New Evidence on Sex Segregation and Sex Differences in Wages from Matched Employee-Employer Data

Kimberly Bayard, *Board of Governors of the Federal Reserve System*

Judith Hellerstein, *University of Maryland and National Bureau of Economic Research*

David Neumark, *Public Policy Institute of California, Michigan State University, and National Bureau of Economic Research*

Kenneth Troske, *University of Missouri—Columbia and the Institute for the Study of Labor*

We use new matched employer-employee data to estimate the contributions of sex segregation and wage differences by sex within occupation, industry, establishment, and occupation-establishment cells to the overall sex gap in wages. In contrast to earlier data used to study this question, our data cover all industries and occupations across all regions of the United States. We find that segregation of women into lower-paying occupations, industries, establishments, and occupations within establishments accounts for a sizable fraction of the sex gap in wages. Nonetheless, approximately one-half of the sex gap in wages remains attributable to the individual's sex.

We thank Fran Blau, Bill Evans, Claudia Goldin, Erica Groshen, Larry Katz, Robert LaLonde, Jonathan Leonard, Brooks Pierce, Peter Schmidt, and seminar/

[*Journal of Labor Economics*, 2003, vol. 21, no. 4]
© 2003 by The University of Chicago. All rights reserved.
0734-306X/2003/2104-0006$10.00

I. Introduction

Women have consistently earned lower wages than men in U.S. labor markets, although this gap has narrowed in recent decades (Blau 1998). Understanding the sources of sex differences in wages is vital to determining why the wage gap between men and women persists. Previous research has focused on the impact of the occupational segregation of men and women on the wage gap (e.g., Macpherson and Hirsch 1995), the effect of industry segregation (e.g., Fields and Wolff 1995), and, to a lesser extent, on the segregation of men and women into different employers (Blau 1977; Bielby and Baron 1984; Carrington and Troske 1998). These studies all find evidence that the wage gap falls considerably after accounting for segregation.

Evidence on the contribution to the wage gap of within-establishment, within-occupation segregation is far harder to find. Indeed, we are not aware of any empirical work on this issue that uses large data sets representative of a wide array of industries. The reason for this is the paucity of data sets containing detailed demographic information for multiple workers in the same establishment. As a result, studies of the effects of establishment and occupation-establishment segregation have used unusual, quite narrow data sets. For example, the best-known study is by Groshen (1991); it uses surveys of wages for a subset of occupations in five specific industries included as part of the Bureau of Labor Statistics (BLS) Industry Wage Surveys (IWS). In earlier work, Blau (1977) used the BLS Area Wage Surveys to provide a decomposition of the sex gap in wages, including evidence on the importance of an individual's sex within occupation, establishment, and job cell. Her data covered subsets of three broad occupations in three large northeastern cities.

The focus in these studies on a handful of industries or occupations provides something closer to a set of case studies, with the lack of representativeness limiting their usefulness in assessing the forces at work in

conference participants at Princeton University, the University of Kentucky, the University of Maryland, the University of California, Los Angeles, the National Bureau of Economic Research, the Center for Economic Studies at the U.S. Census Bureau, the American Economics Association and Southern Economics Association meetings, the Bureau of Labor Statistics, and the International Symposium on Matched Employer-Employee Data for helpful comments. We also thank Jennifer Foster, Nicole Nestoriak, and Scott Adams for outstanding research assistance. This research was supported by National Science Foundation (NSF) grant SBR95-10876 through the National Bureau of Economic Research (NBER). The research in this article was conducted while the last three authors were research associates with the Center for Economic Studies, U.S. Bureau of the Census. Part of this research was conducted with restricted access to Bureau of Labor Statistics (BLS) data on site. Research results and conclusions expressed are ours and do not indicate concurrence by the Bureau of the Census, the Center for Economic Studies, the Bureau of Labor Statistics, or the Public Policy Institute of California.

generating the sex wage gap in the United States. Our goal in this article is to use a much broader and more nationally representative data set to estimate the contributions of sex segregation by industry, occupation, and occupation-establishment cell (job cell) to the sex wage gap. For our analysis, we construct and use an extended version of the Worker-Establishment Characteristics Database (WECD) to decompose the source of male-female wage differentials. Like the WECD, this data set uses the U.S. Census Bureau's Standard Statistical Establishment List (SSEL) to identify the employers of individuals who responded to the long form of the 1990 Decennial Census. However, whereas the WECD is limited to manufacturing plants, this new data set (the New Worker-Establishment Characteristics Database, or NWECD) includes workers and establishments from all sectors of the economy and all regions.[1] Nonetheless, because of the constraints imposed by matching employees to employers, some nonrepresentative characteristics of the data set are unavoidable.

Using the NWECD, we provide new estimates of the role of various dimensions of sex segregation in generating sex differences in wages. Although in some respects our evidence may be viewed as complementary to that in the earlier studies, in our view, the NWECD, while having some shortcomings, is clearly better suited to characterizing the effects of sex segregation in U.S. labor markets. Our results indicate that a sizable fraction of the sex gap in wages is accounted for by the segregation of women into lower-paying occupations, industries, establishments, and occupations within establishments. We also find, however, that a very substantial part of the sex gap in wages remains attributable to the individual's sex.

II. The Data

The data used in this study come from a match between worker records from the 1990 Sample Edited Detail File (SEDF) to establishment records in the 1990 Standard Statistical Establishment List (SSEL). The 1990 SEDF consists of all household responses to the 1990 Decennial Census long form. As part of the Decennial Census, one-sixth of all households receive a "long-form" survey, which asks a number of questions about each member of the household ("person questions") as well as about the housing unit ("housing questions"). Those receiving the long form are asked to identify each employed household member's (1) occupation, (2) employer location, and (3) employer industry in the previous week. The Census Bureau then assigns occupational, industrial, and geographic codes to long-form responses. Thus, the SEDF contains the standard demographic information for workers collected on the long form of the Decennial

[1] See Troske (1998) and Bayard et al. (1999) for descriptions of these data sets.

Census, along with detailed location information and a three-digit census industry code for each respondent's place of work.

The SSEL is an annual list of business establishments maintained by the U.S. Census Bureau. The SSEL contains detailed location information and a four-digit SIC code for each establishment, along with a unique establishment identifier that is common to other Census Bureau economic surveys and censuses. It also includes information on total payroll expenses, employment, and whether or not the establishment is part of a multi-establishment firm.

We matched workers and establishments using the detailed location and industry information available in both data sets. We did this because we did not actually have the employer name available on both establishment and worker records. Briefly, the first step in the matching process was to keep only establishments unique to an industry-location cell. Next, all workers indicating that they work in the same industry-location cell as a retained establishment were linked to the establishment. The matched data set is the NWECD. Because the SEDF contains only a sample of workers and because not all workers are matched, the matched data set includes a sample of workers at each establishment. Complete details of the matching procedure are provided in the appendix.

In our matched sample, we impose some restrictions on both individuals and establishments. We include only individuals who report usually working between 30 and 65 hours per week and 30 or more weeks in the last year (1989). These restrictions on hours and weeks are intended to pick out full-time, full-year workers who are less likely to have changed jobs in the past year, as well as those whose hours are so high that they may have held multiple jobs. We make these restrictions for three reasons. First, because the Decennial Census collects data on earnings from all jobs, rather than wages on the current job, we need to try to eliminate variation in wages that stems from multiple job holding at a point in time or during the previous calendar year.[2] Second, because the 1990 Decennial Census asks workers to report the address of the establishment where they worked in the previous week, while the earnings data are for the previous calendar year, job changing may lead to inaccurate measurement of earnings in the matched data. Imposing restrictions that get us closer to full-year, full-time workers should disproportionately eliminate work-

[2] Multiple job-holding rates are virtually identical among men and women. The 1996 Current Population Survey (CPS) data indicate rates of 6.2% for men and 6.1% for women (Stinson 1997). Thus, although multiple job holding could affect our data, it is unlikely to influence the sex differences we estimate.

ers who have changed jobs.[3] Finally, the IWS data, with which we eventually are interested in drawing some comparisons, cover only full-time workers. We also restrict the sample to workers aged 18–65, with a constructed hourly wage ((annual earnings/weeks worked)/usual hours worked per week) in the range $2.50–$500, and we exclude those working in establishments in public administration (in order to restrict our focus to the private sector).

We also require that establishments have total employment of at least 25 workers. We do this for two reasons: first, when we compared average establishment-level worker earnings in the matched observations in the SEDF with average payroll expenses in the SSEL, these corresponded much more closely for establishments with 25 or more workers; second, the IWS industry samples included mainly establishments with 25 or more workers. In addition, to ensure that we have a reasonable basis for estimating the characteristics of an establishment's workforce, we required that the number of matched workers be at least 5% of employment as reported in the SSEL. Finally, we eliminated the less than .1% of establishments that reported earnings exceeding more than $600,000 per worker.

Table 1 documents the effects of these various matching rules and exclusion restrictions on the sample size, the number of matched workers, and average earnings and employment calculated from both the SSEL and SEDF data. We define measures of establishment earnings per worker from data in both the SSEL and the SEDF. For the SSEL, earnings per worker are constructed as Total Annual Payroll/Total Employment. For the SEDF, establishment earnings per worker are created by averaging

[3] We used the March 1990 Basic CPS file and Income Supplement to attempt to gauge the extent to which these restrictions accomplish this. We first extracted a sample that corresponds to our SEDF sample along other dimensions. For this sample, we estimated the proportion that had not "changed jobs" over the period from the beginning of the previous calendar year to the March interview. We identified such individuals as those who held only one job over the course of the previous calendar year (the survey instructs respondents to ignore multiple jobs held at the same time) and who are working in the same three-digit industry at the March interview as in the last job held in the previous calendar year (the only available information on job change for this interval [Stewart 1998]). Overall, 76.52% of the sample satisfies these criteria. When we impose the weeks and hours restrictions for the previous calendar year, this percentage rises to 78.62%, indicating that the weeks and hours restrictions tend to screen job changers, but not in a very disproportionate fashion. Thus, measurement error in the wage because of job changing remains a problem. We do note, however, that this phenomenon is very similar for men and women, as the percentages that had not changed jobs (with or without the restrictions) differ by no more than .5 percentage point. Thus, it seems that any measurement error is not systematically related to sex.

