Gender segregation and the evolution of the earnings gap in the U.S.*

Vivienne Groves† and Kristin McCue‡

April 15, 2015

Abstract

Changes in women's economic roles revolutionized the 20th century workforce. An extensive body of literature has documented a dramatic rise in female workforce participation, entry of women into previously male dominated industries and occupations, and a fall in the earnings gap between male and female workers. Yet research on how the gender earnings gap is influenced by the segregation of male and female workers across industries and firms has been limited. We use matched employer-employee data from Census's Longitudinal Employer-Household Dynamics (LEHD) database to examine segregation and its effect on the gender earnings gap in the U.S. We find that gender segregation declined only slightly between 1992 and 2011. Segregation of men and women across industries explains much of the observed segregation of male and female workers across firms. But even within industries, men and women are more segregated than would be predicted based on a random matching of workers to firms. Our estimates suggest that measures of systematic segregation are sensitive to choice of reference group used to normalize segregation measures, which may have led to an over estimate of the extent of segregation in previous work. Our regression analysis of earnings shows that men work in higher paid industries and firms, but that most of the gender earnings gap–and most of its decline over this period–is the result of within-firm differences in pay.

*__This is a preliminary draft, please do not cite or circulate.__

†Stanford Graduate School of Business. Email: vivienne.groves@stanford.edu.

‡Center for Economic Studies, United States Census Bureau. Email: kristin.mccue@census.gov. We would like to thank Eddie Lazear, Paul Oyer, and Lanier Benkard for their invaluable guidance. We would also like to thank Renee Bowen, Caroline Hoxby, Chad Jones, David Kreps, Charles O'Reilly III, and Andrzej Skrzypacz for their helpful advice. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

Contents

1 Introduction

The latter half of the 20th century saw a dramatic convergence in the roles of men and women in the U.S. labor market. Female workforce participation roughly doubled between 1950 and 1990, reducing the participation gap of 25 to 54 year old workers from 58.7% to 19.6% (Mosisa and Hipple (2006)); women became less likely to work in clerical and service positions, and more likely to enter managerial and professional jobs that were once overwhelmingly dominated by men (Blau and Kahn (2006), Blau and Kahn (2013)); and the educational attainment gap between men and women narrowed and, in recent years, reversed among young workers.[1] As a result, the gender wage gap declined substantially (Blau and Kahn (2000), Blau and Kahn (2006)). But these trends have appeared to slow down, and a significant gender pay gap remains.

Studies of gender segregation across employers using data from the 1980s and 1990s found that firm segregation, too, was an important factor in the earnings gap (Carrington and Troske (1995), Carrington and Troske (1998b), Bayard et al. (2003)). That is, even within narrow industries, men and women tended to work for different employers, and men on average worked for higher paying firms than women did. Dramatic changes in women's careers along other dimensions suggest that we should expect gender segregation to have declined as well, and for that decline to help explain the narrowing in the gender gap in pay. But the scattered time periods for which we have evidence and the differences across the data sets used in this literature make it unclear whether this is in fact the case.

We use data from Census's Longitudinal Employer-Household Dynamics (LEHD) database to examine gender segregation across firms and its relationship to the earnings gap over the period 1992-2011. Thus our first contribution is simply to document the evolution of segregation and the earnings gap in the U.S. using a broadly representative time-series. We find a modest decline in segregation, accompanied by a more substantial decline in the earnings gap. We show that shifts in the distribution of employment played an important role in determining trends in segregation over this period. The decline of the manufacturing sector and the growth in the health care and education sectors constituted a large shift of employment away from predominately male firms and towards those that are predominately female. Moreover, the fall in segregation within the retail sector—a relatively segregated sector, despite employing men and women in equal numbers, and one that accounts for a large and increasing share of employment—made a substantial contribution to the decline in segregation over the time period.

[1] A note published by the Census Bureau (2011) reports that, among 25 to 29 year olds in 2010, 36% of women held at least a bachelor degree, compared to 28% of men. Among all adults 25 or older, men retained a slight advantage: 30.3% versus 29.6% for women.

Our second contribution is to document the relative importance of workers' industries and firms of employment in accounting for segregation and the earnings gap. While important components of both segregation and the earnings gap occur across firms within industries, we find a more important role for differences between the distribution of men and women across industries. Nonetheless, most of the decline in the earnings gap remains when we include firm fixed effects, and thus reflects a within-firm decline in the gap. While reduced gender discrimination is one candidate explanation for the decline, there are alternatives. (Interestingly, education is shown to play little role.) Observed gender differences in hours of work and occupations undoubtedly account for a portion of the remaining gap. We plan to incorporate measures of these variables from household surveys in future work to account for their effects.

We also examine how the earnings gap evolves over time for cohorts of workers. We find that the gap increases substantially over workers' careers, but that it has fallen across successive cohorts of workers at any given age.

We follow relatively recent studies of segregation by distinguishing segregation resulting from job matching from that occurring by chance. In addition to reporting segregation indices, we simulate the distribution of firm gender shares that would result if, holding the within-industry firm size distribution fixed, each firm's probability of hiring a female worker were equal to the share of women in the firm's industry. We carry this out for varying levels of industry detail (4-digit North American Industry Classification System (NAICS) industry, 3-digit NAICS, NAICS sector, and pooling all industries) to illustrate the sensitivity of our findings to our choice of proxy for the relevant labor pool in simulating segregation resulting from chance. We then use the information index (Theil (1972)) as our preferred measure of segregation. Reardon and Firebaugh (2002) have shown that the this index can be decomposed into segregation across groups and segregation within groups—a property that the more commonly used Gini and dissimilarity indices do not have. Capitalizing on this property, we show that between NAICS sector segregation accounts for almost half of observed segregation, with between 3-digit and 4-digit NAICS industries also making non-trivial contributions. Using the normalization proposed in Carrington and Troske (1997), we confirm previous findings that there is gender segregation in U.S. firms beyond that expected by chance. Yet, we show that the normalized information index can differ noticeably depending on the grouping within which the randomization is carried out, suggesting that the limitations of earlier data sources led to some overestimation of the importance of gender segregation. Thus, the decomposability of the information index and our more representative sample allows us to highlight the importance of taking into account the method for randomly distributing workers amongst firms when interpreting values of normalized segregation

indices.

In the next section we discuss the LEHD data that we use to measure segregation and the wage gap. In Section 3, we detail our methods for measuring segregation and randomly allocating workers to firms, and present our segregation results in Section 4. We provide the estimation method and results for our wage regression in Section 5. Finally, we discuss plans for future research and conclude in Sections 6 and 7.

2 Data

We base our empirical analysis on a sample of employers from the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) database which draws much of its data from complete sets of unemployment insurance (UI) earnings records for U.S. states. Workers' quarterly UI earnings records have been matched to characteristics of their employers drawn from quarterly administrative UI reports and from the Census Bureau's business censuses and surveys.[2] Each quarterly wage record includes a UI account number which links individuals to their firm of employment within a state in a specific quarter. We have basic demographic data (gender, age, place of birth) for all workers. This allows us to measure gender shares of employment on a quarterly basis for each employer in our sample.

Where firms have more than one location within a state, the LEHD data identify each separate location (or establishment) along with its industry, employment level, and total payroll. However, in most states, workers employed by multi-establishment firms cannot be definitively assigned to their specific location of work, so here we measure segregation across firms rather than establishments. Where a firm operates establishments in more than one industry in a state, we assign the industry code that accounts for the most employment.[3]

Our sample includes both private sector employers and state and local governments, but does not include the federal government. While all U.S. states and the District of Columbia now participate in the LEHD program, data availability in earlier years varies across states. The main results presented here are based on data from eight states for which we have data from at least 1992 onwards.[4] We take a 10% sample of employers from those states, and in quarter t use information on all employees

[2] Abowd et al. (2009) gives a full description of the database.

[3] While we do not make use of them here, the LEHD program has developed imputations that assign individuals to specific establishments within a state through multiple imputations based on a rich set of information including the location of the firm's establishments in that state, the worker's place of residence, and the employment histories of both worker and establishment. See Abowd et al. (2009) for details.

[4] The states are: California, Idaho, Illinois, Maryland, North Carolina, Oregon, Washington, and Wisconsin. In future drafts we plan to provide results using a 22 state sample for 1996-2011, which adds Colorado, Connecticut, Florida, Hawaii, Kansas, Louisiana, Maine, Minnesota, Missouri, Montana, New Jersey, New Mexico, Rhode Island, Texas.

of sampled employers who satisfy our age restrictions and have earnings in quarters $t-1$, t, and $t+1$. This last restriction is imposed to narrow the sample to workers who were employed at both the beginning and end of a quarter. We do this primarily so that we can be reasonably sure that most sample members worked all 13 weeks in the quarter, in which case gender differences in earnings are not affected by gender differences in weeks of work, but rather reflect differences in weekly earnings. Imposing that restriction also excludes very transient employees, which we see as desirable.

We include workers aged 15 to 75 when measuring segregation, but narrow the age range to 18 to 60 for our analysis of the earnings gap.[5] We exclude agricultural employers because coverage of agriculture is incomplete and varies from state to state. In constructing segregation measures, we aggregate a few detailed industries with very small samples to maintain a consistent set of industry cells across all quarters.[6] The resulting data set includes information for years 1992 to 2011 on approximately 170,000 employers and 2.7 million workers per quarter. We do not use data from 1992 for the earnings regressions because we require at least one year of data on an employee's work history to use in constructing job tenure.

3 Measuring segregation

Before giving details about how we measure segregation, it is worth emphasizing that what we mean by segregation is differences in the *distribution* of men and women across employers who hire from the same pool of labor. For instance, if 10% of available workers are female and all firms in the population have 10% female employees, by our definition that population is not segregated at all, despite all firms having mostly male employees. In contrast, a collection of firms whose populations of male and female workers are of equal size, but that is characterized by a large number of predominantly male or predominantly female firms is considered segregated. Firm segregation is inherently a characteristic of a group of firms, not an individual firm. Thus, measured segregation will depend on how the population is defined or, equivalently, how firms are grouped together. We focus on the share female in a firm's industry as providing our best proxy for the gender share of the firm's hiring pool. We realize that this is imperfect: ideally we would use gender shares by occupation and occupational shares for each employer, but those data are only available for survey samples. In practice, many firms will hire primarily from local labor markets, but we ignore this

[5] In future work we plan to (i) use the narrower age range for our segregation results as well as the earnings gap; (ii) increase the sampling rate so as to have a reasonably sized subsample of workers who are in the household survey samples in order to be able to examine hours and occupation.

[6] With a higher sampling rate this may not be necessary. We plan to review this in the future.

here. Our reasoning is that geographic variation in gender shares is likely to be quite limited.[7] Note that we are not assuming that workers remain in the same industry over time, just that movement of workers between industries do not have have important short-term effects on gender shares.

Although a large body of literature has examined the relative strengths and weaknesses of numerous indices of segregation,[8] there is no clear consensus as to the best measure. Duncan and Duncan (1955) popularized the dissimilarity index in the sociology literature, while the economics literature relies more on the Gini index of segregation (a measure that is analogous to the Gini index of income inequality). The information index (Theil (1972)) is less commonly used, but Reardon and Firebaugh (2002) show that it has many desirable properties.[9] Among them are two that do not hold for the Gini and dissimilarity indices: additive organizational decomposability and group decomposability. Together these allow us to decompose segregation within the U.S. into segregation between sectors, 4-digit NAICS industries, and 3-digit NAICS industries, and segregation within each of these partitions. These properties are the reason we focus most of our discussion on results using the information index, although we do present Gini and dissimilarity indices in some instances so that we can compare our findings with those in the existing literature.

In the next section we describe the calculation of the information index and its decomposition into within and between industry and sector segregation. In Sections 3.2 and 3.3 we detail the calculation of the Gini and dissimilarity indices. Since we are interested in comparing segregation to that which would occur naturally if workers were randomly allocated to firms, we outline a method for calculating a randomized information index in Section 3.4, and in Section 3.5 we use this to normalize our segregation indices. Finally, in Section 3.6, we discuss our empirical method for randomly allocate workers to firms.

3.1 Information index

We are interested in characterizing the overall level of segregation in the U.S. We begin by outlining how we calculate segregation of men and women amongst a single collection of firms, ignoring any partitioning into more detailed industries. This can be used to measure gender segregation across the U.S., within any NAICS sector, or within any more detailed industry. To illustrate the decomposition, we describe the calculation in terms of a 4-digit industry. We then demonstrate how the information index for the parent 3-digit industry can be written as the sum of segregation

[7] We would make a different choice if we were measuring segregation by race, ethnicity, or foreign-born status, where some geographic areas have very different population shares.

[8] See for example Reardon and Firebaugh (2002), Boisso et al. (1994), Carrington and Troske (1997), Duncan and Duncan (1955), Hutchens (1991), Hutchens (2004), Silber (1989a), Silber (1989b), Silber (1992).

[9] For instance, that exchanging a woman in a male dominated firm for a man in a female dominated firm results in an increase in the index.

between its component 4-digit industries and the segregation within each of those 4-digit industries. This decomposition can be applied to split overall segregation into the sum of between and within segregation for any industry partition. We conclude this section by describing in full the decomposition of overall segregation into a nested series of industry partitions, for which we report results in Section 4.

