage Gaps without and with Establishment Fixed Effects

	Sample means		Regression results			
	Whites	Blacks	Without establishment fixed effects		With establishment fixed effects	
	(1)	(2)	(3)	(4)	(5)	(6)
Black	0	1	-0.204 (0.002)	-0.127 (0.002)	-0.232 (0.002)	-0.164 (0.002)
Less than a high school degree	0.10	0.18		...		...
High school degree	0.31	0.32		0.196 (0.002)		0.096 (0.002)
Some college or Associates degree	0.33	0.37		0.331 (0.002)		0.205 (0.002)
Bachelors degree or above	0.25	0.14		0.744 (0.002)		0.534 (0.002)
Number of observations	1,503,640	115,236	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 (3) and (4) and a constant is included.

Table 7: 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)
Co-worker segregation					
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 3. 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 8: 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-white segregation conditional on 4 language and 4 education groups
	%Hispanic	%Hispanic	%Hispanic	%Hispanic
	(1)	(2)	(3)	(4)
Co-worker segregation				
Observed segregation				
Hispanic workers (H_H^O)	39.4	39.4	39.4	39.4
White workers (W_H^O)	4.5	4.5	4.5	4.5
Difference (CW^O)	34.9	34.9	34.9	34.9
Conditional random segregation				
Hispanic workers (H_H^O)	26.8	29.2	26.9	31.0
White workers (W_H^O)	5.5	5.3	5.4	5.1
Difference (CW^C)	21.3	23.9	21.4	25.9
Effective conditional segregation, $[\{CW^O - CW^C\}/\{100 - CW^R\}]\times 100$	**16.7**	**13.5**	**16.6**	**11.1**
Number of workers	1,625,953	1,625,953	1,625,953	1,625,953
Number of establishments	293,989	293,989	293,989	293,989

See notes to Table 3. 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 5. The education groups used in column (4) are the same as those in Table 5.

Table 9: Hispanic-White Wage Gaps and the Importance of English Language Proficiency

	Sample means		Regression results			
	Whites	Hispanic	Without establishment fixed effects		With establishment fixed effects	
	(1)	(2)	(3)	(4)	(5)	(6)
Hispanic	0	1	-0.277 (0.002)	-0.204 (0.002)	-0.255 (0.002)	-0.221 (0.002)
Speak English "not at all"	0.0002	0.05		...		...
Speak English "not well"	0.0036	0.14		0.210 (0.009)		0.138 (0.009)
Speak English well	0.0072	0.184		0.396 (0.009)		0.256 (0.009)
Speak English very well	0.989	0.626		0.471 (0.009)		0.330 (0.009)
Number of observations	1,513,277	112,676	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 in columns (3) and (4) and the category speak English not at all is omitted.

Table 10: 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 3.

Table 11: 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 Business Register. Effective segregation equals $\{CW^O - CW^R\}/\{100 - CW^R\}]\times 100$.

Table 12: Effective Segregation Measures Based on Co-Worker Segregation and Duncan Index

Table 3	Co-Worker	Duncan
Segregation by education for whites	16.5	24.5
Segregation by education for blacks	15.0	17.6
Table 4		
Black-white segregation	14.0	18.4
Table 5		
Black-white segregation conditional on 2 education groups (% explained)	13.9 (0.7%)	17.6 (4.1%)
Black-white segregation conditional on 4 education groups (% explained)	13.6 (2.9%)	16.4 (10.8%)
Black-white segregation conditional on 1-digit occupations (% explained)	12.9 (7.9%)	16.2 (11.8%)
Table 7		
Hispanic-white segregation	19.8	19.8
Table 8		
Hispanic-white segregation conditional on 2 language groups (% explained)	16.7 (15.7%)	16.4 (16.8%)
Hispanic-white segregation conditional on 4 language groups (% explained)	13.5 (31.8%)	12.9 (34.6%)
Hispanic-white segregation conditional on 1-digit occupations (% explained)	16.6 (16.2%)	13.8 (30.0%)
Hispanic-white segregation conditional on 4 language and 4 education (% explained)	11.1 (43.9%)	8.73 (55.8%)
Table 10		
Hispanic workers, poor English-Hispanic workers, good English	29.1	37.4
Hispanic workers, poor English-non-Hispanic workers, poor English	49.5	52.8

All calculations are within MSA/PMSA. See notes to corresponding tables.

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.