Table 1
Construction of the New Worker-Establishment Characteristics Database (NWECD)

	Number of Establishments (SSEL) (1)	Number of Workers (SEDF) (2)	Average Number of Matched Workers (3)	Average Establishment Employment from SSEL Data (4)	Average Establishment Earnings/Worker from SSEL Data (5)	Average Establishment Earnings/Worker from SEDF Data (6)	Average Earnings/Worker from SEDF Data (7)
A. Establishments in SSEL with positive employment	5,593,379			21.10 (.45)	19,310.37 (10.98)		
B. Establishments in unique industry-location cells	388,787			41.17 (2.30)	17,102.78 (28.97)		
C. Workers among long-form respondents in SEDF		17,311,211					20,977.51 (6.75)
D. Respondents matched to establishments in B	201,944	1,720,423	8.52 (.08)	66.46 (4.43)	18,123.98 (205.00)	17,094.00 (38.83)	20,831.89 (17.28)
E. Discard matches with imputed industry or location, or number of matched workers greater than SSEL employment; discard workers with zero or missing earnings, or working outside United States	156,332	1,117,424	7.15 (.07)	83.24 (5.72)	18,218.33 (181.34)	19,416.60 (47.81)	22,582.18 (21.21)
F. Exclude workers with hours < 30 or > 65, weeks in 1989 < 30, age < 18 or > 65, wage < \$2.50 or > \$500, public administration	129,021	845,036	6.55 (.07)	83.59 (6.70)	19,224.23 (218.75)	23,112.66 (53.66)	25,611.43 (23.88)
G. Discard establishments with employment < 25, number of matched workers < 5% of SSEL employment, and earnings outliers; final sample	32,931	637,718	19.37 (.26)	180.84 (2.68)	20,983.40 (63.64)	23,327.76 (71.73)	25,978.52 (26.58)

NOTE.—SSEL = U.S. Census Bureau's Standard Statistical Establishment List. SEDF = 1990 Sample Edited Detail File. Standard errors of means are reported in parentheses.

the annual wages and salaries of all workers matched to the establishment.[4] The table shows that 7% of establishments can be assigned to unique industry-location cells. These establishments are nearly twice as large as those in the overall sample (with an average of 41.17 workers vs. 21.10 for the full SSEL sample), but they have average earnings that are lower by about $2,200. This does not contrast with standard size-wage effects (Brown and Medoff 1989), since there are no controls for industry, and so forth, and the ability to assign establishments to unique industry-location cells is not random with respect to these characteristics.

The next three rows (C–E) provide information on the observations on workers in the SEDF. Out of a total of 17,311,211 workers, we match 1.1 million, or 6.5%, to establishments, once we discard unreliable matches or workers without earnings data. There are, of course, numerous establishments to which no workers are matched, reflected in the decline in the number of matched establishments from 388,787 to 201,944, based on the simple match, and 156,332, once other restrictions are imposed. Naturally, the establishments to which workers in the SEDF are matched tend to be larger, with an average employment of 83.24. The last three columns compare earnings data. Average establishment earnings per worker reported in the SSEL are about $1,200 lower than the corresponding figure estimated from the SEDF ($18,218 vs. $19,416); this is presumably in part attributable to the fact that, in the SEDF, individuals can report earnings from more than one job. We also find, comparing columns 6 and 7 of row E, that average earnings per worker in the SEDF data are about $3,170 higher than the average establishment earnings estimated from the same (matched) data. These numbers can differ because the earnings per establishment figures are not weighted by the number of matched workers in estimating average establishment earnings per worker; thus, this result likely stems from the concentration of higher-earning workers in larger establishments. Row F drops workers based on the restrictions on hours, weeks, age, wages, and so forth, with little impact except to drop those with lower earnings.

[4] It would be ideal to use actual hourly wages whenever possible, but these are not available in the SEDF. We therefore examined data from the March 1990 Basic CPS file and Income Supplement to gauge the possible sensitivity of the results to using a constructed wage. In particular, we extracted the outgoing rotation group with similar restrictions to those imposed on the SEDF sample we use. We restricted attention to hourly workers for whom the reported March hourly wage is an actual hourly rate, not constructed. For this same sample, we also constructed an hourly wage based on earned income in 1989 divided by an estimate of total hours worked, paralleling the SEDF measure. We then estimated standard wage regressions with similar controls to those used in the SEDF (except for the percent-female variables). The estimated wage regressions—and in particular the coefficients of the sex dummy variable—were very similar using these two wage measures, indicating that use of this constructed wage in the SEDF is unlikely to be problematic.

The final sample is described in row G, which we obtain after dropping establishments with fewer than 25 employees, with an insufficiently small percentage of matched employees, and with earnings outliers. We end up with a sample of 637,718 workers matched to 32,931 establishments. These establishments are, of course, much larger than those represented in the previous rows, and they have an average of 19.37 workers matched to them. We also find that, in this subset of larger establishments, average establishment earnings estimated from the SSEL and the SEDF are considerably closer ($20,983 vs. $23,328).

Descriptive statistics for the matched sample are reported in column 1 of table 2. The sample is approximately 47% female and 7% black, with an average age of 40. The percentage currently married is 71%. With respect to education, 21.1% have a bachelor's degree or higher, and 50% report no college education. Column 2 reports descriptive statistics for the entire SEDF file, with the weeks and hours restrictions imposed. Most of the demographic characteristics are quite close in the matched sample and the full sample. Geographically, individuals living in metropolitan statistical areas (MSAs) are less likely to be in the matched sample, presumably because in urban areas individuals are less likely to work in unique industry-location cells. Turning to occupation, laborers are overrepresented in the matched sample, and support occupations are underrepresented. Similarly, the industry composition of the sample is heavily weighted toward manufacturing, with 52% of workers in this industry versus 24% in the full sample, while retail is grossly underrepresented, presumably because many retail establishments are in locations in which similar establishments are located (such as malls). In the empirical analysis, we address the potential consequences of the overrepresentation of manufacturing establishments. The remaining columns of table 2 compare the descriptive statistics for the NWECD and SEDF separately by sex. The patterns of overrepresentation and underrepresentation are similar, with manufacturing overrepresented for both men and women, retail underrepresented, and so forth. In addition, the sex differences look similar across the two samples. For example, women's earnings are 61.6% of men's in the NWECD, as compared with 60.9% in the SEDF.

We noted above that small establishments (those with fewer than 25 employees) are dropped from the sample. If women are overrepresented in smaller establishments, then, given that smaller establishments also pay lower wages, we may understate the contribution of establishment segregation to the wage gap. To examine this question, we looked at information on the representation of women in establishments of different sizes, using the May 1988 CPS Survey of Employee Benefits Supplement. Across all industries, women are not overrepresented in smaller establishments. The percent female in establishments with fewer than 25 employees is just over 43%, as compared with 44.6% overall. However, there are slight differences across the manufacturing and nonmanufacturing sectors, with women somewhat underrepresented in the smaller establishments in the manufacturing sector (26.1% female in establishments with fewer than 25 employees, as compared with 32.2% overall) and only slightly underrepresented in these smaller establishments in the nonmanufacturing sector (45%, as compared with 48.7% overall). On the other hand, only a small proportion (2.9%) of the manufacturing workforce is employed in plants with fewer than 25 employees. The magnitudes suggest that it is unlikely that our establishment-size cut-off has much impact. At any rate, the underrepresentation of women in smaller establishments, coupled with lower wages in smaller establishments, suggests, if anything, upward bias in our estimate of the contribution of establishment segregation.

The fact that the NWECD is not a representative sample of U.S. workers is not surprising given the requirements for a match and given the size restrictions imposed on matched establishments. For our purposes, however, the most important question is whether the NWECD is not representative in ways that will bias the wage regressions we estimate. To partially answer this question, we report in table 3 estimates from basic wage regressions with and without industry and occupation controls. Columns 1 and 2 provide benchmark estimates from wage regressions, first with no controls and then with the basic demographic and human capital controls but without industry and occupation controls, respectively, using workers from the SEDF. These are followed by specifications including interactions between the female dummy variable and age and its square, and then adding in the industry and occupation controls.

Not surprisingly, the results from the SEDF are very similar to those from other large, nationally representative data sets (such as the Current Population Survey [CPS]). The male-female wage gap in column 2 is 31.6%, but it falls to 23.8% in column 4 when we control for broad occupation and industry categories. Similarly, the black-white wage gap is significant in both regressions, but it is considerably smaller in column 4. The estimates show evidence of quadratic age profiles and positive returns to education, although the returns to education are smaller in column 4. The specifications with the female-age interactions in columns 3 and 5 indicate slower wage growth with age for women over most of the age range; this is expected, and it is likely attributable to age overstating experience and tenure more for women than for men and perhaps also lower human capital investment per unit of time in the labor market among women. Columns 6–10 replicate the specifications of columns 1–5, but they use the NWECD data. The male-female wage gap in column 7 is 36.2%, which is somewhat larger than that in the SEDF (31.6%), but the difference in the male-female wage gap between the two data sets is virtually eliminated once we control broadly for industry and occupation

Table 2
Descriptive Statistics for New Worker-Establishment Characteristics Database (NWECD) and Full 1990 Sample Edited Detail File (SEDF), Full-Time, Full-Year Workers

	Women and Men Combined		Women		Men	
	NWECD (1)	SEDF (2)	NWECD (3)	SEDF (4)	NWECD (5)	SEDF (6)
Annual earnings	25,978.52	27,259.79	19,512.50	19,931.10	31,674.73	32,735.24
	(21,227.38)	(29,275.13)	(13,696.45)	(15,477.93)	(24,759.25)	(35,330.59)
Log hourly wage	2.349	2.336	2.150	2.148	2.525	2.476
	(.539)	(.600)	(.493)	(.528)	(.516)	(.612)
Demographics:						
Female	.468	.428				
Age	39.589	38.440	39.423	38.230	39.735	38.598
	(10.979)	(11.272)	(10.986)	(11.239)	(10.970)	(11.295)
Black	.070	.077	.082	.094	.059	.064
Currently married	.712	.661	.647	.596	.769	.710
High school degree	.369	.325	.359	.334	.378	.319
Some college	.191	.216	.190	.229	.191	.207
Associate's degree	.095	.078	.121	.091	.071	.069
Bachelor's degree	.128	.159	.129	.157	.127	.160
Advanced degree	.083	.085	.078	.078	.088	.091
Location:						
Metropolitan statistical area	.575	.764	.549	.762	.597	.765
New England	.044	.056	.042	.058	.046	.056
Mid-Atlantic	.143	.153	.148	.152	.139	.155
East North Central	.273	.195	.259	.189	.285	.200
West North Central	.122	.092	.123	.093	.121	.091