In an abuse of notation we let A be the number of sectors, B be the number of 3-digit industries, and C be the number of 4-digit industries. The reader should keep in mind that the number of 3-digit industries within any sector and the number of 4-digit industries within any 3-digit industry will, in reality, differ by sector and 3-digit industry, respectively, and that we take this into account when performing our calculations.[10]

Suppose a 4-digit industry c, within 3-digit industry b, which lies in sector a, has J firms. Let $T(j,c,b,a,US)$ be the number of employees in firm j, $T(c,b,a,US)$ be the number of employed people in 4-digit industry c, $T(b,a,US)$ be the number of employed people in 3-digit industry b, $T(a,US)$ be the number of employed people in sector a, and $T(US)$ be the number of workers in the U.S. Also, let $\pi_f(c,b,a,US)$ be the proportion of workers in 4-digit industry c who are women and $\pi_m(c,b,a,US)$ be the proportion who are male, and define $\pi_f(b,a,US)$, $\pi_m(b,a,US)$, $\pi_f(a,US)$, $\pi_m(a,US)$, $\pi_f(US)$ and $\pi_m(US)$ analogously.

Then, the information index for firms in industry c is given by

$$H(c,b,a,US) = \sum_{j=1}^{J} \frac{T(j,c,b,a,US)}{T(c,b,a,US)E(c,b,a,US)}\Big(E(c,b,a,US) - E(j,c,b,a,US)\Big) \qquad (1)$$

where

$$\begin{aligned} E(c,b,a,US) &= \pi_f(c,b,a,US)\log\left(\frac{1}{\pi_f(c,b,a,US)}\right) + \pi_m(c,b,a,US)\log\left(\frac{1}{\pi_m(c,b,a,US)}\right), \text{ and} \\ E(j,c,b,a,US) &= \pi_f(j,c,b,a,US)\log\left(\frac{1}{\pi_f(j,c,b,a,US)}\right) + \pi_m(j,c,b,a,US)\log\left(\frac{1}{\pi_m(j,c,b,a,US)}\right) \end{aligned}$$

The entropy index $E(j,c,b,a,US)$ is a measure of diversity in firm j. We assign it a value of zero when workers are either entirely male or entirely female, and it is maximized when the share of men and women in firm j are equal. The information index compares the entropy index of each firm within the 4-digit NAICS industry to the industry entropy index and takes an employment-weighted average across firms.

Ignoring, for the moment, any 4-digit industry partitions, we can use the same method to

[10] See `http://www.census.gov/eos/www/naics/` for examples.

calculate the information index for a 3-digit industry:

$$H(b,a,US) = \sum_{c=1}^{C}\sum_{j=1}^{J}\frac{T(j,c,b,a,US)}{T(b,a,US)E(b,a,US)}\Big(E(b,a,US)-E(j,c,b,a,US)\Big) \quad (2)$$

where

$$E(b,a,US) = \pi_f(b,a,US)\log\left(\frac{1}{\pi_f(b,a,US)}\right)+\pi_m(b,a,US)\log\left(\frac{1}{\pi_m(b,a,US)}\right)$$

is the entropy index for sector b. Additionally, if we put aside the partition of 3-digit industries into firms, we can calculate the between 4-digit industry segregation within a 3-digit industry:

$$H_{\text{between}}(b,a,US) = \sum_{c=1}^{C}\frac{T(c,b,a,US)}{T(b,a,US)E(b,a,US)}\Big(E(b,a,US)-E(c,b,a,US)\Big) \quad (3)$$

We show in Section A.1 that segregation within a 3-digit industry can be decomposed into the sum of between 4-digit segregation and a weighted sum of within 4-digit segregation:

$$H(b,a,US) = H_{\text{between}}(c,b,a,US)+\sum_{c=1}^{C}\frac{T(c,b,a,US)E(c,b,a,US)}{T(b,a,US)E(b,a,US)}H(c,b,a,US) \quad (4)$$

where the weights for each 4-digit industry segregation measure, $H(c,b,a,US)$, depend on the size of the 4-digit industry as well as the diversity in that industry. Similarly, segregation within a sector can be be decomposed into between and within 3-digit industry segregation, and overall segregation can be decomposed into between and within sector segregation:

$$H(a,US) = H_{\text{between}}(b,a,US)+\sum_{b=1}^{B}\frac{T(b,a,US)E(b,a,US)}{T(a,US)E(a,US)}H(b,a,US), \text{ and} \quad (5)$$

$$\underbrace{H(US)}_{(z)} = \underbrace{H_{\text{between}}(a,US)}_{\text{Between sector }(d)}+\underbrace{\sum_{a=1}^{A}\frac{T(a,US)E(a,US)}{T(US)E(US)}H(a,US)}_{\text{Weighted sum within sector }(e)} \quad (6)$$

Equation (6) gives a decomposition for total U.S. segregation into between and within sector segregation. Using equations (4) and (5) we get two alternate decompositions of overall segregation based on recursively partitioning the population of firms into 3-digit and and then 4-digit NAICS industries:

$$
\begin{aligned}
H(US) &= H_{\text{between}}(a, US) + \underbrace{\sum_{a=1}^{A} \frac{T(a,US)E(a,US)}{T(US)E(US)} H_{\text{between}}(b,a,US)}_{\text{Weighted sum between 3-digit } (f)} \\
&\quad + \underbrace{\sum_{a=1}^{A}\sum_{b=1}^{B} \frac{T(b,a,US)E(b,a,US)}{T(US)E(US)} H(b,a,US)}_{\text{Weighted sum within 3-digit } (g)}
\end{aligned} \tag{7}
$$

and

$$
\begin{aligned}
H(US) &= H_{\text{between}}(a, US) + \sum_{a=1}^{A} \frac{T(a,US)E(a,US)}{T(US)E(US)} H_{\text{between}}(b,a,US) \\
&\quad + \underbrace{\sum_{a=1}^{A}\sum_{b=1}^{B} \frac{T(b,a,US)E(b,a,US)}{T(US)E(US)} H_{\text{between}}(c,b,a,US)}_{\text{Weighted sum between 4-digit } (h)} \\
&\quad + \underbrace{\sum_{a=1}^{A}\sum_{b=1}^{B}\sum_{c=1}^{C} \frac{T(c,b,a,US)E(c,b,a,US)}{T(US)E(US)} H(c,b,a,US)}_{\text{Weighted sum within 4-digit } (k)}
\end{aligned} \tag{8}
$$

In Section 4, we report the information index for the U.S. that arises from our data and decompose this into the weighted sums of between and within sector, 3-digit industry, and 4-digit industry segregation. The letter symbols, e, d, f, g, h, k, and z represented underneath the various terms in equations (6), (7), and (8) are used in Table 2 as a marker to identify precisely which measure of segregation is being reported.

This ability to decompose the information index is one of its major benefits, and is why we prefer it to more commonly used measures. Although we can measure the Gini and dissimilarity indices for any subset of our data, neither the Gini nor the dissimilarity index for the U.S. is a weighted sum of the between and within sector or industry segregation. While we primarily use the information index, we also apply the dissimilarity and Ginis indices to some of our results for comparison purposes, so we briefly review the calculation of the dissimilarity and Gini indices in the next two sections.

3.2 Duncan and Duncan's dissimilarity index

First introduced by Duncan and Duncan (1955), the dissimilarity index measures the share of workers that would need to change employers in order to create a workforce in which each firm

perfectly mimicked the diversity of the population of workers within an industry. It takes values in $[0, 1/2]$.

Consider the dissimilarity index for a 4-digit NAICS industry c. If we let t^g be the number of workers of gender $g \in \{f, m\}$ within industry c, and t^g_j be the number of workers of gender $g \in \{f, m\}$ at firm $j \in \{1, \ldots, J\}$, the dissimilarity index is given by

$$D = \frac{1}{2} \sum_{j=1}^{J} \left| \frac{t^f_j}{t^f} - \frac{t^m_j}{t^m} \right|$$

3.3 Gini index of segregation

The Gini index of segregation is analogous to the Gini index of income inequality. Let π^j_g be the share of gender $g \in \{f, m\}$ in firm j. If we reindex firms so that they are in ascending order of share female, $\pi^1_m < \pi^2_m < \cdots < \pi^J_m$, and we plot the cumulative share female on the horizontal axis, $\sum_{j=1}^{\hat{j}} \pi^j_f$, and the cumulative share male on the vertical axis, $\sum_{j=1}^{\hat{j}} \pi^j_m$, we get what is referred to as a segregation curve. The Gini index is equal to twice the area between this curve and the diagonal. It thus takes values in $[0, 1]$. Hutchens (1991) showed that when firms are indexed in ascending order of share female, the Gini index can be written in the following simple form

$$G = 1 - \sum_{j=1}^{J} \pi_{jf} \left(\pi_{jm} + 2 \sum_{k=j+1}^{J} \pi_{km} \right)$$

3.4 Randomized indices of segregation

We present our results primarily using summary measures of observed segregation relative to what we would expect from a random allocation of workers to firms.[11] Constructing random allocations requires some assumptions about the relevant labor pool for particular employers, and in particular, the probability with which each worker assigned to a firm will be female. We construct our primary measures of segregation assuming that the gender composition of a 4-digit NAICS industry is a reasonable proxy for the gender distribution of potential hires for all firms in that industry. We make this assumption because we have in mind that individuals' educational investments and occupational choices account for much of the across-industry differences in employment distributions by gender, and we are focused on differences in the firms that men and women end up working for, given differences in human capital investment. Because this is at best an approximation, we also construct three other series of simulations of the information index in which 3-digit NAICS industry

[11] Except in the limit, when firms are all very large in size and number, any random allocation of workers to employers will produce some segregation.

gender shares, sector gender shares, and the entire labor force gender share are used to set each firms' probability of hiring a woman.

Let $H(US)_1^*$ be the value of the information index for the U.S. when firms are randomly assigned workers, and for each worker that the firm is assigned, the probability that the worker is female is given by the mean share of women in our dataset. Let $H(US)_2^*$ be the randomized value of the information index for the U.S. associated with an assignment of workers in which each firm's probability of drawing a female worker is equal to the mean share of women in the sector in which that firm is located. Define $H(US)_3^*$ and $H(US)_4^*$ analogously. Each of $H(US)_1^*, H(US)_2^*, H(US)_3^*$, and $H(US)_4^*$ provides a value of the information index with which to compare our observed information index for the U.S., $H(US)$, and to determine the extent to which chance plays a role in the segregation of workers across firms and industries.

Further, let $H(a, US)_2^*$ be the value of the randomized information index for sector a where the probability that a firm draws a female worker is equal to the share of women in sector a observed in the data. Define $H(b, a, US)_3^*$ as the randomized segregation index for 3-digit industry b associated with an allocation of workers to firms where the probability that a firm draws a female worker is equal to the share of women in the 3-digit industry, and define $H(c, b, a, US)_4^*$ analogously. (Here the subscript reminds us of the industry or sector which determines the share of women). We show in Section A.2 that since each randomized value of the information index is the average of a number of sample information index values (as described in more detail in Section 3.6), we can write

$$\underbrace{H(US)_2^*}_{(q^*)} = H_{\text{between}}(a, US) + \underbrace{\sum_{a=1}^{A} \frac{T(a,US)E(a,US)}{T(US)E(US)} H(a,US)_2^*}_{\text{Weighted sum randomized within sector } (m^*)} \tag{9}$$

$$\begin{aligned} \underbrace{H(US)_3^*}_{(r^*)} &= H_{\text{between}}(a, US) + \sum_{a=1}^{A} \frac{T(a,US)E(a,US)}{T(US)E(US)} H_{\text{between}}(b, a, US) \\ &\quad + \underbrace{\sum_{a=1}^{A}\sum_{b=1}^{B} \frac{T(b,a,US)E(b,a,US)}{T(US)E(US)} H(b,a,US)_3^*}_{\text{Weighted sum randomized within 3-digit industry } (n^*)} \end{aligned} \tag{10}$$

$$\begin{aligned}
\underbrace{H(US)_4^*}_{(s^*)} &= H_{\text{between}}(a,US) + \sum_{a=1}^{A} \frac{T(a,US)E(a,US)}{T(US)E(US)} H_{\text{between}}(b,a,US) \\
&\quad + \sum_{a=1}^{A}\sum_{b=1}^{B} \frac{T(b,a,US)E(b,a,US)}{T(US)E(US)} H_{\text{between}}(c,b,a,US) \\
&\quad + \underbrace{\sum_{a=1}^{A}\sum_{b=1}^{B}\sum_{c=1}^{C} \frac{T(c,b,a,US)E(c,b,a,US)}{T(US)E(US)} H(c,b,a,US)_4^*}_{\text{Weighted sum randomized within 4-digit industry } (p^*)}
\end{aligned} \tag{11}$$

Thus, we can decompose $H(US)_4^*$, the total segregation in the U.S. associated with a random allocation of workers that fixes the share of women in 4-digit industries, into

(i) segregation between sectors as observed in the data,

(ii) a weighted sum of segregation between 3-digit industries as observed in the data,

(iii) a weighted sum of segregation between 4-digit industries as observed in the data, and

(iv) a weighted sum of segregation within 4-digit industries associated with a random allocation of workers.

We can decompose $H(US)_3^*$ and $H(US)_2^*$ similarly. When each firm in the U.S. has the same chance of any employee being a woman, as is the case with $H(US)_1^*$, no such decomposition applies since no between sector or industry segregation is retained from our original data.

In Section 4 we report the randomized Gini and dissimilarity indices associated with 4-digit gender shares, but cannot decompose these as we do with the information index.

3.5 Normalized indices of segregation

To compare the randomized and actual values of our information index, we follow Carrington and Troske (1997) and calculate a normalized information index, given by

$$\hat{H} = \begin{cases} \frac{H-H^*}{1-H^*} & \text{if } H \geq H^* \\ \frac{H-H^*}{H^*} & \text{if } H < H^* \end{cases} \tag{12}$$

where H is the actual value of the information index and H^* is the chosen randomized value. When $H \geq H^*$, the value $\hat{H}$ will lie in the interval $[0,1]$ and indicates the extent to which the industry is more segregated than a random allocation of workers to firms would predict, expressed as a fraction

of one less the segregation that is attributable to randomness. If $H < H^*$, then $\hat{H} \in [-1, 0)$, and $\hat{H}$ measures the extent to which the industry is more even than random, expressed as a proportion of one less the evenness that is attributable to a random assignment of workers to firm.

3.6 Randomly allocating workers to firms

To generate estimates of differences in the gender distribution of workers across firms due to chance for a particular industry/quarter cell, we use the number of sample employees in that cell, and the number of women among them, as if they were population counts for the pool of potential employees.[12] Let t_{jt} be the number of employees of firm j at time t. We take a draw of size t_{jt} from that distribution for each employer j at time t. This gives a count of (randomly assigned) female employees for each of the sample employers, while exactly replicating the sample firm size distribution. We repeat this process 100 times and average across the 100 draws.

4 Segregation results

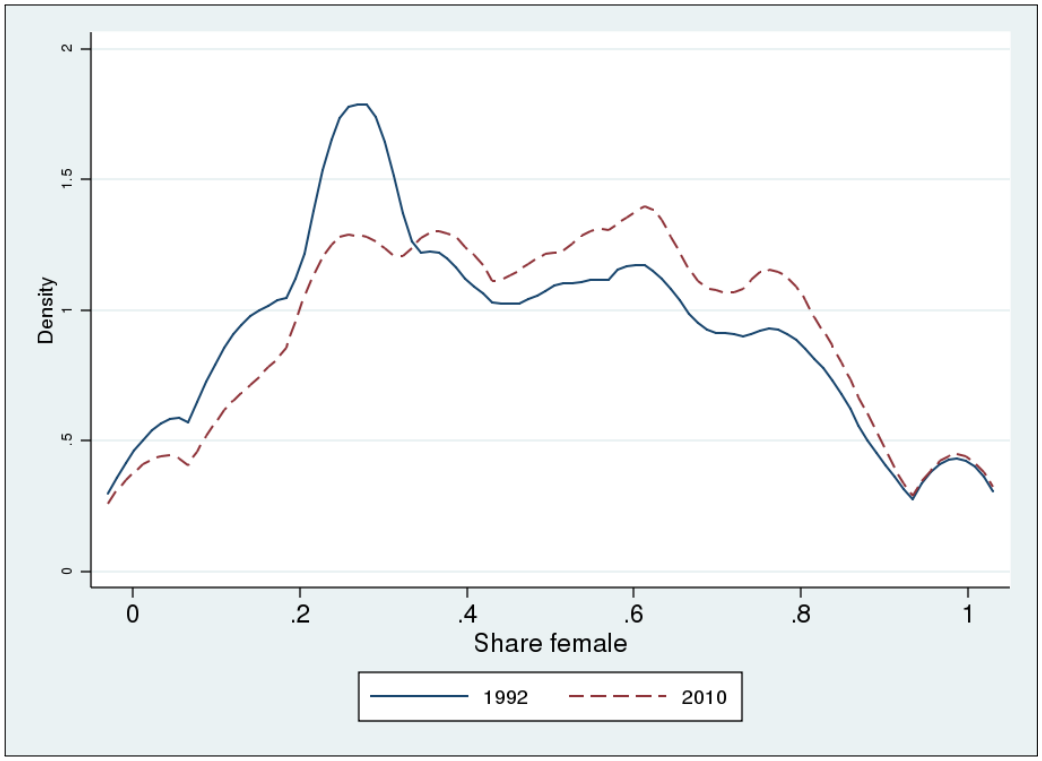
Segregation indices provide a summary of the gender differences in employment distribution that interest us. They reflect how far an economy lies from perfect segregation (where firms hire exclusively men or exclusively women) and how far it lies from perfect integration (where each firm is a microcosm of the underlying diversity within an economy). But they do not make clear, for instance, whether segregation is being driven by a small number of very homogenous firms or a large number of slightly homogenous firms. To provide a better sense of the underlying variation, we graphically describe the distribution of employment across firms characterized by the share of women in their workforces in Section 4.1. In Section 4.2 we then turn to indices based on these distributions to describe recent trends.

4.1 Graphical representation of segregation

Figure 1 shows kernel density estimates of the distribution of employment by the share of women within the firm for 1992 and 2010. In both years there is considerable lumpiness in the distribution, which is a function of both the firm size distribution and differences across industries in the gender composition of their workforces. To illustrate these distributional effects, the next four figures plot

[12] We implement this using Stata to produce pseudo-random draws of t_j employees *without* replacement from a finite population of size T that has share female equal to π. This is based on the hypergeometric distribution which describes the probability of drawing k women in n draws from a finite population. For T large relative to t_j and with π not close to 0 or 1, the hypergeometric distribution is very similar to the binomial distribution.

Figure 1: Actual shares of women in firms



a series of randomized segregation measures for each of the two years, where the different measures reflect the use of share female from progressively narrow industries in the random assignment.

In Figure 2 the randomization is based on the overall share female, which rose in our sample from 46% in 1992 to 50% in 2010. Unsurprisingly, a random assignment of men and women across all firms would result in most firms employing roughly half female and half male employees, and a kernel density estimate in 2010 that first order stochastically dominates the kernel density estimate of 1992 due to a rise in female workforce participation over that period. The less expected feature of this graph is the weight in the tails of the distribution, which reflects the non-negligible share of overall employment accounted for by very small firms: these have a high probability of having mostly female or mostly male employees due to chance.

In Figure 3, the randomization uses NAICS sector gender shares and so reflects gender differences in employment patterns at a fairly aggregate level.[13] Were the employment weights and share of women within each industry to remain fixed, and only the national female participation rate to change, we would expect the 2010 kernel density estimate to resemble a rightwards translation of

[13]The sectors are: mining (NAICS 21); utilities (22); manufacturing (31-33); wholesale trade (42); retail trade (44-45); transportation and warehousing (48-49); information (51); finance and insurance (52); real estate and rental and leasing (53); professional, scientific and technical services (54); management of companies and enterprises (55); administrative and support and waste management and remediation services (56); educational services (61); health care and social assistance (62); arts, entertainment, and recreation (71); accommodation and food services (72); other services (81); and public administration (92).

Figure 2: Kernel density estimate of share of women in firms when employees randomly allocated within U.S. firms.

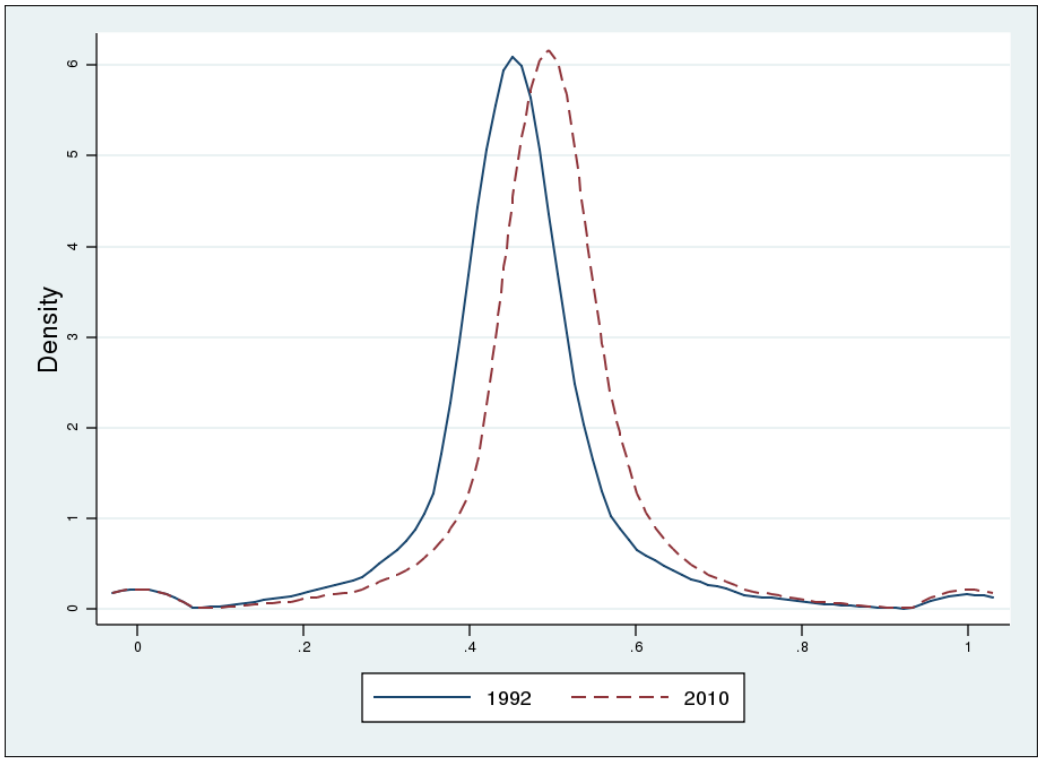


Figure 3: Kernel density estimate of share of women in firms when employees randomly allocated within their respective NAICS sector

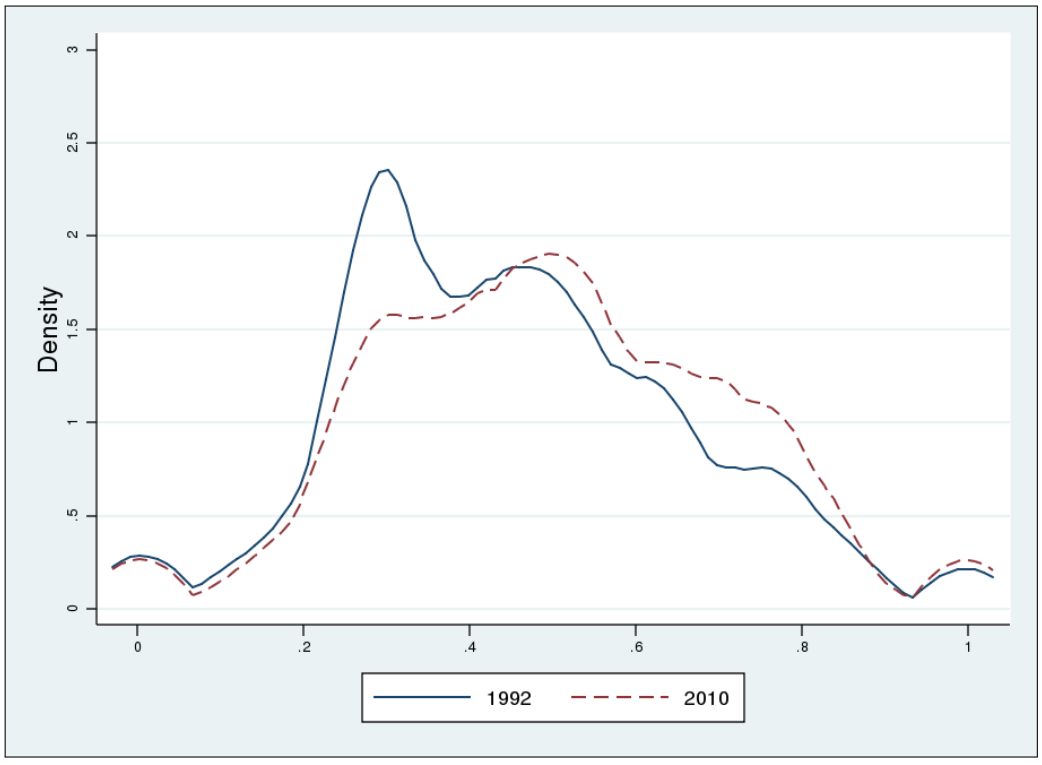
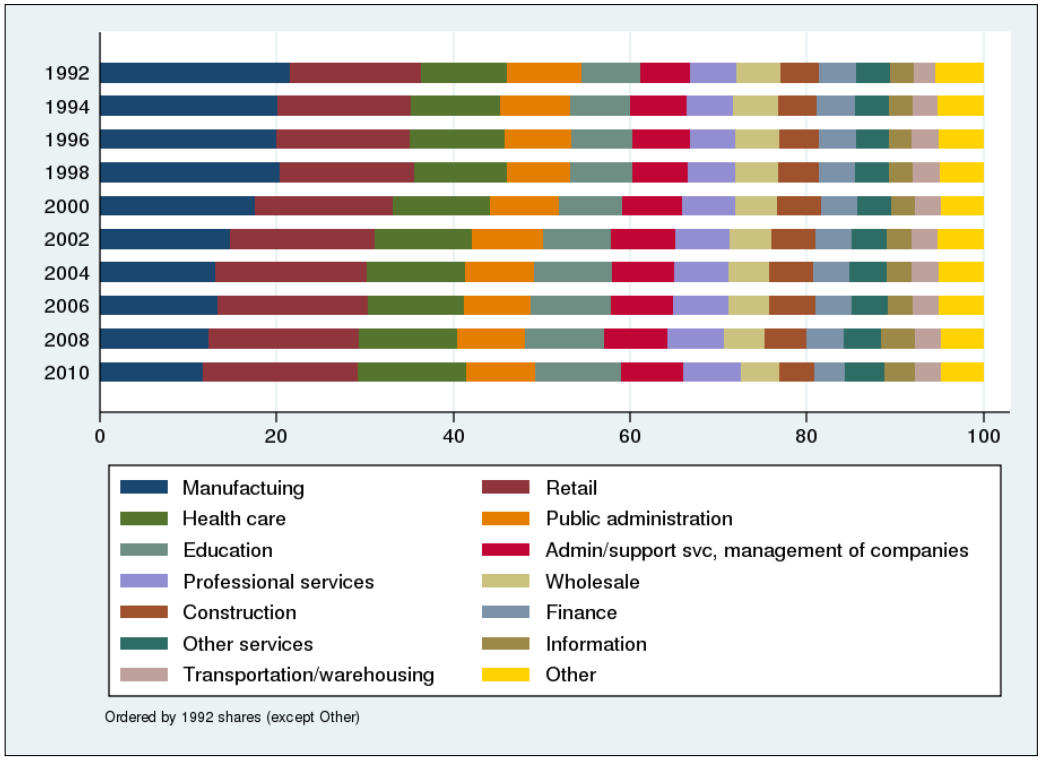