South Atlantic	.137	.165	.147	.174	.128	.159
East South Central	.083	.059	.085	.061	.081	.058
West South Central	.086	.095	.088	.095	.084	.094
Mountain	.038	.050	.036	.049	.039	.051
Pacific	.075	.134	.072	.130	.078	.137
Occupation:						
Manager	.270	.281	.308	.310	.238	.260
Support	.230	.306	.338	.432	.134	.213
Service	.086	.088	.115	.122	.060	.063
Farming	.004	.016	.001	.005	.007	.024
Production	.145	.135	.042	.028	.236	.215
Laborer	.265	.173	.196	.104	.326	.225
Industry:						
Agriculture	.004	.018	.002	.009	.006	.026
Mining	.006	.009	.001	.003	.011	.014
Construction	.0001	.067	.00003	.015	.0002	.106
Manufacturing	.517	.239	.349	.176	.666	.286
Transportation	.062	.085	.036	.054	.086	.108
Wholesale	.012	.052	.007	.034	.017	.065
Retail	.039	.139	.051	.152	.029	.129
FIRE	.007	.076	.011	.109	.003	.052
Business services	.004	.043	.004	.033	.005	.050
Personal services	.002	.023	.003	.034	.002	.014
Entertainment services	.002	.010	.001	.009	.002	.011
Professional services	.343	.238	.535	.373	.175	.138
N	637,718	10,830,247	298,677	4,631,357	339,041	6,198,890

NOTE.—Means are reported. Standard deviations are in parentheses.

Table 3
Regressions for Log Wages for SEDF and NWECD Workers

	SEDF Workers					NWECD Workers				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Female	−.328	−.316	−.317	−.238	−.240	−.375	−.362	−.364	−.230	−.234
	(.0004)	(.0003)	(.0003)	(.0004)	(.0004)	(.001)	(.001)	(.001)	(.001)	(.001)
Age		.063	.076	.049	.061		.060	.074	.046	.058
		(.0001)	(.0001)	(.0001)	(.0001)		(.0003)	(.0005)	(.0003)	(.0004)
Age^2/100		−.062	−.073	−.047	−.058		−.059	−.071	−.043	−.055
		(.0001)	(.0001)	(.0001)	(.0001)		(.0004)	(.001)	(.0004)	(.001)
Black		−.078	−.079	−.034	−.037		−.091	−.094	−.051	−.053
		(.001)	(.001)	(.001)	(.001)		(.002)	(.002)	(.002)	(.002)
Currently married		.097	.086	.065	.057		.076	.066	.049	.042
		(.0003)	(.003)	(.0003)	(.0003)		(.001)	(.001)	(.001)	(.001)
High school degree		.186	.186	.114	.116		.163	.162	.107	.108
		(.0005)	(.0005)	(.0004)	(.0004)		(.002)	(.002)	(.002)	(.002)
Some college		.297	.291	.172	.170		.264	.260	.157	.156
		(.001)	(.001)	(.001)	(.0005)		(.002)	(.002)	(.002)	(.002)
Associate's degree		.377	.370	.207	.203		.410	.405	.209	.206
		(.001)	(.001)	(.001)	(.001)		(.002)	(.002)	(.002)	(.002)
Bachelor's degree		.585	.576	.360	.355		.555	.548	.342	.338
		(.001)	(.001)	(.001)	(.001)		(.002)	(.002)	(.002)	(.002)
Advanced degree		.762	.754	.520	.514		.681	.674	.498	.495
		(.001)	(.001)	(.001)	(.001)		(.002)	(.002)	(.003)	(.003)
Metropolitan statistical area		.215	.214	.175	.175		.162	.160	.125	.124
		(.0003)	(.0003)	(.0003)	(.0003)		(.001)	(.001)	(.001)	(.001)
Female × age			−.026		−.024			−.027		−.024
			(.0002)		(.0002)			(.001)		(.001)
Female × (age^2/100)			.022		.021			.023		.022
			(.0002)		(.0002)			(.001)		(.001)
Industry and occupation controls	No	No	No	Yes	Yes	No	No	No	Yes	Yes
R^2	.073	.356	.363	.466	.470	.121	.400	.407	.533	.537

Note.—Standard errors of regression estimates are reported in parentheses. The industry and occupation controls include dummy variables for the full set of three-digit census industry and occupation codes. There are 10,830,247 observations in the SEDF regressions and 637,718 observations in the NWECD regressions. In columns 3, 5, 8, and 10, the interactions of the female dummy variable and the age variables are created using the age variables minus their sample means, so the estimated coefficient of the female dummy variable measures the sex difference evaluated at the sample means of the control variables.

(23.8% in the SEDF vs. 23.0% in the NWECD). The same is true in the specifications that include the female-age interactions. It makes sense that the estimates would match better after controlling for industry and occupation because of the overrepresentation of some industries and occupations and the underrepresentation of others in the NWECD. Looking across the columns of table 3, there are some other minor differences between the two data sets, but, for the most part, the wage regression results from the NWECD come close to replicating those from the SEDF, particularly once controls for industry and occupation are added. Note that, in the empirical work below, we always include some sort of controls for industry and occupation since one of our interests is in the effects of industry and occupation segregation on male-female wage differences.

Therefore, while the NWECD data are not representative of the underlying population of U.S. workers, this data set represents a substantial improvement over existing data sources used to study the role of sex segregation along a number of dimensions in the workplace. The NWECD covers essentially the entire array of industries, occupations, locations, and so forth, in the U.S. economy. Moreover, wage regression estimates from the NWECD do not differ substantively from those obtained from a representative sample of the U.S. population of workers. Nonetheless, it remains to future work to attempt to construct even more representative samples of matched employee-employer data.

III. Methods

In our initial empirical work, we assume that the wage gap between men and women is a function of individual human capital characteristics and characteristics of the "femaleness" of where a worker works, as represented by the percent female in a worker's occupation, industry, establishment, and occupation within an establishment (job cell). That is, we estimate wage regressions of the following form:

$$\begin{aligned} w_{\text{poiej}} = {} & \alpha + \beta F_p + \gamma \text{OCC\%F}_o + \delta \text{IND\%F}_i \\ & + \lambda \text{EST\%F}_e + \theta \text{JOB\%F}_j + X_{\text{poiej}}\Phi + \epsilon_{\text{poiej}}, \end{aligned} \tag{1}$$

where w is the log hourly wage, F is a dummy variable equal to one if individual p is female, OCC%F is the percent female in occupation o, IND%F is the percent female in industry i, EST%F is the percent female in establishment e, and JOB%F is the percent female in job cell j. A vector of control variables is represented by X.

With the estimated coefficients of equation (1) in hand, we can construct a wage decomposition expressing the difference in average log wages between women and men as follows:

$$\begin{aligned} w_f - w_m = {} & \beta' + \gamma'(\mathrm{OCC\%F}_f - \mathrm{OCC\%F}_m) \\ & + \delta'(\mathrm{IND\%F}_f - \mathrm{IND\%F}_m) + \lambda'(\mathrm{EST\%F}_f - \mathrm{EST\%F}_m) \\ & + \Theta'(\mathrm{JOB\%F}_f - \mathrm{JOB\%F}_m) + (X_f - X_m)\Phi', \end{aligned} \tag{2}$$

where the primes on the coefficients indicate the estimates, and the f and m subscripts on the variables indicate the means for women and men, respectively. This decomposition gives the proportion of the wage gap that is due to the segregation of women into particular (generally lower-wage) occupations ($\gamma'(\mathrm{OCC\%F}_f - \mathrm{OCC\%F}_m)$), industries ($\delta'(\mathrm{IND\%F}_f - \mathrm{IND\%F}_m)$), establishments ($\lambda'(\mathrm{EST\%F}_f - \mathrm{EST\%F}_m)$), and job cells ($\Theta'(\mathrm{JOB\%F}_f - \mathrm{JOB\%F}_m$); differences in other observable characteristics ($(X_f - X_m)\Phi'$); and, most important, sex differences in wages controlling for segregation along all four dimensions (and therefore implicitly within job cells), as well as these other characteristics, captured in β'. These decompositions can therefore be thought of as traditional Oaxaca (1973) decompositions, imposing the restriction that the coefficients are the same for men and women. We present most of our results imposing this restriction, but, as we discuss below, we also repeated the basic analysis using the unrestricted decomposition, and we did not find qualitative differences in the results.

While establishments are well defined, industry and occupation can be defined at a variety of levels of disaggregation. Since the question of primary concern is within- versus across-job wage differences, we are interested in trying to use narrow occupational classifications. If, however, we use highly disaggregated occupations, we may end up with very small job cells (establishment-occupation cells), particularly since we only have a sample of workers in each plant, which may cause measurement error problems. Consequently, we report evidence from specifications using a variety of levels of occupational disaggregation, beginning with 13 broad census occupations and then using successive levels of disaggregation of occupations used by the Census Bureau, down to the finest level of disaggregation into 501 occupations (of which 491 are represented in our data). Each detailed census occupation code corresponds generally to a mix of three-digit and four-digit Standard Occupation Classification (SOC) codes, often combining two or three four-digit occupations into a census occupation.[5] To preview the results, we find that, while using different levels of occupational disaggregation does change the quantitative results, the qualitative conclusion is not strongly affected by the level of occupational detail that we use.

The percent-female variables in equation (1) are all estimated directly from the data. The percentages female in the occupation and industry are estimated from the full SEDF sample, so measurement error is likely to be minimal. However, the percentages female in the plant and job cell are estimated by necessity from the matched data in the NWECD. On average, 19.37 workers are matched to a plant, so job-cell estimates, in particular, are often based on a small number of observations. In order to eliminate potential measurement error, we also report results in which we estimate the coefficient on the female dummy variable, β, controlling for fixed occupation, industry, establishment, and job-cell effects, rather than controlling for the percent female in each of these categories; this amounts, of course, to putting in job-cell dummy variables, since these absorb occupation, industry, and establishment effects. In the absence of bias stemming from measurement error in the percent-female variables, we would not expect estimates of β obtained using these fixed effects to differ much from estimates using the percent-female variables if the percent-female variables do a reasonable job of characterizing how wages are affected by the sorting of workers into different industries, occupations, establishments, and job cells. Using job-cell dummies, however, avoids the measurement error inherent in the percent-female variables, and therefore this should provide more reliable estimates of the within-job-cell sex difference in wages (β).[6] Nonetheless, most of the results we report use

[5] For an example of what this occupational disaggregation entails, consider one of our 13 census occupation codes, Technicians and Related Support Occupations. At the level of 72 total occupations, this category constitutes three separate occupations: (1) Health Technologists and Technicians; (2) Technologists and Technicians, Except Health Engineering and Related Technologists and Technicians; Science Technicians; and (3) Technicians, Except Health, Engineering, and Science. At the level of 491 total occupations, these three categories are further disaggregated into 22 distinct occupations, including such occupations as dental hygienists, survey and mapping technicians, and legal assistants. Because we do not look at establishment-industry cells (since all workers in an establishment are presumably in the same industry, and they are so by construction in our data set), we face no constraint in disaggregating industries finely, and, hence, we always use the detailed census industry codes.

[6] We could, in principle, implement a formal correction for the measurement error bias that results from the sampling error in this case, as explained in Cockburn and Griliches (1987) and Mairesse and Greenan (1999). Cockburn and Griliches find that, when they construct a consistent estimate of the measurement error covariance matrix, allowing the error variances to differ across observations (as is necessary in this case, because the cell sizes differ), the resulting error-corrected covariance matrix is near-singular. Mairesse and Greenan are able to successfully invert their error-

the percent-female variables, both to follow some of the earlier literature and because the estimated effects of the percent-female variables are of interest in their own right—for example, in inferring the potential effects of a policy of comparable worth (Johnson and Solon 1986).

IV. Results Using the NWECD

A. Basic Analysis

Table 4 begins by reporting the results of wage regression estimations using 13 broad occupation categories. The first panel (A) reports results with no control variables. We report results from this simple specification for two reasons. First, it allows us to focus on the effects of segregation and to see how much of the sex gap in wages can be eliminated by controlling solely for measures of segregation. Second, no information on other characteristics of workers is available in the IWS, so this specification provides the closest comparison. However, to better contrast these estimates with standard wage regression estimates in other studies, the second panel (B) reports results adding the same basic set of human capital and other control variables used in the previous table. Finally, to provide a comparison with most other studies of sex segregation—in which data only on the percent female in the industry and occupation are available (e.g., Johnson and Solon 1986; Sorensen 1989; Fields and Wolff 1995; Macpherson and Hirsch 1995)—the third (C) and fourth (D) panels report results using only these segregation measures.

Beginning with panel A, column 1 simply reports the estimate of the raw wage gap from a regression of log hourly wages on the female dummy variable; the raw gap in these data is −.375.[7] Column 2 then reports wage regression estimates introducing the four percent-female variables; we divide the percent-female variables by 100, and therefore in the tables and accompanying discussion refer instead to the proportion female. Controlling for segregation by industry, occupation, establishment, and job cell, the sex gap in wages falls by about one-third, to −.244. Wages are lower in establishments with a higher proportion female, and within establishments they are lower in occupations with a higher proportion female (the job-cell effect). In this specification without other controls (notably education), occupational and industry segregation have the opposite of the usual effects, with wages higher in occupations and industries with a higher proportion female. However, as panel B shows, this result is reversed when individual-level controls are added. In addition, as just noted, most studies of sex segregation do not control for segregation at the level of the establishment and job cell.

corrected covariance matrix, but their model contains many fewer covariates that are measured with error.

[7] All of our coefficient estimates are highly significant, so while standard errors are reported, we do not continually discuss their statistical significance.

Table 4
Estimated Log Wage Differentials by Sex and Proportion Female in Occupation, Industry, Establishment, and Job Cell

	Coefficient Estimate (1)	Coefficient Estimate (2)	Mean Difference, Women − Men (3)	Absolute Contribution to Wage Gap, (2) × (3) (4)	Relative Contribution to Wage Gap (5)
A. Full decomposition, with no controls:					
Female	−.375	−.244	1.00	−.244	.651
	(.001)	(.002)			
Proportion female in occupation		.180	.180	.032	−.087
		(.013)			
Proportion female in industry		.122	.248	.030	−.081
		(.026)			
Proportion female in establishment		−.188	.338	−.064	.170
		(.019)			
Proportion female in job cell		−.243	.536	−.130	.347
		(.008)			
R^2	.121	.140			
B. Full decomposition, with basic controls:					
Female	−.375	−.193	1.00	−.193	.514
	(.001)	(.002)			
Proportion female in occupation		−.103	.180	−.019	.050
		(.006)			
Proportion female in industry		−.171	.248	−.043	.113
		(.018)			
Proportion female in establishment		−.173	.338	−.059	.156
		(.014)			
Proportion female in job cell		−.098	.536	−.053	.141
		(.004)			
R^2	.121	.432			
C. Limited decomposition, with no controls:					
Female	−.375	−.341	1.00	−.341	.910
	(.001)	(.002)			
Proportion female in occupation		.086	.180	.016	−.041
		(.003)			
Proportion female in industry		−.198	.248	−.049	.131
		(.003)			
R^2	.121	.126			
D. Limited decomposition, with basic controls:					
Female	−.375	−.241	1.00	−.241	.643
	(.001)	(.002)			
Proportion female in occupation		−.143	.180	−.026	.069
		(.005)			
Proportion female in industry		−.395	.248	−.098	.261
		(.012)			
R^2	.121	.427			

NOTE.—The sample size is 637,718. Standard errors of regression estimates are reported in parentheses; all standard errors are adjusted for nonindependence of residuals within establishments. In this table, 13 occupational categories are used. In panels B and D, the control variables listed in table 3, cols. 2 and 7, are included.

Thus, these studies no doubt overstate the role of occupational and/or industry segregation per se. This is demonstrated in panel D; when the establishment and job-cell segregation variables are dropped and the basic controls are included, the negative effects of occupation and especially industry segregation are stronger.[8]

The decomposition of the sex gap in wages requires not only the regression coefficients but also the mean differences between women and men of the right-hand-side variables, which are reported in column 3 of each panel. Women, of course, are in occupations, industries, establishments, and job cells with a higher proportion female. Columns 4 and 5 present the decomposition results. Column 4 reports the absolute contribution of each variable to the wage gap, and column 5 reports the relative contribution. In panel A, the estimates in column 5 indicate that nearly two-thirds (65.1%) of the wage gap is attributable to sex differences in wages that remain after accounting for segregation by occupation, industry, establishment, and job cell. Just over one-third (34.7%) is due to segregation into lower-paying occupations within establishments. Upon including the basic controls, in panel B, the estimated coefficient of the female dummy variable declines by about 20% in absolute value (from −.244 to −.193), while the estimated coefficient of the proportion female in the job cell declines more sharply (from −.243 to −.098). In terms of the decomposition, after accounting for the effects of sex segregation by occupation, industry, establishment, and job cell, the sex difference in wages remains large, contributing approximately one-half of the sex gap (51.4%). The fact that this figure is smaller than in panel A, without the basic controls, implies some differences in observables between men and women, conditioning on the segregation measures; this raises the possibility, which we can, of course, not address directly, that some unobservable differences remain. The contribution of establishment segregation remains about the same, while the contribution of segregation within jobs within establishments falls by over half (to 14.1%). Finally, panels C and D indicate that controlling for establishment and job cell segregation is

[8] Macpherson and Hirsch (1995) caution that the percent female in the worker's occupation may be a proxy for other job-related characteristics, so that the estimated negative effect on wages may partially reflect compensating differentials based on workers' preferences and perhaps also different skill requirements. Their evidence is consistent with this, as the longitudinal estimate of the effect of percent female in the occupation is much smaller than the cross-sectional estimate. Sorensen (1989) presents similar evidence for women only, based on a comparison of OLS estimates with estimates that account for selectivity into employment and into female-dominated occupations (although one can raise questions regarding identification of this model). In both papers, despite the evidence of bias, occupational segregation continues to lower wages.

important, as the share of the sex wage gap accounted for by an individual's sex is higher when segregation along these dimensions is ignored.

Our basic results therefore suggest that, while segregation does explain a substantial fraction of the sex gap in wages, a large proportion of the sex wage gap is still attributable to the sex of the worker. We now turn to evaluating the robustness of these results.

B. The Effects of the Degree of Occupational Disaggregation

The results in table 4 are based on 13 highly aggregated occupations. Because sex segregation (reflected in the mean difference in the proportion female in women's vs. men's occupations) is likely to be more severe at a more detailed occupational level, the decomposition results may be sensitive to the level of occupational aggregation used. To explore this question, in columns 1–3 of table 5 we report results for increasing degrees of occupational disaggregation. From this point on, we report the results from specifications including the control variables so as to provide the most reliable estimates of the decomposition. Column 1 replicates the key results from table 4 (corresponding to panel B, cols. 2, 3, and 5). In the second column, we increase the number of occupational classifications to 72, which amounts to disaggregating each of the 13 original occupations into anywhere from two to 14 distinct occupations. In the third column, we disaggregate as much as our data allow and use the most detailed Census occupation codes.