Table 1: **Employment shares and share female by sector**

Sector	Sector share of employment		Share female in sector	
	1992	2010	1992	2010
Manufacturing	21.5	11.6	30.7	29.1
Retail	14.9	17.6	50.2	52.0
Health care	9.7	12.3	78.8	78.0
Public administration	8.4	7.8	44.9	47.3
Education	6.7	9.8	63.3	68.0
Administrative and support services, management of companies	5.5	7.0	42.0	44.2
Professional services	5.3	6.6	48.0	46.1
Wholesale	4.9	4.4	32.7	33.3
Construction	4.4	3.9	16.6	18.0
Finance	4.3	3.5	66.3	63.7
Other services	3.8	4.5	54.5	59.1
Information	2.6	3.4	45.9	39.0
Transportation and warehousing	2.4	3.0	27.3	28.3
Other	5.5	4.8	36.6	42.3

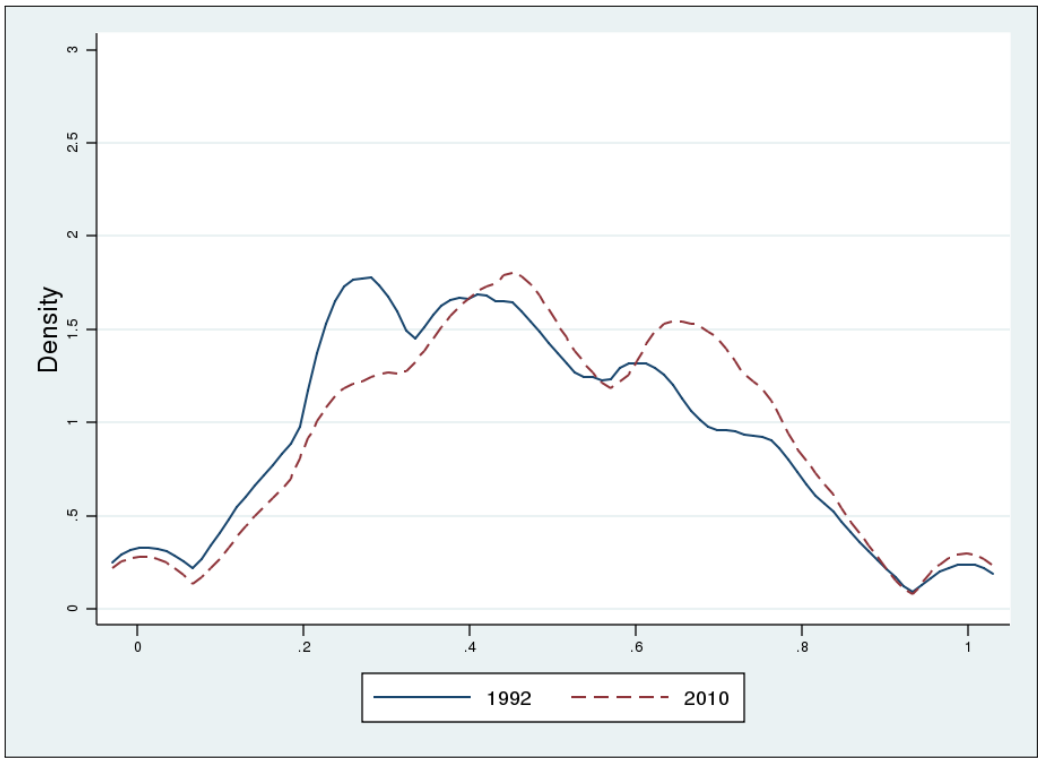
Notes: The other category is made up of the remaining sectors: real estate and leasing; arts, entertainment, and recreation; mining; and utilities.

Figure 4: Sector employment shares



the 1992 kernel density estimate, as is the case in Figure 2.[14] Figure 4 and Table 1 show, however, that the share of total employment attributable to manufacturing has fallen by almost half over this period (from 21.5% to 11.6%), while the shares for health care and education have increased by 2.6 and 3.1 percentage points, respectively. The decline of the manufacturing sector with its roughly 70% male workforce, and growth in health care and education, where women account for roughly 78% and 65% respectively, drive much of the distributional changes we observe in Figure 3.

Figure 5: Kernel density estimate of share of women in firms when employees randomly allocated within their respective NAICS 3-digit industries



Figures 5 and 6 use 3-digit and 4-digit NAICS industry categories, respectively, and the resulting randomized distributions look increasingly similar to the actual distributions for the two years, shown in Figure 1. This reflects a finding that shows up repeatedly in our results: much of the overall tendency for men and women to work for different firms reflects differing industry distributions rather than within-industry differences, and much of that difference is across broad industries rather than detailed industries.[15]

[14] We are working on developing methods to reweight our random-assignment distributions to illustrate the separate contributions of (a) the rise in the female participation rate, (b) changes in the share of women within industries, and (c) shifts in industry shares of employment. In particular, we plan to produce densities adjusted to capture the effects of (i) 1992 industry employment shares, 1992 female in industry to female employees in U.S. ratios, and a 2010 female participation rate; and (ii) 1992 employment weights, and 2010 female industry shares.

[15] We are currently investigating (and plan to investigate further) whether we can determine how much of the sorting across broad industries is based on occupation.

Figure 6: Kernel density estimate of share of women in firms when employees randomly allocated within their respective NAICS 4-digit industries