The first two rows of the table show, as we would expect, that given the greater degree of sex segregation in more detailed occupations, the mean sex difference in the proportion female by occupation and job cell is larger in each successive column. (The figures for industry and establishment are unchanged, of course.) Turning to the wage regression estimates, the estimated coefficient of the female dummy variable, or the effect of an individual worker's sex, declines a bit as more detailed occupations are used, from −.193 in column 1 to −.151 in column 3, with the corresponding relative contribution to the wage gap falling from 51.4% to 40.2%. Nonetheless, a sizable sex wage gap persists. Among the segregation measures, the contributions of establishment and job-cell segregation are most affected by the level of occupational disaggregation. The percentage of the sex gap accounted for by establishment segregation rises from 15.6% to 17.7%, and the percentage accounted for by job-cell segregation rises from 14.1% to 23.9%.

C. Not-Elsewhere-Classified Occupations

It turns out that workers in the census are often assigned to not-elsewhere-classified (n.e.c.) occupations when there is not a detailed occupational classification in which it seems appropriate to classify a worker.

Table 5
Estimated Log Wage Differentials by Sex and Proportion Female in Occupation, Industry, Establishment, and Job Cell, Varying Degrees of Occupational Disaggregation

	All Occupations			Excluding n.e.c. Occupations
	(1)	(2)	(3)	(4)
Number of occupations	13	72	491	451
Mean differences, women − men:				
Proportion female in occupation	.180	.340	.396	.418
Proportion female in job cell	.536	.657	.744	.746
Coefficient estimates:				
Female	−.193	−.164	−.151	−.143
	(.002)	(.002)	(.002)	(.002)
Proportion female in occupation	−.103	−.011	−.041	−.042
	(.006)	(.005)	(.004)	(.005)
Proportion female in industry	−.171	−.187	−.171	−.138
	(.018)	(.018)	(.018)	(.018)
Proportion female in establishment	−.173	−.169	−.196	−.218
	(.014)	(.014)	(.014)	(.014)
Proportion female in job cell	−.098	−.142	−.120	−.116
	(.004)	(.003)	(.003)	(.004)
Age	.056	.056	.056	.056
	(.001)	(.001)	(.001)	(.001)
Age^2/100	−.054	−.054	−.054	−.055
	(.001)	(.001)	(.001)	(.001)
Black	−.084	−.083	−.083	−.080
	(.004)	(.004)	(.004)	(.004)
Currently married	.074	.074	.074	.070
	(.001)	(.001)	(.001)	(.002)
High school degree	.164	.161	.164	.155
	(.002)	(.002)	(.002)	(.002)
Some college	.281	.275	.276	.264
	(.003)	(.003)	(.003)	(.003)
Associate's degree	.460	.456	.458	.460
	(.005)	(.005)	(.005)	(.005)
Bachelor's degree	.604	.595	.596	.583
	(.005)	(.005)	(.005)	(.005)
Advanced degree	.769	.757	.758	.755
	(.006)	(.006)	(.006)	(.006)
Metropolitan statistical area	.150	.148	.148	.148
	(.005)	(.005)	(.005)	(.005)
Relative contribution to wage gap:				
Female	.514	.438	.402	.406
Proportion female in occupation	.050	.010	.043	.050
Proportion female in industry	.113	.124	.114	.098
Proportion female in establishment	.156	.153	.177	.210
Proportion female in job cell	.141	.249	.239	.245
N	637,718	637,718	637,718	567,965

NOTE.—Other than the level of occupational disaggregation, the specifications are the same as in panel B, table 4. Column 1 reproduces results from panel B, cols. 2, 3, and 5, of table 4. Standard errors of regression estimates are reported in parentheses; all standard errors are adjusted for nonindependence of residuals within establishments. n.e.c. refers to not elsewhere classified.

Such "residual" occupations are likely to be more heterogeneous and less sex segregated than standard occupations because they presumably lump together many jobs. If so and if these occupations are quantitatively important, the estimated coefficient of the dummy variable for females may overstate the within-job-cell wage gap because for n.e.c. occupations the female dummy variable may capture, to some extent at least, a wage gap across jobs within n.e.c. occupations.[9]

To explore this issue, column 4 of table 5 reports estimates after dropping workers in the n.e.c. occupations. As the table indicates, a fairly high proportion of workers are in these occupations (10.9%). In addition, estimated segregation—captured by the mean differences between women and men in the proportion female in the occupation or job cell—is higher once these workers are excluded. However, the differences are small. Finally, the wage equation estimates indicate a slightly lower effect of sex per se on the wage gap, with the estimated coefficient falling (in absolute value) from −.151 to −.143, although the relative contribution of sex to the wage gap actually rises slightly from 40.2% to 40.6%. Thus, the n.e.c. occupations do not in any way drive the results.[10]

D. The Effects of Other Forms of Disaggregation

The results reported thus far follow much of the literature in using a single wage regression for men and women, with a dummy variable for women. A more flexible decomposition procedure is to use separate wage regressions for men and women, following Oaxaca (1973). Using this less restrictive decomposition (regardless of whether we view the male or female wage structure as the "no-discrimination wage structure," as explained in Neumark [1988]), we found some differences in the estimated coefficients on the proportion-female variables between men and women. Although the apparent differences in the effects of segregation for men and women

[9] Examples of n.e.c. occupations include engineers, n.e.c. (within the occupational category of engineers, which includes aerospace, chemical, nuclear, etc.); social science teachers, n.e.c. (within the occupational category of postsecondary teachers, which includes psychology, economics, history, political science, and sociology teachers); and office machine operators, n.e.c. (within the occupational category of duplication, mail, and other office machine operators, which includes duplicating machine operators and mail preparing and paper handling machine operators).

[10] To assess the robustness of our basic results and those using different levels of occupational disaggregation, we also estimated these same specifications including controls for establishment size and whether the establishment belongs to a multi-unit firm. To the extent that these reflect establishment-level characteristics, they may "overcontrol" for establishment-level differences, because they may capture dimensions of sex segregation. On the other hand, these variables may be related to unobserved human capital, calling for their inclusion along with the other human capital controls. The estimates scarcely changed upon including these additional control variables.

are of independent interest and bear further exploration, for the purposes of this article, the important point is that using separate wage equations for men and women did not eliminate the significant role of an individual's sex in determining wages, conditional on all of the segregation measures. Compared with single-equation estimates, the estimated overall effect of an individual's sex was as large (in absolute value) using the female wage structure, and it was larger using the male wage structure.

Another potential disaggregation issue is the difference between urban and nonurban labor markets. For example, there may be some monopsony power in nonurban markets, and this may particularly influence women's wages, although the relative influences of monopsony power on the various components of the sex wage gap are not obvious. We reestimated the specifications separately for establishments located in MSAs and those located outside of MSAs. The results were qualitatively similar in the two subsamples. Using the most disaggregated occupations, an individual's sex accounted for 42.5% of the sex wage gap in the MSA subsample and 39.6% in the non-MSA subsample. In both subsamples, job-cell segregation contributed the second largest share, and establishment segregation contributed the third largest share.

Earlier, we noted that the sample of establishments in the NWECD is disproportionately weighted toward manufacturing. If the effects of sex segregation or an individual's sex are different in manufacturing and nonmanufacturing industries, then in order to obtain unbiased estimates of population parameters, one might want to weight the data to make them more representative. Of course, many other factors influence selection into the NWECD sample, some of which may be related to unobservable characteristics, making it unclear how the weights might be constructed. We explored the most salient nonrepresentativeness—the preponderance of manufacturing plants—by estimating the equations and decomposition separately for the manufacturing and nonmanufacturing sectors. There are some differences between the two sectors. In general, there is more segregation in the nonmanufacturing sector, except at the job-cell level; this may reflect the greater variety of industries making up this sector. There are some differences in the relative contributions to the wage gap of occupation, industry, establishment, and job-cell segregation, as well as sex per se. In particular, the contribution of job-cell segregation is larger in nonmanufacturing (25.5% vs. 15.8%), while the contribution of establishment segregation is larger in manufacturing (16.2% vs. 9.2%). Furthermore, the sex gap that remains conditional on the various segregation measures is a bit larger in nonmanufacturing than in manufacturing (42.1% vs. 35.2%). However, the estimates are relatively close, and the key finding that the individual's sex accounts for a large share of the wage gap persists in both sectors. Thus, the most pronounced source of nonrepresentativeness has little influence on the qualitative findings.

Table 6
Estimated Log Wage Differentials by Sex, with Fixed Occupation, Industry, Establishment, and Job Cell Effects

	(1)	(2)	(3)
A. No controls:			
Number of occupations	13	72	491
Estimated coefficient of female dummy variable	−.235 (.002)	−.196 (.003)	−.180 (.003)
Relative contribution to wage gap	.628	.524	.481
Estimate relative to specification using percent female variables	.96	.93	.91
B. Basic controls:			
Estimated coefficient of female dummy variable	−.205 (.001)	−.176 (.002)	−.162 (.002)
Relative contribution to wage gap	.548	.469	.433
Estimate relative to specification using percent female variables	1.06	1.07	1.07

E. The Effects of Eliminating Measurement Error

As discussed above, sampling error in the estimates of the percentage or proportion female in the establishment and job cell may be quite severe. One approach that eliminates any role of measurement error in the proportion female variables is to include, in place of the segregation measures, a full set of job-cell dummy variables that captures occupation, industry, establishment, and, of course, job-cell fixed effects. This specification has the added benefit that, unlike the wage decompositions used to this point, it does not impose a particular functional form on the way segregation affects wages, although it sacrifices estimates of the effects of sex segregation.

These estimates are reported in table 6. Panel A reports results excluding the other controls, and panel B reports results including them. In panel A, using the job-cell dummy variables, the estimated effect of the individual's sex is only slightly smaller than in table 5 at each level of disaggregation of occupations.[11] The last row of the panel reports the ratio of the estimated coefficient of the female dummy variable using fixed effects to that using the segregation variables. While below one, the estimates range from .96 to .91, decreasing as we disaggregate occupations further. Because the female dummy is positively correlated with the mismeasured proportion female in establishment and job-cell variables, attenuation bias in these latter coefficients would tend to bias the estimated coefficient of the female dummy variable away from zero. This bias should be larger the more severe the measurement error is, which is consistent with a smaller ratio using more disaggregated occupations. Of course, as more control variables are added, any such predictions of the effects of

[11] The results were similar excluding the n.e.c. occupations.

measurement error become less definitive. Thus, in panel B, in which the specifications include the other control variables, the estimated coefficients of the dummy variable for females are actually a shade larger than the corresponding estimates in table 5.

The key finding, however, is that a large share of the wage gap remains within job cells (or, alternatively, attributable to an individual's sex).[12] The finding in the earlier tables was not a spurious result stemming from mismeasurement of the segregation variables.[13] Note also that, in this table, adding the basic controls has less of an impact on the relative contribution of the within-job-cell sex wage gap to the overall gap than was the case using the segregation measures. This is natural, as the very large set of dummy variables for each job cell is likely to capture far more heterogeneity than the limited set of direct segregation measures. Finally, the results in table 6 also indicate that the finding of a sizable within-job-cell sex difference in table 5 is not attributable to the functional form used to estimate the impact of segregation on wages.

One potential objection to the estimates with fixed job-cell effects is that different individuals identify the coefficient of the female dummy variable than in the specifications using the segregation measures. In particular, in the former, only women and men in integrated job cells contribute identifying information; clearly, if all women and men were in completely sex-segregated job cells, the coefficient on the female dummy variable would be unidentified. However, we reestimated the specification using the segregation measures but including only individuals in integrated job cells. The key result is unchanged; for example, for the specification corresponding to column 3 of table 5 (using the most disaggregated oc-

[12] When job-cell dummy variables are included in the regression specifications, the estimated effects of the individual's sex in these specifications are, literally, within-job-cell sex differences in wages. Up to this point, we have not used this label in describing the effect of an individual's sex; however, because the results from this specification are quite similar to those using the proportion-female variables, from this point on we use the more transparent "within-job-cell" expression.

[13] We also used this specification to verify that the relationship we estimate between wages and sex is driven by the rate of pay. Because our wage variable is a constructed wage, it is possible that it is not rates of pay that differ by sex within job cells but, rather, weeks or hours. For women to have lower constructed wages within job cells, however, it would have to be the case that their weeks or hours were higher within job cells, which seems unlikely (although our "stylized facts" do not refer to within-job-cell differences). To check this, we estimated specifications for log weeks and for log hours, including a dummy variable for females and job-cell fixed effects. For the different levels of occupational disaggregation, the estimated coefficient on the female dummy variable in the weeks regression ranged from $-.005$ to $-.007$, while in the hours regression it ranged from $-.030$ to $-.044$. These negative coefficients imply that dividing through by weeks and hours tends to make constructed wages look, if anything, more equal by sex within job cell.

cupations), the within-job-cell sex gap in wages remains large (−.147 as compared with −.151 in table 5).

F. Control Variables and Results for Different Types of Women and Men

Finally, the sex wage gap may vary with measured human capital characteristics and other controls, even within job cells. To examine how this affects the conclusions, table 7 reports results from specifications with fixed job cell effects where we augment the set of control variables to allow the within-job-cell sex wage gap to differ by race, age, and marital/childbearing status, by interacting the sex dummy variable with these controls.[14] We include each of these interactions in separate regressions because including a full set of interactions between sex and the other variables in one regression makes it difficult to evaluate the results.

The results in panel A of table 7 come from a regression where we include interactions between race and sex, and these indicate that there is a significant within-job-cell sex gap in wages for both blacks and nonblacks. Consistent with the findings in other studies, though, the sex gap in wages is smaller for blacks than for nonblacks (see, e.g., Bayard et al. 1999). The results in panel B indicate that there is a sex gap in wages that is significant at all ages, with a low of 8.1% for the youngest age category that rises monotonically to a high of 22.1% for the oldest age category. In panel C, we report results from a regression where we allow the effects of an individual's sex to differ by marital status and past childbearing. In this specification, in which we are trying to capture the effects of marriage and childbearing, we also include a dummy variable measuring whether a woman has ever had children.[15] We do this because the Decennial Census does not have any direct information on experience or tenure, and for women past childbearing is negatively associated with wages, especially in the absence of experience or tenure controls (Korenman and Neumark 1992). The within-job-cell sex wage gap for single, childless women is 9.2%, and the gap is much larger for married women, whether or not they have children. Finally, panels D and E disaggregate the results by region and industry. While the within-job-cell sex wage gaps are quite stable across regions, they vary considerably across industries, although all are positive and sizable.

A fundamental question that arises with regard to interpreting within-

[14] Note that we have substituted age dummy variables for the linear and quadratic terms.

[15] This variable is available in the SEDF only for women. One can, of course, measure whether there are currently children in the household for men and women, but this variable does not capture the effects of past childbearing and child rearing. Using the variable only available for women implies that we simply restrict the effect on men's wages of children to be zero.

Table 7
Estimated Log Wage Differentials by Sex, with Fixed Occupation, Industry, and Job Cell Effects, Interactive Specifications, Most Disaggregated Occupations Only

	Coefficient	SE
A. Race:		
Black	−.121	.007
Nonblack	−.164	.003
B. Age (years):		
≤ 25	−.081	.007
26–35	−.123	.004
36–45	−.174	.004
46–55	−.217	.005
≥ 56	−.221	.007
C. Marriage/children:		
Single, no children	−.092	.005
Married, no children	−.162	.004
Married, with children	−.198	.003
D. Region:		
Northeast	−.169	.007
North Central	−.160	.005
South	−.157	.005
West	−.161	.008
E. Industry:		
Agriculture	−.132	.062
Mining	−.158	.047
Construction	−.143	.006
Manufacturing	−.169	.004
TCU	−.181	.013
Wholesale	−.234	.048
Retail	−.249	.016
FIRE	−.344	.050
Services	−.142	.005

NOTE.—The specifications include the basic controls substituting age group dummy variables for the linear and quadratic age variables. For each panel, a separate specification is estimated that also includes the female dummy variable interacted with the listed variables, and it is those interactions that are reported; thus, the table reports the within-job-cell sex wage gap for each of the indicated categories.

job-cell sex wage gaps is whether they could be due solely to unmeasured human capital differences. We read the evidence in table 7 as providing some support for a role for human capital in determining sex wage gaps. For example, the widening of the sex gap in wages with age is consistent with a cumulative widening of the human capital gap between men and women over time. This widening gap stems from the smaller positive impact of age on women's wages as compared with men's wages (see table 3), which could be interpreted as lower human capital investment or as reflecting more intermittent accumulation of experience among women. Similarly, the smaller gap for single women with no children as compared with married women with children is consistent with models of human capital investment based on household specialization, and the gap between single and married childless women could reflect planned or expected future labor market interruptions for the latter, as in the Polachek model (1975). Nonetheless, all of the results in table 7 still indicate that a statistically significant sex gap in wages exists even within job cells and within categories of workers defined by race, age, and marital/childbearing status. While we cannot rule out a human capital explanation for the remaining sex gap in wages, it is important to emphasize again that this gap exists after we have controlled for a potentially huge set of job-related characteristics, presumably including skill requirements, with the inclusion of the job cell dummy variables. Indeed, we may want to interpret the within-job-cell sex differences for the youngest women, or single childless women, as lower-bound estimates of within-job wage discrimination, estimates that are on the order of 8%–9%.

V. Comparison with Groshen's Estimates

The findings from the NWECD indicating that occupational segregation is quantitatively unimportant and that within-job-cell sex differences in wages contribute a large (in fact, the largest) share of the sex wage gap contrast sharply with Groshen's (1991) findings. Using IWS data from the 1970s and 1980s on five specific industries (Miscellaneous Plastic Products, Nonelectrical Machinery, Life Insurance, Banking, and Computer and Data Processing), Groshen reports that within-job-cell sex differences account for only −1.0% to 6.6% of the wage gap, while the effect of job-cell segregation ranges from explaining −2.7% to 32.5% of the wage gap, with the percentage above 20% for three of the five industries. The proportion female in the occupation accounts for the largest share, ranging from 40.6% to 74.8%. The estimated contributions of job-cell segregation are not very different from ours, but the estimated contributions of occupational segregation and within-job-cell sex differences contrast strongly.[16]

There is a natural explanation for the difference in the estimated role of within-job-cell sex wage gaps. In particular, the IWS data may indicate a much smaller role for within-job-cell sex differences in wages because occupation classifications in the IWS are much more narrowly defined than Census occupation codes, and they are even industry specific. As an example, in the Miscellaneous Plastics Products industry, there are separate occupation codes for "Compression-Molding Machine Operators," "Extrusion Press Operators," "Injection-Molding Machine Operators," "Preform-Machine Operators," and "Vacuum-Plastics-Forming Machine Op-

[16] Since the IWS data do not contain the basic control variables we used, comparisons with the NWECD estimates should be based on the estimates without these controls (in panel A of table 4); these controls are also excluded in the NWECD estimations presented in this section.

erators," all of which get aggregated into a single three-digit Census occupation (719, "Molding and Casting Machine Operators"). If it is only at the level of disaggregation of the IWS occupations that the within-job-cell sex wage gap disappears, then the results using the NWECD and IWS might coincide if they were based on the same level of occupational disaggregation. In this section, we explore this possibility by aggregating IWS occupation categories into Census occupations (at the most detailed level of Census occupation codes), using in-house BLS documentation.[17] We then perform the decompositions using the IWS data based on these broader occupations and compare the results to those in the NWECD.

To carry out this exercise, we obtained from the BLS the original IWS data that Groshen studied. To establish a baseline, we first verified that we could replicate Groshen's results using the IWS data. Having done this, to draw a sharper comparison, we drew subsamples in our NWECD data set for the five IWS industries, and we restricted our analysis to three of the five industries (Miscellaneous Plastics Products, Nonelectrical Machinery, and Banking) for which we have reasonable-sized samples in the NWECD.[18] We also used the original IWS documentation to determine which classes of occupations the IWS covered, and we further restricted our NWECD sample to workers in these occupations. Even restricting the NWECD data to the industries and occupations covered by Groshen's IWS analysis, the differences between the two data sets remain sharp. In the two manufacturing industries, 41%–49% of the wage gap is due to occupational segregation in the IWS, as compared with 6%–17% in the corresponding NWECD data. In Banking, the IWS estimates indicate an even larger role for occupational segregation, contributing 71% of the sex gap in wages, as compared with 36% in the NWECD data. And, for two of the three industries, the NWECD results replicate the finding from the full NWECD that the sex gap within job cells accounts for a large share of the wage gap—41.4% in Nonelectrical Machinery and 43.7% of this gap in Banking, based on the NWECD, as compared with −1% and 2.4%, respectively, in the IWS data. For Plastics, the estimated contributions are closer, but the contribution is still larger by a factor of three in the NWECD data (13.5% vs. 4.7%). Thus, the NWECD data assign a much less prominent role to occupational segregation and a much more prominent role to sex wage gaps within establishments and occupations.