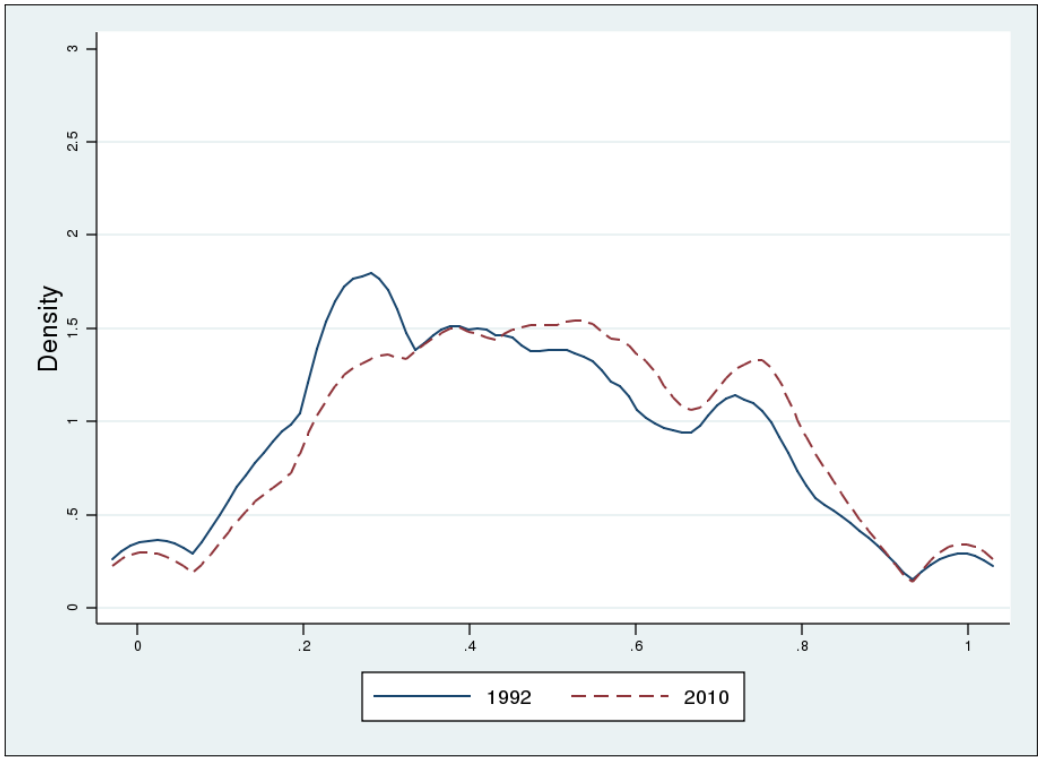


Figure 7: Randomized segregation for different industry partitions 1992

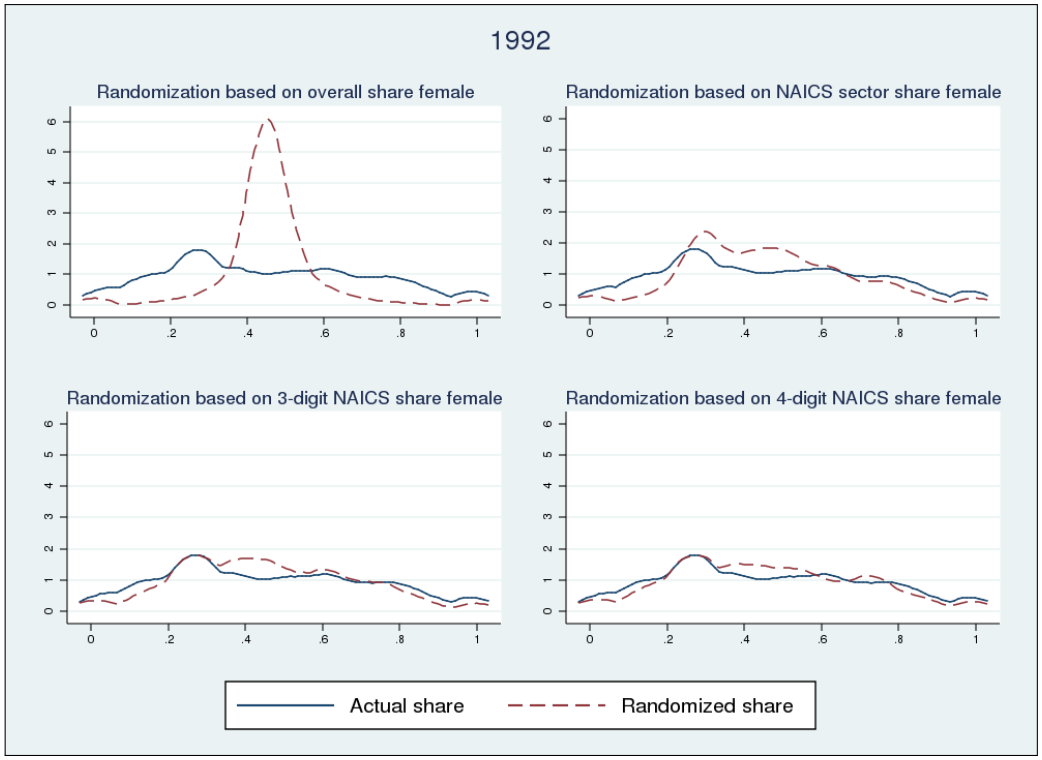


Figure 8: Randomized segregation for different industry partitions 2002

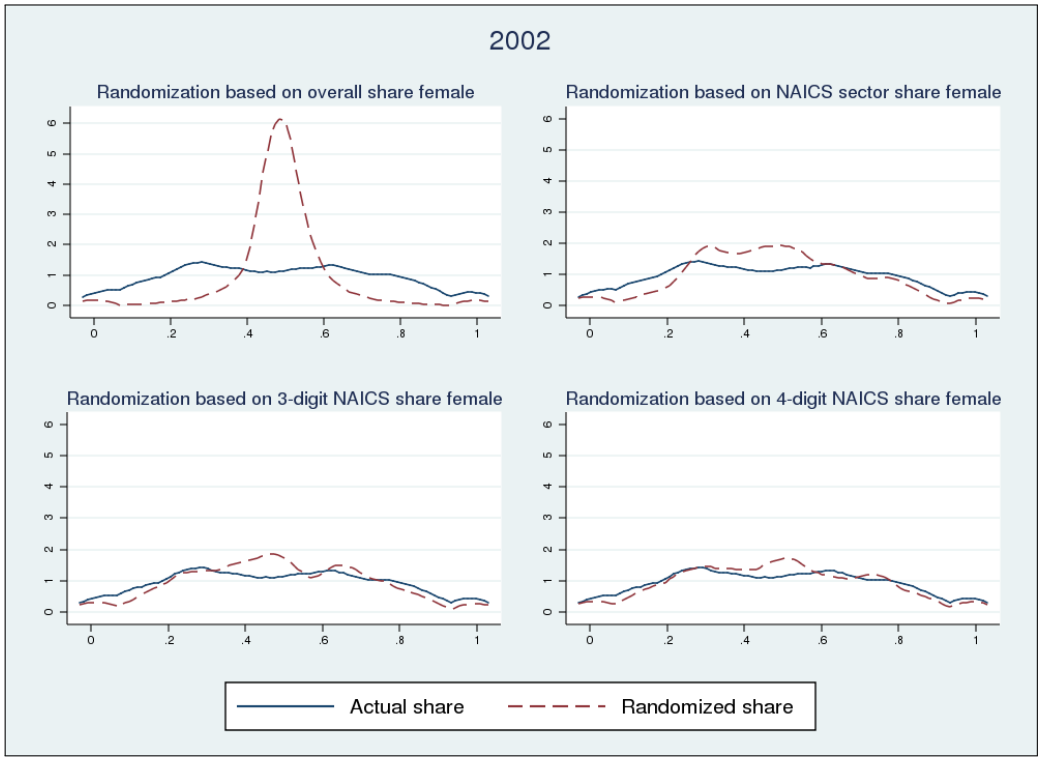
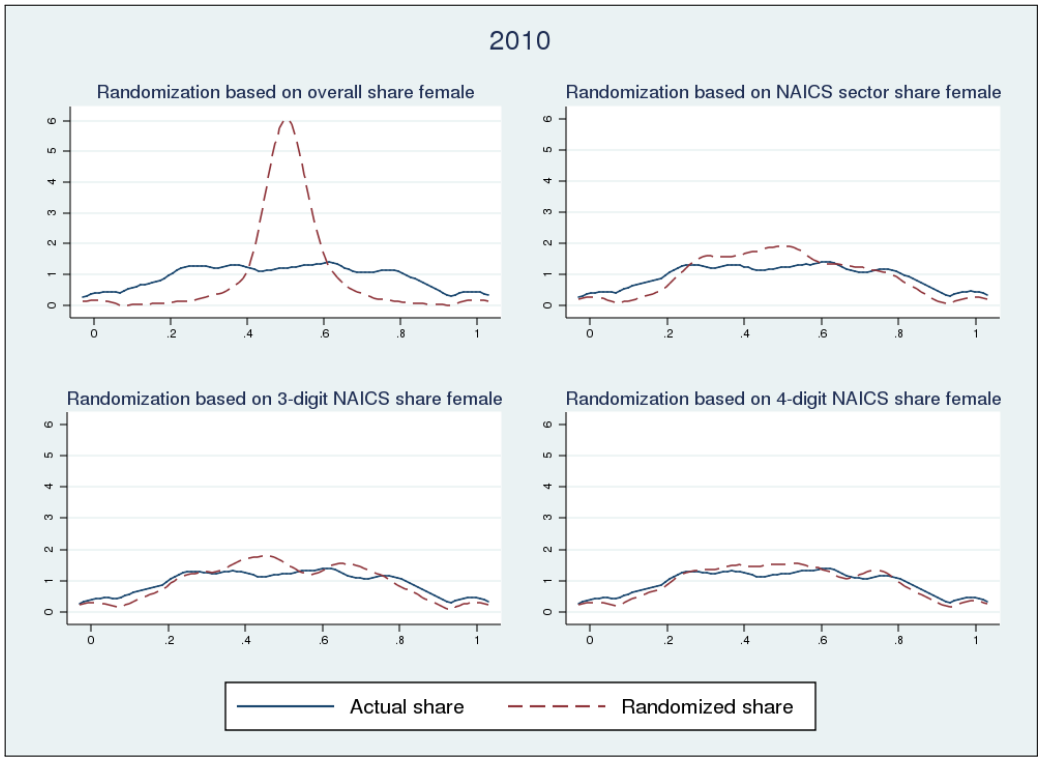


Figure 9: Randomized segregation for different industry partitions 2010



Figures 7 (and 15, see Appendix B.2) plot the actual shares of females in firms in 1992 against the randomized distributions discussed above, while Figures 8 (and 16) do the same for 2002, and Figures 9 (and 17) do the same for 2010. Regardless which of the random allocation of workers with which we compare the actual distribution of female shares, it is clear that there is systematic segregation of workers above what is caused by randomness in each of the three years. The fat tails of the actual distribution of workers provides evidence – more male dominated and female dominated workplaces and fewer gender balanced workplaces exist than in an economy where workers are randomly allocated.

Next, we present segregation estimates based on our data, consider their trends over time, and compare our findings to those of the previous literature.

4.2 Segregation indices

We start by examining trends and demonstrate that our broad findings are not sensitive to our particular choice of segregation index. We then compare our estimates of segregation to those of previous papers.

Figure 10: Information, dissimilarity, and Gini indices

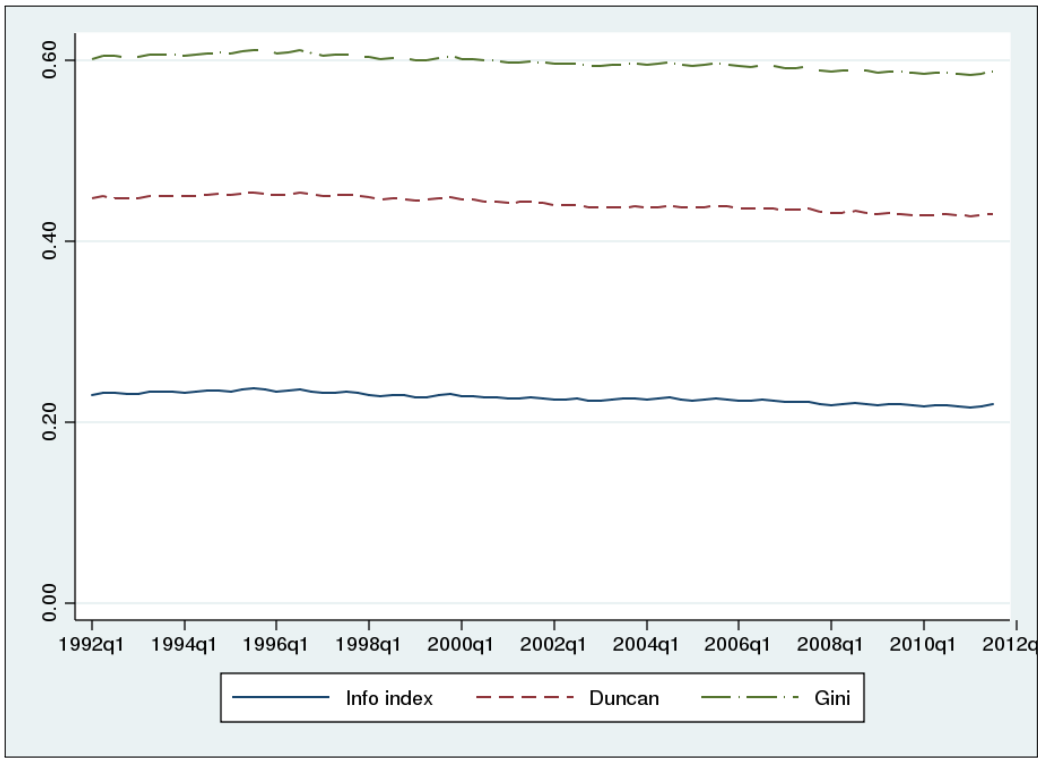
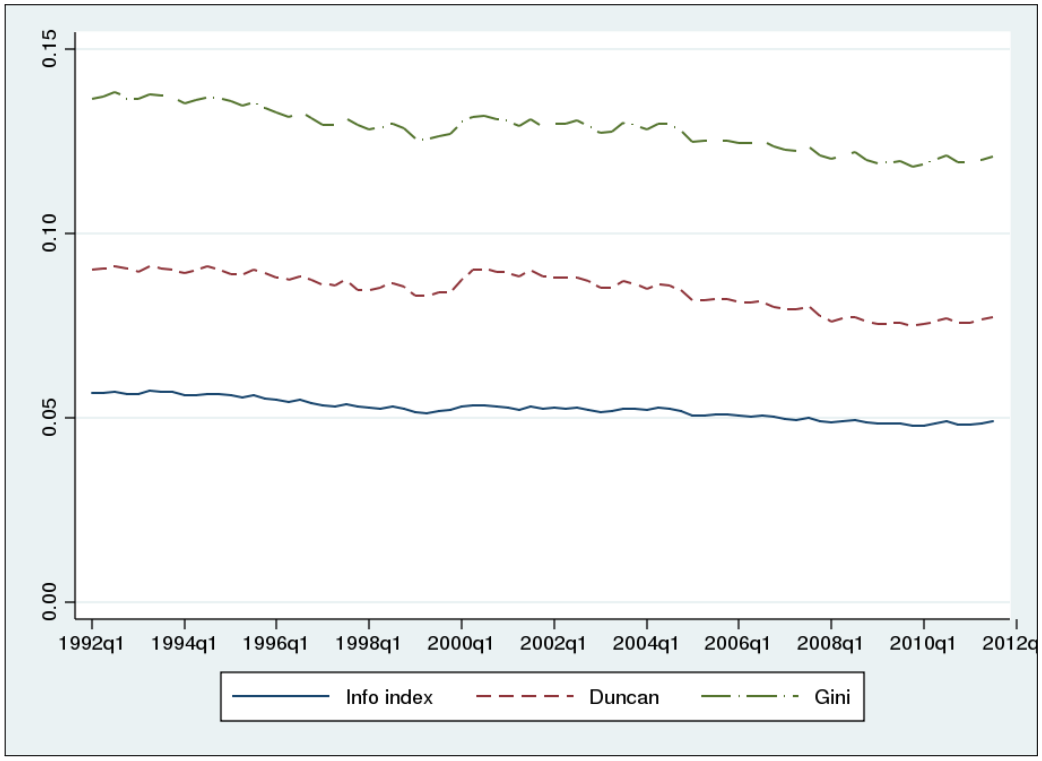


Figure 10 plots quarterly estimates of the information index based on our full sample for the first quarter of 1992 through the third quarter of 2012 (solid blue line). Surprisingly, we do not see much

decline in the information index over the 18 year time span of our data, despite substantial changes in women's relative investment in human capital and role in the labor market across cohorts.

Figure 11: Normalized information, dissimilarity, and Gini indices



This figure also includes timeseries for the dissimilarity and Gini indices calculated using the same data. In Figure 11 we present normalized versions of each of these indices. Since each is a somewhat arbitrary index of segregation (as opposed to a concrete measure of something quantifiable), we read nothing into the fact that one measure gives a higher number than another. What is reassuring is that each follows the same downward trend over the time period, and that each exhibits the same peaks and troughs for the normalized values. The variation around the trend is less pronounced for the information index, but its scale is also smaller relative to the other two indices.

Our segregation estimates for the Gini and dissimilarity indices are reasonably close to those presented in the existing literature, given the differences in the datasets used. Carrington and Troske (1995) report a value of 0.66 for the dissimilarity index, which we would expect to be greater than our value of 0.45 in 1992—assuming the decline in Figure 10 was part of a longer term pattern—since their estimate draws from 1982 data. Possibly more important, however, is that their analysis is restricted to small firms, which we would expect to have more segregation due to chance than our sample which includes many larger firms. If we estimate the dissimilarity index using only firms with fewer than 100 employees (only a rough approximation to their sample), our estimate rises

to 0.54. In their 1998b paper, Carrington and Troske calculate a Gini index of 0.59 for inter-plant gender segregation in manufacturing in 1990 and a value of 0.43 for the dissimilarity index. Again, these are somewhat greater than our estimates for manufacturing in a nearby year (1992) of 0.43 for the Gini index and 0.30 for the dissimilarity index.[16] Bygren (2013)'s dissimilarity index estimate of roughly 0.48 suggests that there might be slightly more gender segregation in Stockholm than the U.S. (although, there are many other reasons why these two figures may differ).

Comparing our normalized indices to other findings, our estimates appear very low relative to those of Carrington and Troske (1998a) and Bygren (2013). Those authors report a normalized dissimilarity index of 0.33 for manufacturing plants in 1990, and a normalized dissimilarity index of 0.39 for Stockholm in the early 2000s, respectively. Our most comparable dissimilarity estimates based on randomization within 4-digit industries are 0.07 for manufacturing in 1992, and 0.08 for all firms in 2002. Given that the original indices are closer together and we are using the same normalization, differences in the randomized indices explain the discrepancy.[17] The main difference in methods is in the group used to proxy for the labor pool from which firms hire in simulating random hiring. While Carrington and Troske and Bygren assign workers to firms according to the overall share of women within their dataset, we assign workers to firms using the share of women in each 4-digit NAICS industry. If one accepts that 4-digit NAICS industry gender shares are a better proxy for female hiring probabilities of firms than sector gender shares, then this would suggest that past estimates of normalized segregation indices have overstated segregation.

4.3 Decomposing the information index

Table 2 presents the components of the decomposition of the information index that we outlined in Section 3 for selected years. The actual, randomized, and normalized values appear in the separate panels of the table. The bottom line in the first panel gives the overall information index for our sample, while the lines above it present components of that total. Looking at the bottom line, the information index declined from 0.232 in 1992 to 0.218 in 2010. The lines above indicate whether particular components changed over time. For example, within 4-digit industry, segregation steadily declined from 0.109 in 1992 to 0.102 in 2010, though the change was modest. Summing up the effects

[16] While their manufacturing sample is not explicitly limited to large firms, it is constructed by matching survey data on individuals to their employers. This sample of employers is selected proportional to size, so small employers are likely to be under-represented in their matched sample. While we find that in general segregation is somewhat lower in large firms, in manufacturing that problem seems to be reversed. That is, segregation is on average higher among large manufacturing firms than among small manufacturing firms. This may simply reflect differences across detailed manufacturing industries. In future work we hope to more closely mimic the sample scheme used in Carrington and Troske (1998b) to determine if our resulting estimates are more similar to those reported by the authors.