Having established the differences in the results using comparable industries and occupations in the two data sets, we next turn to the evidence

[17] There is, of course, no way to do the reverse exercise, disaggregating the NWECD occupations to correspond to those in the IWS.

[18] There are 582 workers working in 105 establishments in Plastics, 3,220 workers in 191 establishments in Nonelectrical Machinery, and 1,830 workers in 390 establishments in Banking. The NWECD samples sizes in Computer and Data Processing and Life Insurance are much smaller.

on the role of occupational disaggregation. The results are reported in table 8. In columns 1 and 2, we report (for comparison purposes) the estimated coefficients and relative contributions of each component of the decomposition to the wage gap in the IWS using IWS occupation definitions. In columns 3 and 4, we report the same set of results but use the IWS occupations aggregated up to the same census occupations that we use in the NWECD. Finally, in columns 5 and 6, we report estimated coefficients and relative contributions using the NWECD data for workers in these occupations in each of the three IWS industries.

The results for Plastics are given in panel A. The coefficient on the female dummy using IWS data aggregated into census occupations is −.033, as reported in column 3. This is slightly larger than the estimated coefficient of −.011 in column 1, based on IWS occupation classifications, and it raises the relative contribution of the female dummy to the total sex wage gap from 4.7% to 13.8%, as reported in columns 2 and 4. The estimated coefficient on the female dummy variable for the NWECD is −.026, which is similar to the IWS result, as is the estimated relative contribution of an individual's sex (13.5%, which closely matches the aggregated IWS estimate of 13.8%). This is a relative contribution, however, so it is partially driven by the fact that, in the NWECD, the proportion female in the establishment contributes virtually nothing to the wage gap. But while the results for Plastics do suggest that aggregation has an effect on relative wage gaps, the effect of an individual's sex in the NWECD in Plastics is a small fraction of what it is in the other two industries.

The IWS and NWECD comparisons for Nonelectrical Machinery appear in panel B. Comparing columns 1 and 3, we see that aggregation changes the coefficient on the female dummy variable in the IWS from 0.003 to −.022. This, of course, changes both the sign and the magnitude of the relative contribution of an individual's sex, from −1.0% to 7.3%. In column 5, however, the estimated coefficient on the female dummy variable in the NWECD sample is −.123, much larger than that in the IWS for the same level of occupational disaggregation (−0.022). The relative contribution of an individual's sex in the NWECD is also much larger, 41.4%, as reported in column 6. So while aggregation does change the estimated contribution of an individual's sex in the IWS, its contribution is still much smaller than we find in the NWECD.

The results in Banking, reported in panel C, are equally stark. The coefficient on the female dummy variable in Banking in the IWS rises (in absolute value) from −.009 to −.026 with occupational aggregation, which raises the relative contribution from 2.4% to 7.0%. However, the coefficient on the female dummy variable in the NWECD is −.301, and the relative contribution of sex is 43.7%. So, as in Nonelectrical Machinery, aggregating IWS occupations into census occupations in Banking does have an effect on the estimated contribution of an individual's sex to the

Table 8
Effects of Aggregation in IWS on IWS versus NWECD Comparison

	IWS Data, IWS Occupations		IWS Data, Census Occupations		NWECD	
	Coefficient Estimate (1)	Relative Contribution to Wage Gap (2)	Coefficient Estimate (3)	Relative Contribution to Wage Gap (4)	Coefficient Estimate (5)	Relative Contribution to Wage Gap (6)
A. Miscellaneous plastics products:						
Female	−.011	.047	−.033	.138	−.026	.135
	(.003)		(.003)		(.053)	
Proportion female in occupation	−.245	.497	−.172	.265	−.186	.172
	(.004)		(.004)		(.114)	
Proportion female in establishment	−.117	.156	−.122	.162	−.015	.018
	(.004)		(.004)		(.113)	
Proportion female in job cell	−.093	.300	−.141	.435	−.159	.675
	(.005)		(.005)		(.070)	
R^2	.332		.308		.079	
N	70,355		70,355		582	
B. Nonelectrical machinery:						
Female	.003	−.010	−.022	.073	−.123	.414
	(.004)		(.004)		(.033)	
Proportion female in occupation	−.452	.406	−.150	.090	−.104	.063
	(.006)		(.007)		(.075)	
Proportion female in establishment	−.331	.479	−.318	.460	−.291	.265
	(.007)		(.008)		(.133)	
Proportion female in job cell	−.057	.124	−.191	.377	−.147	.259
	(.007)		(.008)		(.045)	
R^2	.358		.299		.098	
N	54,873		54,873		3,220	
C. Banking:						
Female	−.009	.024	−.026	.070	−.301	.437
	(.004)		(.003)		(.053)	
Proportion female in occupation	−.685	.711	−.603	.608	−1.01	.363
	(.008)		(.009)		(.077)	
Proportion female in establishment	−.385	.048	−.376	.047	−.072	.019
	(.012)		(.013)		(.084)	
Proportion female in job cell	−.160	.217	−.216	.274	−.184	.182
	(.008)		(.009)		(.071)	
R^2	.401		.387		.404	
N	74,501		74,501		1,830	

NOTE.—Standard errors of estimates are reported in parentheses. Following Groshen, in the estimates in this table we do not correct standard errors for correlation across individuals in an establishment. However, the corrected standard errors on the female dummy variable are virtually identical to the uncorrected standard errors.

wage gap, but the effect is small and is nowhere near large enough to explain the discrepancies between the results for the IWS and NWECD.

To summarize, in each of the three industries, aggregating up from detailed IWS occupations to census occupations does increase the relative contribution of the female dummy variable to the overall sex wage gap.[19] But the qualitative conclusions one can draw from the analysis of the IWS are not affected by the aggregation of occupations. In the IWS, an individual's sex accounts for a relatively small portion of the overall wage gap, even when occupations are aggregated up to Census occupation codes. Moreover, in two of the three industries studied (Nonelectrical Machinery and Banking), the IWS results are markedly different from the NWECD results, even when the level of occupational classification used is identical. Thus, these results for the IWS show that differing levels of detail in occupational classifications cannot explain the discrepancies between results from the sample of NWECD workers in the three IWS industries, and results using the IWS data.

Additional exploration of differences between these two quite different data sources raised some caution flags regarding the IWS. Because the IWS for Banking was conducted in 1980, it can be compared with data from the 1980 Decennial Census. We extracted data from the 5% Public Use Microsample (PUMS) for all workers reporting that they worked in the Banking industry in 1979 in occupations represented in the IWS. Using similar individual-level sample restrictions to those in the NWECD, we obtained a sample of 37,710 workers in Banking in 1980. The mean wage of banking workers in the PUMS is $5.47 per hour, which is similar to the estimate of $5.60 per hour that we obtained in our IWS sample, while the percent female in banking is 79.9% in the PUMS, again very similar to the 82.8% we obtained in the IWS sample. In contrast, the unadjusted sex wage differences in the PUMS and IWS are vastly different. In the IWS, the sex wage difference is −.372, while in the PUMS the sex wage difference is −.614, almost double that in the IWS. In contrast, the PUMS estimate is very similar to the −.689 sex wage difference in the NWECD. Of course, such a large discrepancy in unadjusted sex wage differences could lead to vastly different conclusions from a wage decomposition using the two data sets.

We can only speculate on why it is that the sex wage gap in Banking in 1980 is so much smaller in the IWS than in the 1980 PUMS. The IWS in Banking is likely to contain a nonrepresentative sample of banks, given that it was only conducted in 29 large metropolitan areas, although it does not appear that these banks were nonrepresentative in terms of the

[19] As expected, this aggregation reduces the estimated contribution of occupational segregation to the wage gap and increases the estimated contribution of job-cell segregation.

average wages they paid or in terms of the sex mix of their workers. Another possibility is that at least some banks were hesitant to provide data suggesting that within narrowly defined occupations there was a pay gap between men and women,[20] and therefore both the overall sex wage gaps and the within job-cell sex wage gaps calculated in the IWS are artificially low. Regardless, given that the data for one of the three IWS industries for which we have data from another source seem not to be comparable to a nationally representative sample of workers in that same year, we think one should be very cautious in drawing conclusions about the importance of segregation in determining sex wage differences using the IWS. In contrast, as we discussed above, the NWECD, while not entirely representative, comes much closer to matching the critical moments from the distributions of wages and worker characteristics in the U.S. population, and, of course, it provides broad industry coverage.[21]

VI. Conclusions

We assembled a large matched employer-employee data set covering essentially all industries and occupations across all regions of the United States. We use this data set to reexamine the question of the relative contributions to the overall sex gap in wages of sex segregation versus wage differences by sex within occupation, industry, establishment, and occupation-establishment cells. This is especially important given that earlier research on this topic relied on data sets that covered only a narrow range of industries, occupations, or regions.

Our results indicate that, although a sizable fraction of the sex gap in wages is accounted for by the segregation of women into lower-paying occupations, industries, establishments, and occupations within establishments, a substantial part of this gap remains attributable to the individual's sex. Overall, our estimates indicate that approximately one-half of the sex wage gap takes the form of wage differences between men and women within narrowly defined occupations within establishments. These findings contrast sharply with the conclusions of previous research (especially Groshen 1991) using more limited data, which indicated that sex segregation accounted for essentially all of the sex wage gap. While we do not attempt in this article to determine the underlying forces that cause men and women to have different wages within narrowly defined occupations in the same establishments, further research into the sources of within-

[20] The Equal Pay Act places the burden of proof on the employer to show that unequal pay for equal work is based on a factor other than sex, such as seniority.