[17] For example, our random dissimilarity index for 1992 is 0.25, while Carrington and Troske report a 1990 value of 0.16.

Table 2: **Information index estimates**

Measure		Year 1992	1995	2000	2005	2010
Estimates based on data						
Between partition segregation						
Between sector	d	0.081	0.084	0.085	0.083	0.082
Weighted sum between 3-digit	f	0.027	0.028	0.024	0.022	0.019
Weighted sum between 4-digit	h	0.015	0.017	0.016	0.016	0.015
Sum between (sector, 3-digit, 4-digit)	$d+f+h$	0.123	0.128	0.125	0.121	0.116
Within partition segregation						
Weighted sum within sector	$e=f+g$	0.151	0.152	0.144	0.142	0.136
Weighted sum within 3-digit	$g=h+k$	0.124	0.124	0.120	0.120	0.117
Weighted sum within 4-digit	k	0.109	0.108	0.104	0.104	0.102
Information index for U.S.						
Within and between	$z=(d+f+h)+k$	0.232	0.236	0.229	0.225	0.218
Randomized estimates						
Within partition segregation						
U.S.	l^*	0.056	0.056	0.055	0.059	0.059
Weighted sum within sector	m^*	0.059	0.060	0.058	0.062	0.062
Weighted sum within 3-digit	n^*	0.060	0.061	0.059	0.061	0.062
Weighted sum within 4-digit	p^*	0.063	0.063	0.061	0.063	0.063
Information index for U.S.						
Fixing U.S. share of women	l^*	0.056	0.056	0.055	0.059	0.059
Fixing sector shares of women	$q^*=m^*+d$	0.140	0.144	0.143	0.146	0.144
Fixing 3-digit shares of women	$r^*=n^*+d+f$	0.168	0.172	0.168	0.166	0.163
Fixing 4-digit shares of women	$s^*=p^*+d+f+h$	0.185	0.191	0.185	0.184	0.179
Normalized estimates						
Fixing U.S. share of women	$\frac{z-l^*}{1-l^*}$ if $z \geq l^*$ $\frac{z-l^*}{l^*}$ if $z < l^*$	0.187	0.191	0.183	0.177	0.169
Fixing sector shares of women	$\frac{z-q^*}{1-q^*}$ if $z \geq q^*$ $\frac{z-q^*}{q^*}$ if $z < q^*$	0.106	0.108	0.100	0.093	0.087
Fixing 3-digit shares of women	$\frac{z-r^*}{1-r^*}$ if $z \geq r^*$ $\frac{z-r^*}{r^*}$ if $z < r^*$	0.077	0.077	0.073	0.071	0.066
Fixing 4-digit shares of women	$\frac{z-s^*}{1-s^*}$ if $z \geq s^*$ $\frac{z-s^*}{s^*}$ if $z < s^*$	0.057	0.056	0.053	0.051	0.049

of differing employment distributions across sectors, 3-digit industries within sectors, and 4-digit industries within 3-digit industries, we find that between-industry segregation rose over the first part of our period, from 0.123 in 1992 to to 0.128 in 1995, but later fell to end at 0.116 in 2010. Thus, while within detailed industries, the distribution of men and women across firms became more similar (contributing to the decline in the information index from 0.232 in 1992 to 0.218 in 2010), changes across industries offset that trend somewhat.

Consistent with our findings in Section 4, the sum of between sector and weighted between 3-digit and 4-digit NAICS industry segregation consistently accounted for more than half of the total segregation over the 18 year period (0.123 in 1992 and 0.116 in 2010). Of this between sector and

industry segregation, around two thirds is due to between sector segregation, with the contribution becoming less sizable as we consider segregation between more detailed industries.

Table 2 also shows that for each of the five years presented, the actual within-industry component of the information index for our sample is greater than the corresponding randomized value. This gap demonstrates that, on average, workers are less integrated than they would be if we randomly allocated them to firms within 4-digit NAICS industries, 3-digit NAICS industries, sectors, or within the U.S. The difference between the randomized value of the overall/sector/industry segregation and the actual value is the greatest when each firm's probability of hiring a worker is determined by the share of women within the coarsest partition of workers. The method for determining the share of women within a firm's hiring pool has a substantial effect on the interpretation of how different the observed within partition segregation is to a random allocation. On the one hand, if the share of women is determined by 4-digit NAICS industry female shares, the randomized within 4-digit industry segregation is roughly 60% of the observed actual within 4-digit industry segregation (compare, for instance, the randomized value of 0.063 in 1992 to its actual counterpart of 0.109). On the other hand, if the share of women is given by the national share, all segregation is considered within group segregation, and the randomized segregation is roughly one fourth of observed segregation for the U.S. (compare, for instance, a randomized segregation of 0.056 in 1992 to the observed segregation for the U.S. of 0.232). Notice also that as we move from within the U.S. segregation through to within 4-digit NAICS industry segregation, the observed value of segregation decreases (from 0.232 to 0.109 in 1992) as we increasingly eliminate systematic between sector and industry segregation, yet the randomized value of segregation increases slightly (from 0.056 to 0.063 in 1992).

All of this helps explain the observation in Table 2 that the four different randomizations give very different values for the normalized information index. If we are to believe that the mean share of women in the U.S. is a good proxy for the probability with which each firm would be allocated a female worker under a randomized process, the resulting normalized information index is over 3 times as large as the value we obtain when we fix firm female hiring probabilities at the mean share of women within 4-digit industries. This highlights very clearly the major influence that the choice of randomization process can have on the value of a normalized segregation index.

In Section 4.1 we showed that the rise in the education and health sectors and the fall in the manufacturing sector led to the presence of more female dominated firms within the U.S; but, more female firms does not necessarily imply less segregation. To help understand which sectors contribute most to the segregation of workers across firms, Table 3 shows the information index and

Table 3: **Information indices by sector**

Sector	Information index		Weighted info. index		Between contribution	
	1992	2010	1992	2010	1992	2010
Manufacturing	0.120	0.097	0.023	0.010	0.023	0.015
Retail	0.211	0.149	0.032	0.026	-0.001	0.000
Health care	0.151	0.124	0.011	0.011	0.024	0.030
Public administration	0.097	0.077	0.008	0.006	0.000	0.000
Education	0.059	0.062	0.004	0.006	0.003	0.009
Administrative and support services, management of companies	0.228	0.165	0.012	0.011	0.001	0.001
Professional services	0.199	0.216	0.010	0.014	-0.000	0.000
Wholesale	0.188	0.187	0.009	0.007	0.004	0.004
Construction	0.249	0.188	0.007	0.005	0.015	0.012
Finance	0.132	0.151	0.005	0.005	0.003	0.002
Other services	0.407	0.489	0.016	0.021	0.000	0.001
Information	0.093	0.072	0.002	0.003	-0.000	0.001
Transportation and warehousing	0.213	0.175	0.004	0.005	0.004	0.004
Other	0.174	0.161	0.007	0.006	0.005	0.002

Notes: The other category is made up of the remaining sectors: real estate and leasing; arts, entertainment, and recreation; mining; and utilities.

the weighted information index for each sector in 1992 and 2010, as well as the contribution that each sector makes to between sector segregation in the U.S. The sum of the weighted information indices for each sector gives the portion of the total information index for the U.S. that is attributable to within sector segregation (0.151 in 1992 and 0.136 in 2010), while the sum of the between contributions for each sector adds to 0.081 in 1992 and .082 in 2010. Each sector's contribution to between sector segregation is a measure of how that sector's entropy index (or diversity) differs to the entropy index of the country as a whole, weighted by employment. The manufacturing sector—a large and male dominated industry, with roughly 30% women and 21.5% employment in 1992—accounts for more than one quarter of between sector segregation. With no more than 3% employment, the transportation and warehousing sector plays a much smaller role in total between sector segregation, however, despite having an even lower share of women than manufacturing. Retail, with almost 15% employment, also plays a very small role in between sector segregation due it's gender share of workers that closely mimics the national average. In line with our discussion in Section 4.1 we observe that those sectors which contributed the most to changes in between sector segregation between 1992 and 2010 were manufacturing (and also construction)—lowering total segregation—and health care and education—raising segregation between sectors.

In contrast, the decline that we observe in within sector segregation in Table 2 is determined not by the share of women within industries, but the manner in which those women are distributed

across firms. Each sector's weighted information index—which, recall, is the sector information index weighted by the sector share of employment and the ratio of the sector entropy index to the national entropy index—informs us of the contribution each sector makes to total within sector segregation. Those sectors that make a large contribution will have some combination of a high information index, a high entropy index (indicating that the segregation is not simply due to between sector segregation which has already been accounted for), and high employment. The retail sector, for instance, with a moderately sized information index, a gender ratio that very closely reflects the female share of the workforce, and a large share of employment is the biggest contributor to segregation. Since segregation and diversity are frequently confused, this is not immediate; and it is interesting to note that this is despite being one of the smallest contributors to between sector segregation. Thus, we conclude that while the retail sector employs men and women in roughly equal shares, it does not distribute them evenly amongst firms. Additionally, the Other Services sector has a relatively small share of employment (3.8% and 4.5% in 1992 and 2010, respectively), but its contribution to segregation is quite high due to the very large information index and it's relatively high entropy index associated with a fairly gender balanced workforce of 55.9% and 59.1% women in 1992 and 2010, respectively. This is most likely explained by the fact that the Other Services sector is a somewhat arbitrary collection of industries that are likely to be quite different in nature and unlikely to have similar shares of women. Finally, we see that the fall in within sector segregation between 1992 and 2010 comes predominately from the fall in the information index of the retail sector, as well as the slight decline in the information index for the manufacturing sector and the large drop in employment in manufacturing, but is countered by the simultaneous rise in employment and information index for other services.

Our analysis thus far demonstrates the existence of systematic gender segregation in the U.S.; that differences between men and women in how they sort across industry generates a large share of this; and that there is only a the slight decline in segregation over the period 1992 to 2011. But, it does not address the implications of this segregation for workers. One major affect of gender differences in place of employment is that they may increase earnings disparities between men and women. Where differences in the distribution of employment are correlated with differences in pay, segregation will be associated with a gender gap in earnings. We examine this possibility in Section 5 by estimating the earnings gap in the U.S.

5 Evolution of the earnings gap

Existing studies of segregation (Bayard et al. (2003), Carrington and Troske (1998b), Carrington and Troske (1997)) document that on average men work for higher paying employers, and that this accounts for a substantial share of the gender gap. We estimate this effect on the gender earnings gap in the U.S. and compare it to the influence of the segregation of workers across industries. Our regression models are outlined in the next section and our estimation results follow. In Section 5.2, we estimate effects for the entire cross-section of workers at any given time, and then examine variation across birth cohorts and over the life-cycle of workers in Section 5.3.

5.1 Specifications for earnings regressions

To help describe our basic approach to measuring the importance of gender segregation in explaining the gender gap in earnings, consider the following statistical model of earnings:

$$y_{ijt} = \alpha + \gamma F_i + X'_{it}\beta + Z'_{jt}\phi + \epsilon_{ijt}$$

where y_{ijt} denotes individual i's log quarterly earnings at firm j in quarter t, F_i is a dummy variable indicating whether i is female, X_{it} denotes other individual characteristics, Z_{jt} denotes employer characteristics, and ϵ_{ijt} represents residual variation in earnings. We begin by estimating a regression of individual log earnings that excludes employer characteristics. X_{it} consists of:

- a series of dummy variables for single year of age, to control for predictable life-cycle effects
- controls for the individual's race, ethnicity, and whether they were foreign born
- dummies for state of employment to control for regional differences in earnings
- year/quarter dummies to control for overall trends and seasonal/cyclical variation in earnings.

Because we exclude employer characteristics, this estimate of the gender gap ($\hat{\gamma}_{base}$) will include any systematic effects on pay due to differences between men and women in the kinds of employers they work for. Note that there are no dramatic differences in demographic characteristics (aside from gender) between working men and women, so this base earnings gap differs little from what we obtain when we control only for age. We then compare $\hat{\gamma}_{base}$ to other estimates of γ from a series of additional regressions that each add additional controls.

The first adds job tenure (measured in quarters) to produce an estimate of the gap net of the effects of any gender differences in turnover patterns ($\hat{\gamma}_{tenure}$). The second variation also includes

job tenure, but further adds 6-digit industry dummies to control for differences in pay resulting from gender differences in industry of employment.[18]

We then add firm age and size to yield an estimate of the average gap in earnings between men and women who work for employers with the same observable characteristics ($\hat{\gamma}_{main}$). Finally, we add firm fixed effects to absorb the effects of mean differences in pay across employers ($\hat{\gamma}_{FE}$). We drop controls for location and industry in this specification because they are collinear with the firm fixed effects.

In this last specification, $\hat{\gamma}_{FE}$ is an estimate of the average within-firm gender gap in earnings between men and women with similar observable characteristics. The difference between $\hat{\gamma}_{FE}$ and $\hat{\gamma}_{tenure}$ represents the net effect of segregation patterns on the gender gap. The difference between $\hat{\gamma}_{FE}$ and $\hat{\gamma}_{main}$ represents the part of that effect that is associated with differences in pay across employers in the same industry and within the same firm size/age group.