[21] This is true not only for the NWECD in general but also specifically for Banking. For example, the sex wage difference for NWECD workers in Banking is −.689, whereas for all workers in Banking in the SEDF, it is −.617. This is a very small discrepancy relative to the difference between the unadjusted wage gap in the IWS in 1980 relative to that in the PUMS in 1980.

establishment, within-occupation sex wage differences is apparently much more important than previously thought.

The policy implications of our findings are very different from those drawn from the earlier research. Our results suggest that identifying and eliminating the sources of within-occupation, within-establishment wage differences between men and women can play a large role in reducing wage differences between the two genders. In particular, if, within the narrowly defined occupations that we study, the jobs performed by men and women require substantially equal skill, effort, responsibility, and working conditions yet wages differ by sex, then enforcement of the Equal Pay Act can play a fundamental role in closing the wage gap between men and women. In contrast, if segregation along various dimensions accounts for most of the sex wage gap, then policies along the lines of comparable worth, equal opportunities in employment and promotion, and affirmative action would be central to any further closing of this gap and stronger equal pay provisions would not be effective. Our findings suggest that stronger enforcement of equal pay legislation could further reduce the wage gap between men and women, perhaps substantially.

Appendix

Matching Employees to Employers Using Location and Industry Information

The Census Bureau organizes the United States into different geographic areas, assigning codes to each. For the NWECD, there are five areas of interest: state, county, place, tract, and block.[22] The geographic coding process works primarily as a hierarchy. The Census Bureau assigns unique codes to every state in the country. Within states, each county is also assigned a unique code. In addition, in areas or townships with a population of 2,500 or more, the Census Bureau assigns a place code. Because an area or town can cross county boundaries, we can distinguish between areas in the same place but different counties. Tract codes are unique within counties, and block codes are unique within tracts. The Census Bureau uses the same geographic codes in both the SSEL and the Decennial Census.[23]

[22] In some geographic areas, the Census Bureau uses Block Numbering Areas (BNAs) instead of tracts. For our purposes, a BNA is equivalent to a tract. The Census Bureau assigns tracts and blocks in tandem, so whenever an establishment is assigned a tract code, it is also always assigned a block code.

[23] Two shortcomings of the geographic codes in the SSEL are (1) the absence of tract and block codes before 1992 and (2) the incomplete assignment of these codes to all establishments. To assign tract and block codes to the 1990 SSEL, we extracted each establishment's block and tract code (when available) from the 1992 SSEL and then matched these codes back to the 1990 SSEL. Some establishments that ceased operation between 1990 and 1992 do not appear in the 1992 SSEL, making it impossible to identify block and tract codes for these establishments. Due to address problems, not all establishments had tract and block codes assigned as of 1992. In the 1992 SSEL, Census had assigned tract and block codes to 45% of all establishments.

In addition to geographic codes, the Census Bureau assigns industry codes to the SEDF and the SSEL. The Census Bureau asks long-form respondents to identify their employer's industry, which the Census Bureau codes into one of 236 Census Industry Classification (CIC) codes. Each CIC code corresponds roughly to a three-digit Standard Industrial Classification (SIC) code.[24] In the SSEL, the Census Bureau assigns each establishment a six-digit SIC code based on the plant's primary economic activity.[25] Since the CIC codes are more aggregated than SIC codes, we use a concordance table to assign a CIC to each SIC in the SSEL.[26]

The first step in the matching process is to assign all plants in the SSEL to industry-location cells. We divide the SSEL into plants that are unique in a state-county-place-industry (SCPI) cell, and those that are not, and retain all unique SCPI plants. In cases where there are multiple plants in an SCPI cell, we first retain the cell only if all plants in the cell have tract and block codes. We then keep only those plants that are unique within a state-county-place-tract-block-industry cell. Next, we assign workers in the SEDF to industry-location cells based on information provided in the SEDF. Unlike the SSEL, the Census Bureau assigns detailed geographic codes to all observations in the SEDF.[27] Once we have workers assigned to industry-location cells and have establishments that are unique within a cell, we can match the workers to the particular establishments where they work.

We take a number of additional steps to improve the quality of the match. First, to ensure that workers are matched properly to employers in the NWECD, we discard all workers and establishments from the matched sample where the census imputed either the worker's or the establishment's industry.[28] We also discard all workers from the matched sample if the worker's place-of-work code is imputed and this imputed code is the source of the match.[29] Second, some matches lead to apparent inconsistencies,

[24] An exception is Construction. There is one CIC for Construction, and this corresponds to the equivalent of three two-digit SIC codes.

[25] The last two digits of the SIC code are product codes for goods-producing industries, or type of business codes for service establishments.

[26] A few SICs correspond to more than one CIC. We omitted establishments in these industries.

[27] When long-form respondents omit geographic information, the Census Bureau imputes missing values.

[28] This imputation occurs for plants when an incomplete SIC code (only the first two or three digits) is provided. For such cases, the Census Bureau randomly assigns the remaining digits.

[29] To understand the exclusion based on imputed geographic data, consider the following example. When a worker is matched to an establishment unique in an SCPI cell, the match is based on the state, county, place, and industry of the worker and the establishment. If the worker's block code is imputed, then this imputed code has no bearing on the match, and we retain the match in the data. However, if the match relies on tract- and block-level information, and the worker's place-of-work block code is imputed, then we discard the worker from the matched data set.

prompting us to discard matches when the number of workers matched to an establishment exceeds the number of employed workers as reported by the establishment in the SSEL. Although there may be legitimate reasons for the number of matched workers to exceed reported establishment employment, to avoid potentially incorrect matches, we discard cases where this occurs.[30]

References

Bayard, Kimberly; Hellerstein, Judith; Neumark, David; and Troske, Kenneth. "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*, edited by John Haltiwanger, Julia Lane, James Spletzer, Jules Theeuwes, and Kenneth Troske, pp. 175–203. Amsterdam: Elsevier Science, 1999.

Bielby, William, and Baron, James. "A Woman's Place Is with Other Women: Sex Segregation within Organizations." In *Sex Segregation in the Workplace: Trends, Explanations, Remedies*, edited by Barbara Reskin, pp. 27–55. Washington, DC: National Academy Press, 1984.

Blau, Francine D. *Equal Pay in the Office*. Lexington, MA: Heath, 1977.

———. "Trends in the Well-Being of American Women, 1970–1995." *Journal of Economic Literature* 36 (March 1998): 112–65.

Brown, Charles, and Medoff, James L. "The Employer Size Wage Effect." *Journal of Political Economy* 97 (October 1989): 1027–59.

Carrington, William J., and Troske, Kenneth R. "Sex Segregation in U.S. Manufacturing." *Industrial and Labor Relations Review* 51 (April 1998): 445–64.

Cockburn, Ian, and Griliches, Zvi. "Industry Effects and Appropriability Measures in the Stock Market's Valuation of R&D and Patents." Working Paper no. 2465. Cambridge, MA: National Bureau of Economic Research, 1987.

Fields, Judith, and Wolff, Edward N. "Interindustry Wage Differentials and the Gender Wage Gap." *Industrial and Labor Relations Review* 49 (October 1995): 105–20.

[30] There are several reasons why the number of matched workers might exceed total employment. First, there may be errors in the industry or geographic codes for some workers or establishments in the SEDF or SSEL. Second, there is a time difference built into the Census Bureau surveys of workers and employers. The census asks workers where they worked on April 1 and employers how many workers they employed as of March 12; total employment may differ on the two dates. A third problem is that workers may be incorrectly assigned to locations because of imprecise SEDF questions. Because the SEDF asks workers only where they worked in the past week, workers who were working at a site other than their employer's primary location may be improperly assigned to an establishment. Fourth, in the SSEL, total employment includes only a plant's employees, not its owners. In the SEDF, however, both owners and employees are assigned to a particular establishment.

Groshen, Erica L. "The Structure of the Female/Male Wage Differential: Is It Who You Are, What You Do, or Where You Work?" *Journal of Human Resources* 26 (Summer 1991): 457–72.

Johnson, George, and Solon, Gary. "Estimates of the Direct Effects of Comparable Worth Policy." *American Economic Review* 76 (December 1986): 1117–25.

Korenman, Sanders, and Neumark, David. "Marriage, Motherhood, and Wages." *Journal of Human Resources* 27 (Spring 1992): 233–55.

Macpherson, David A., and Hirsch, Barry T. "Wages and Gender Composition: Why Do Women's Jobs Pay Less?" *Journal of Labor Economics* 13 (July 1995): 426–71.

Mairesse, Jacques, and Greenan, Nathalie. "Using Employee Level Data in a Firm Level Econometric Study." In *The Creation and Analysis of Employer-Employee Matched Data*, edited by John Haltiwanger, Julia Lane, James Spletzer, Jules Theeuwes, and Kenneth Troske, pp. 489–512. Amsterdam: Elsevier Science, 1999.

Neumark, David. "Employers' Discriminatory Behavior and the Estimation of Wage Discrimination." *Journal of Human Resources* 23 (Summer 1988): 279–95.

Oaxaca, Ronald. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14 (October 1973): 693–709.

Polachek, Solomon. "Differences in Post-school Investment as a Determinant of Market Wage Differentials." *International Economic Review* 16 (May 1975): 451–70.

Sorensen, Elaine. 1989. "Measuring the Pay Disparity between Typically Female Occupations and Other Jobs: A Bivariate Selectivity Approach." *Industrial and Labor Relations Review* 42 (July 1989): 624–39.

Stewart, Jay. "Has Job Mobility Increased? Evidence from the Current Population Survey, 1975–1995." Working Paper no. 308. Washington, DC: U.S. Bureau of Labor Statistics, 1998.

Stinson, John F., Jr. "New Data on Multiple Jobholding Available from the CPS." *Monthly Labor Review* 120 (March 1997): 3–8.

Troske, Kenneth R. "The Worker-Establishment Characteristics Database." In *Labor Statistics Measurement Issues*, edited by John Haltiwanger, Marilyn Manser, and Robert Topel, pp. 371–404. Chicago: University of Chicago Press, 1998.