As specified, this series of regressions produces estimates of the gender gap in earnings averaged across age groups and over the period 1992-2011. We are interested in changes over time, and present two different ways of characterizing those changes. First, we simply estimate the gap averaged across age groups at a point in time, but we allow the gap to change from quarter to quarter by adding an interaction term between F_i and a series of dummy variables for each quarter. If changes in employment practices over time have similar affects on workers at all stages in their careers, then this might be the most appropriate way of capturing the evolution of the gender earnings gap. But if, instead, change tends to affect recent labor market entrants more directly than those further along in their careers, comparing the gap at similar ages across different cohorts may be more informative. So our second set of estimates is based on dividing up our sample into birth cohorts, and estimating the series of regressions described above separately for each cohort. Because time and age are collinear within birth cohort, the sequence of earnings gaps estimated for a particular cohort combine the effects of age and of time (e.g. business cycle effects).

5.2 Trends in average gender gap over time

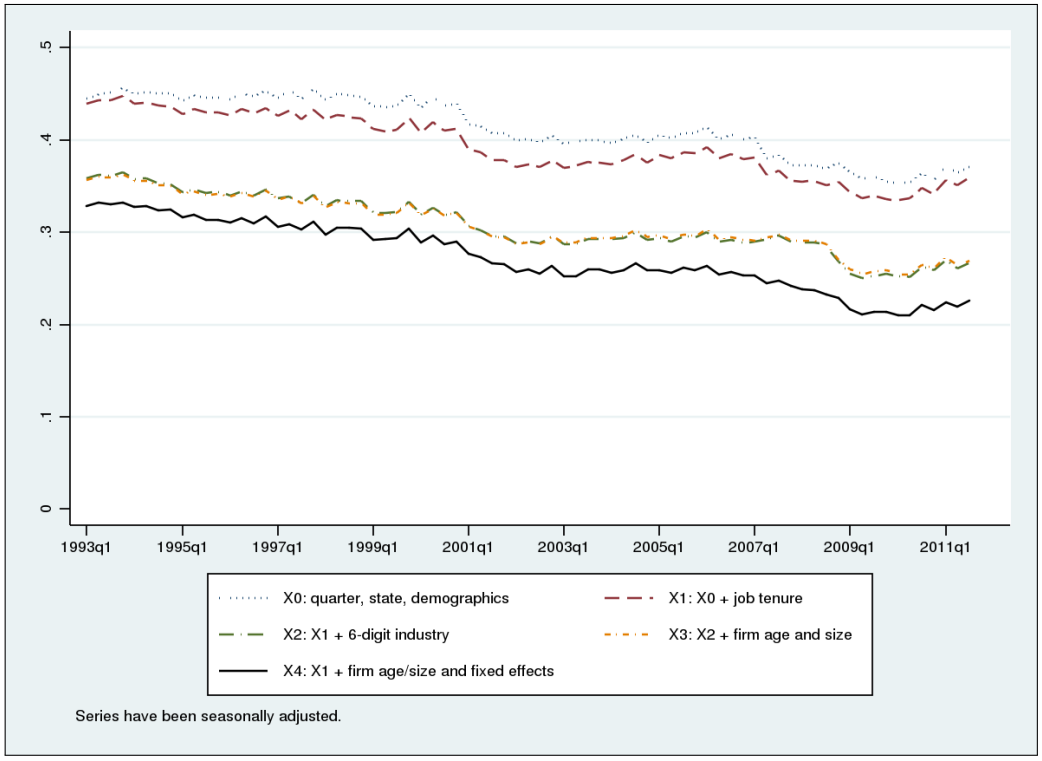
Figure 12 plots the results of our gender earnings gap regressions over the period from first quarter 1993 to third quarter 2011, while Table 4 gives point estimates for selected years in that interval. In each of the regressions we control for quarter, state, and demographics.[19] The dotted blue line

[18]The absence of firms in certain 6-digit industry/state/year cells makes it difficult to measure segregation but is of no consequence in the earnings regressions. It is for this reason that the earnings wage regressions use 6-digit NAICS controls, but our segregation measures depend only on sector, 3-digit, and 4-digit NAICS industries.

[19]Each individual's age is measured as the age at the end of the first quarter so that we compare men and women whose age differs by no more than one year.

shows that the (log) earnings gap between men and women of the same age within the same state falls from 0.46 to 0.36 (10 log points, or roughly one-fifth) over the 18 year time period. Adding controls for job tenure reduces the gap modestly, but explains none of the overall decline in the gap. This reduction in the gap reflects somewhat lower average tenure for women.[20] Controlling for detailed industry of employment has a much larger effect, but that effect is similarly stable over time. Firm size and age, while having significant effects on worker earnings levels, have essentially no effect on the gap because their averages are quite similar for men and women.[21]

Figure 12: Gender gap over time



The last two rows of the table give estimates of the effects of segregation on the earnings gap. The full effect reports the difference $\hat{\gamma}_{tenure} - \hat{\gamma}_{FE}$, giving the combined effect of gender differences in earnings associated with overall differences in the distribution of employment across firms. The within-industry effect gives the component associated with differences in the distribution of employment within very detailed industries. As Figure 12 suggests, gender differences in patterns of employment matter for earnings, but the majority of the effect on the gap comes from differences across rather than within industry. Gender differences in industry of employment explains roughly 20-25% of the overall wage gap: if the share of men and women in each 6-digit industry reflected

[20] Adda et al. (2011) document substantial career costs in Germany associated with taking time out to have children as well as reduced fertility for those women for whom the cost of a labor supply interruption is the greatest.

[21] This is also the case when we include firm age and size but *not* 6-digit industry – see Figure 18 in Appendix B.3.

Table 4: **Gender earnings gap over time**

	Year				
Specification	1993	1995	2000	2005	2010
Demographics only	0.46	0.45	0.45	0.41	0.36
Add tenure	0.45	0.44	0.42	0.39	0.35
Add 6-digit NAICS	0.37	0.35	0.33	0.30	0.27
Add firm size and age	0.37	0.35	0.33	0.30	0.27
Add firm fixed effects	0.34	0.32	0.30	0.27	0.22
Segregation effects					
Full	0.11	0.11	0.12	0.12	0.12
Within industry	0.03	0.03	0.03	0.04	0.04

Notes: Figures represent log differences in earnings averaged across quarters of a year. Demographic characteristics included are indicators for foreign birth, black, Hispanic, Asian, and other race. Regressions also include state dummies, except where firm fixed effects are included. Industry dummies are dropped when firm fixed effects are included. The row labeled 'Full' gives the difference between 2nd and 5th row estimates, the row labeled 'Within industry', between 4th and 5th rows.

the overall share of women in the labor force, the gap would close by .08 to .09 log points in each of the five years represented in Table 4.

These estimates also make clear that these components have little to do with the *decline* in the earnings gap. The portion of the gender gap that is explained by job tenure, industry, firm size and age, and firm fixed effects is approximately constant over time. Thus, the unexplained portion of the gap is driving this decline, and that represents within-firm differences in earnings. While a decline in gender wage discrimination (that is, paying otherwise identical men and women different salaries) could explain these results, so too could many other variables not observed in our data. Card et al. (2013) showed that women do not reap the full monetary benefits of moving to firms with higher fixed effects that men do, and argued that 10 to 15% of the gender wage gap in Portugal can be attributed to differences in the bargaining power of men and women. Bertrand et al. (2010) find that including hours worked in earnings regressions for Chicago MBAs reduces the gender pay gap significantly, and that the difference in working hours for men and women 10 or more years out is striking for MBAs: only 4% of men are working part time, while the figure for women is 22%. The shift of women from traditionally female occupations to higher paid, formerly male-dominated occupations within industries like health care and law is also likely to reduce within-firm wage gaps.[22] In future work we hope to estimate the effect of hours on the earnings gap by considering the subset of workers for which survey data on occupation and hours are available.

Since educational attainment is a major predictor of earnings, and gender differences in education have shifted substantially over time, we might expect education to explain a sizable proportion of

[22]For instance, nurses are paid less than doctors and paralegals are paid less than attorneys.

Table 5: **Gender earnings gap controlling for education**

Specification	Year 1993	1995	2000	2005	2010
Full education sample					
Demographics only	0.50	0.50	0.51	0.50	0.46
Add education	0.50	0.51	0.52	0.50	0.46
Add tenure	0.50	0.49	0.48	0.47	0.44
Add 6-digit NAICS	0.38	0.37	0.36	0.34	0.32
Add firm size and age	0.38	0.37	0.36	0.35	0.32
Add firm fixed effects	0.36	0.35	0.34	0.32	0.29
Subtract firm FEs from dep var	0.36	0.35	0.34	0.32	0.29
Segregation effects					
Full	0.14	0.14	0.14	0.15	0.15
Within industry	0.02	0.02	0.02	0.02	0.03
High school or less					
Demographics only	0.51	0.53	0.52	0.50	0.46
Add education	0.52	0.54	0.52	0.51	0.46
Add tenure	0.51	0.51	0.48	0.47	0.43
Add 6-digit NAICS	0.35	0.35	0.31	0.30	0.26
Add firm size and age	0.35	0.35	0.32	0.30	0.26
Add firm fixed effects	0.37	0.37	0.35	0.33	0.28
Subtract firm FEs from dep var	0.35	0.35	0.34	0.31	0.27
Segregation effects					
Full	0.14	0.15	0.13	0.14	0.15
Within industry	-0.01	-0.02	-0.03	-0.02	-0.02
Some college or more					
Demographics only	0.49	0.50	0.52	0.51	0.47
Add education	0.48	0.49	0.51	0.49	0.46
Add tenure	0.48	0.47	0.49	0.47	0.44
Add 6-digit NAICS	0.38	0.37	0.37	0.36	0.34
Add firm size and age	0.37	0.36	0.37	0.36	0.34
Add firm fixed effects	0.35	0.34	0.34	0.32	0.30
Subtract firm FEs from dep var	0.35	0.34	0.33	0.32	0.30
Segregation effects					
Full	0.13	0.14	0.15	0.15	0.15
Within industry	0.03	0.03	0.03	0.04	0.05

Notes: Figures represent log differences in earnings averaged across quarters of a year. Demographic characteristics included are indicators for foreign birth, black, Hispanic, Asian, and other race. Regressions also include state dummies, except where firm fixed effects are included. Industry dummies are dropped when firm fixed effects are included. Firm fixed effects estimated from the full sample are included as a regressor in the row labeled "Add firm fixed effects". The row labeled "Full' gives the difference between the 3rd and 6th row estimates, the row labeled "Within industry", the difference between the 5th and 6th rows.

the earnings gap. While we do not have education data for all individuals in our sample, we do have information for a reasonably large subsample on whether they have less than a high-school education, graduated from high school, have some college education, or have a college degree. We make use of this subsample to examine whether controlling for education affects our general findings about the importance of segregation for the earnings gap. Note that because the education data were mostly collected during the 1990 and 2000 population censuses and is only used if the respondent was at least 25 years of age when they provided the information, the subsample with education measures is older than the overall sample.

Table 5 shows that adding dummies for education categories to earnings regressions–when using the sample of workers for which we have educational data–has essentially no effect on the earnings gap.[23] When we split the sample into workers who did and did not enter college, we observe two very small effects that in the aggregate cancel each other out. For workers who did not enter college, adding controls for education slightly *increases* the earnings gap. This is not too surprising given that of those workers who do not attend college, women are more likely to graduate from high school. Thus, for high school educated workers, the raw gender earnings gap actually underestimates that part of the gap in earnings for men and women in the same education category that is not explained by education. In contrast, women have historically been somewhat less likely than men to graduate from college so we would expect women to be less educated than men, especially at the beginning of our time series. Here we see that the raw gender earnings gap narrows slightly when we control for education. In sum, the effect of education is surprisingly small. Including education controls does not change our finding that gender differences in industry of employment accounts for the largest share of the gap. Nonetheless, whether including education controls affects our conclusions about the importance of differences within industry is complicated by issues with estimating fixed effects using a sample of workers. In firms with small samples, a worker's own earnings will have a large influence on the estimate of the firm fixed effect if we simply use the same approach used when we have all workers. As an alternative, we estimate the fixed effect using the full sample, and present two ways of controlling for it: including the fixed effect as a covariate (row 6 in each panel of Table 5), and using log(earnings) minus the estimated fixed effect as the dependent variable (row 7 in each panel of Table 5). The latter specification is equivalent to constraining the coefficient on the fixed effect to be equal to one. The estimated coefficient on the fixed effect is very close to one when we include it as a covariate in the full education sample (1st panel) and some-college sample

[23]Note, however, that the overall gender gap does differ somewhat (compare, for instance, the gender gap of 0.50 when we add only demographics in 1993 in the education sample to the correspoding gap of 0.46 in the full sample). Since we do not include education dummies in the first regression, this reflects differences in the sample, most probably due to the fact that workers in this sample are at least 25 years of age.

(3rd panel), so it is unsurprising that the results from the two approaches are very similar in those panels. With the high-school-or-less sample, the estimated coefficient is somewhat lower than one, so the specifications give slightly different results. We find somewhat smaller effects when we add firm fixed effects with this specification and sample, but it is not clear whether this is the result of including education controls or of changing the sample.[24]

The earnings gap in any given year is a weighted average of a number of different cohorts of workers of various ages and experience levels. Although the wage gap may be declining over time, this does not necessarily mean that the earnings gap is falling over time within a fixed group. In the next section we partition workers into cohorts to examine the gap over the life-cycle of workers and how that has changed across cohorts.

5.3 Trends in the gender gap within cohorts of workers

Figure 13 and Table 6 show the earnings gap for five generations of workers: those born in the 40s, 50s, 60s, 70s, and those born in the early 1980s (1980-1982). Figure 13 plots the earnings gap for each year between the ages of 18 and 60 for which we have data for that cohort, while Table 6 presents point estimates for selected ages spaced out over the working years in which we observe members of each cohort. Looking at any single cohort in Figure 13, we find the same general effects of adding our controls that we observed in Figure 12. A substantial portion of the earnings gap is explained by our controls, but a large portion remains unexplained. We also find that the gender earnings gap decline across cohorts: at any given age, the gap is larger for cohorts born in earlier decades. For each of the cohorts studied, we see that once workers reach their early 20s, the wage gap rises over time and eventually peaks. The age at which the peak occurs (and the steepness of the curve before and after the peak) varies across cohort, with some suggestion that among more recent cohorts the gap has peaked at earlier ages, as well as lower levels.[25] The sharp rise and fall in the gender earnings gap between 18 and 22 (when a large share of young workers are enrolled in college) is striking. The pattern suggests that, among the less-skilled workers who participate in the labor market at those ages, men seem to have a sizeable advantage that falls when the college educated enter the labor market (and are included in the average gap). This seems likely to result from work in physically demanding blue collar jobs, in which young men have an advantage. While beyond the scope of this paper, an interesting question is whether this wage premium among young men contributes to lower rates of college completion for men.

[24] In future work we plan to estimate the earnings gap for our education sample excluding education controls so as to isolate any sampling effects before analyzing the effects of education.

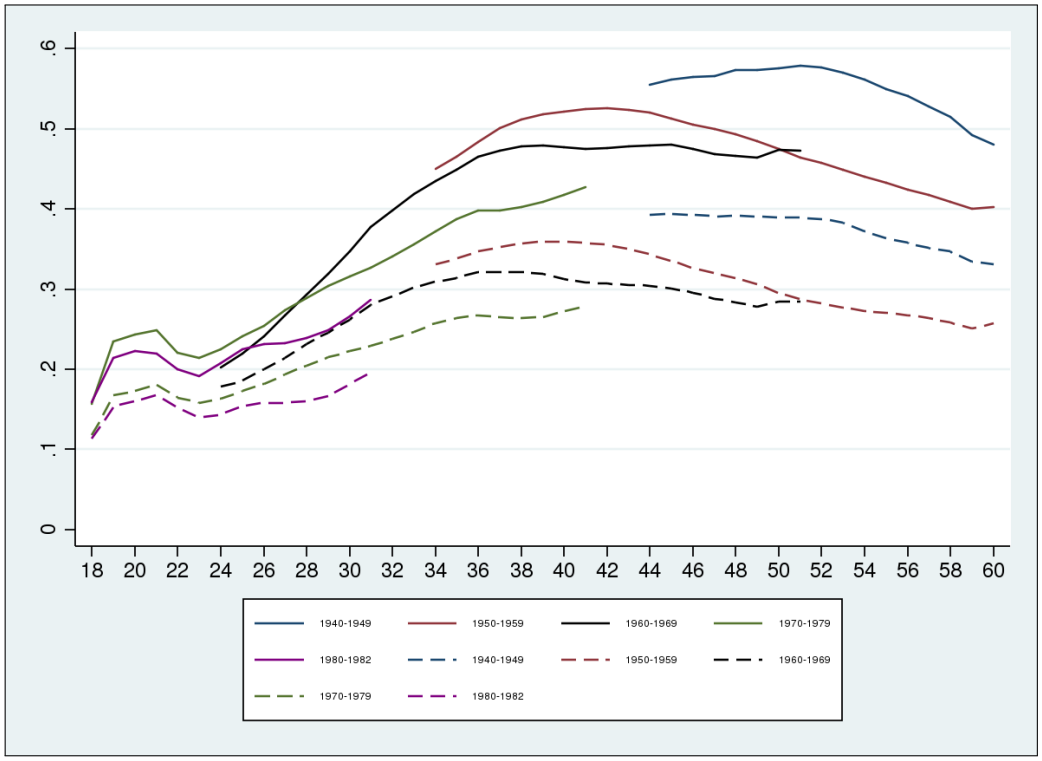
[25] See Goldin (2014) for life cycle earnings gaps that span 40 years.

Table 6: **Gender earnings gap by cohort and age**

	Age								
Birth years	18	24	30	35	40	45	50	55	60
Controlling only for demographics									
1940-1949						0.56	0.58	0.55	0.48
1950-1959				0.47	0.52	0.51	0.47	0.43	0.40
1960-1969		0.20	0.35	0.45	0.48	0.48	0.47		
1970-1979	0.16	0.23	0.32	0.39	0.42				
1980-1982	0.16	0.21	0.27						
Adding tenure									
1940-1949						0.55	0.55	0.52	0.47
1950-1959				0.45	0.50	0.48	0.44	0.41	0.40
1960-1969		0.20	0.34	0.43	0.45	0.45	0.44		
1970-1979	0.16	0.22	0.31	0.38	0.41				
1980-1982	0.16	0.21	0.27						
Adding 6-digit NAICS controls									
1940-1949						0.47	0.47	0.44	0.40
1950-1959				0.38	0.41	0.39	0.35	0.33	0.33
1960-1969		0.17	0.28	0.35	0.36	0.35	0.33		
1970-1979	0.10	0.15	0.23	0.29	0.31				
1980-1982	0.09	0.11	0.17						
Adding firm fixed effects									
1940-1949						0.39	0.39	0.36	0.33
1950-1959				0.34	0.36	0.34	0.30	0.27	0.26
1960-1969		0.18	0.26	0.31	0.31	0.30	0.28		
1970-1979	0.12	0.16	0.22	0.26	0.27				
1980-1982	0.11	0.14	0.18						
Full segregation effects									
1940-1949						0.15	0.16	0.16	0.14
1950-1959				0.12	0.14	0.14	0.14	0.14	0.14
1960-1969		0.02	0.08	0.11	0.13	0.15	0.16		
1970-1979	0.04	0.06	0.09	0.12	0.13				
1980-1982	0.05	0.06	0.09						
Within industry segregation effects									
1940-1949						0.08	0.08	0.08	0.07
1950-1959				0.04	0.05	0.06	0.06	0.06	0.07
1960-1969		-0.01	0.01	0.03	0.04	0.05	0.05		
1970-1979	-0.01	-0.02	0.01	0.03	0.04				
1980-1982	-0.02	-0.03	-0.01						

Notes: Figures represent log differences in earnings averaged across quarters of a year for single-year birth cohorts, giving each birth cohort that contributes data equal weight. Demographic characteristics included are indicators for foreign birth, black, Hispanic, Asian, and other race. Regressions also include state dummies, except where firm fixed effects are included. Firm fixed effects are estimated on a sample that includes all full-quarter employees aged 18-60, with controls for worker demographics, and tenure. Those fixed effects are then included as a regressor in the birth-year specific regressions. Subtracting the fixed effect from workers log earnings gives essentially the same estimated gap. The row labeled "Full' gives the difference between the 1st and 3rd row estimates, the row labeled "Within industry" gives the difference between the 3rd and 4th rows.

Figure 13: Demographics vs. Firm FE (and all else)



6 Further work and extensions

There are a couple of extensions to our earnings regression framework that we hope to address in future work. First and foremost we hope to understand the role that hours and occupation play in generating the earnings gap. Hours, in particular, are likely to explain a significant portion of the within-firm differences in pay between men and women. We do not have these variables for the full LEHD sample of individuals, but can measure them for a subset of individuals who are included in Census household surveys.

In the absence of occupation controls, our estimated indices of segregation are likely to overestimate the extent to which the men and women work side by side. In many industries clear divisions of communication, hierarchy, and earnings exist between workers of different occupation. Much of the integration of men and women in firms in the health care sector, for instance, may be between nurses who are more often than not female and doctors who are still predominantly male. Similarly, women in the manufacturing sector are more likely to fill administrative roles, while men are more likely to be engineers, for instance. An implication of this is that the integration observed in the health care industry is potentially masking divisions of workers by earnings and education, while the integration in the manufacturing sector may be overstating the extent to which men and women

communicate or work in teams. Additional work that decomposes our empirical density functions and information indices into between and within occupational segregation would provide insight into the extent to which men and women are segregated between teams or groups of workers within firms. Moreover, an analysis of this effect over time would provide insight into how women's movement from lower paid occupations to higher paid occupations within industries has contributed to within-group integration. The small size of the sample of workers for whom we have occupational data means that this may not be achievable using our current data set, but it may nonetheless be an interesting topic for future research for academics with more detailed, perhaps European, data.

Finally, an important non-pecuniary benefit of some but not all jobs is hours flexibility, which on average is likely to be particularly valuable to women with children at home. Recent papers by Blau and Kahn (2013) and Goldin (2014) both argue that the demand for and supply of jobs that can accommodate the time demands of child rearing play an important role in determining the evolution of gender differences in careers, though the two papers differ by focusing on the role of national policies (Blau and Kahn (2013)) versus differences in employer practices (Goldin (2014)).[26] In future work we hope to consider an analysis of the subsample of workers for whom we have additional information on human capital and job characteristics to consider whether the association between personal characteristics and workplace segregation appear consistent with an important role for workplace flexibility in generating the observed sorting across employers. To determine the extent to which women's tendency to choose more flexible occupations (such as pharmacy, see Goldin and Katz (2012)) may be contributing to workplace segregation and the gender earnings gap, we hope to follow from Goldin (2014) and use O*NET data to generate an index of flexibility for each occupation. We can then split industries into low, medium, and high hours flexibility groups and compare the levels of segregation and run our earnings regressions with flexibility indices as a control variable.

7 Conclusion

Women's entry into the workforce over the second half of the 20th century has been a game changer, and the economics literature documents this extensively. But the absence of comprehensive linked employer-employee datasets has left a gap in our understanding of gender segregation across U.S. firms. Using LEHD data, we uncover three novel findings regarding segregation. First, we build on

[26]See also Goldin and Katz (2011) for a discussion of how workplace flexibility, and the pecuniary penalties associated with it, varies by occupation; and, Goldin and Katz (2012) for insight into the structural change in the pharmacy sector which has led it to evolve into a profession where workplace flexibility comes at very little cost, resulting in a very female dominated workforce.

previous studies of gender segregation to confirm that systematic gender segregation between firms that has been shown to exist for particular industries or locations also holds for a more general sample. We demonstrate, however, that the choice of reference group used to construct randomized segregation measures has substantial effects on conclusions about the extent of segregation, and our findings suggest that, as a result, earlier studies may have over-estimated systematic segregation. Second, we find that gender segregation in the U.S. has declined modestly over the period 1992 to 2010. Third, we highlight the important role that industry –and, in particular, sector–of employment plays in the segregation of men and women across firms.

Our earnings regressions uncover three findings of a similar nature. We show that industry choice accounts for a large proportion of the gender earnings gap, but that firm choice within industry also plays a significant role. We also document a substantial decrease in the earnings gap, though our set of explanatory variables do little to explain this decline. The earnings gap declines even when we include employer fixed effects, so the decrease reflects a narrowing in the gap between men's and women's earnings when working for the same employer. Finally, we find that, among recent cohorts, the earnings gap is smaller and may peak at somewhat younger ages and lower levels than was the case for older cohorts.

A Proofs

A.1 Decomposing the information index

Rearranging (1) gives

$$\begin{aligned}
H(b,a,US) &= \sum_{c=1}^{C}\sum_{j=1}^{J}\frac{T(j,c,b,a,US)}{T(b,a,US)E(b,a,US)}\Big(E(b,a,US)-E(c,b,a,US)\Big) \\
&\quad +\sum_{c=1}^{C}\sum_{j=1}^{J}\frac{T(j,c,b,a,US)}{T(b,a,US)E(b,a,US)}\Big(E(c,b,a,US)-E(j,c,b,a,US)\Big) \\
&= \sum_{c=1}^{C}\frac{T(c,b,a,US)}{T(b,a,US)E(b,a,US)}\Big(E(b,a,US)-E(c,b,a,US)\Big) \\
&\quad +\sum_{c=1}^{C}\frac{T(c,b,a,US)E(c,b,a,US)}{T(b,a,US)E(b,a,US)}\Bigg\{ \\
&\quad \sum_{j=1}^{J}\frac{T(j,c,b,a,US)}{T(c,b,a,US)E(c,b,a,US)}\Big(E(c,b,a,US)-E(j,c,b,a,US)\Big)\Bigg\} \\
&= H_{\text{between}}(b,a,US)+\sum_{c=1}^{C}\frac{T(c,b,a,US)E(c,b,a,US)}{T(b,a,US)E(b,a,US)}H(c,b,a,US)
\end{aligned}$$

where the last line follows from (2) and (3).

A.2 Relationship between randomized information indices

Consider a 3-digit industry b. For each 4-digit industry c, randomly allocate workers to firms such that the probability that each firm's new employee is a woman is given by the share of women within the 4-digit industry in which that firm is located. Calculate the resulting information indices for each 4-digit industry c and the 3-digit industry b. Repeat N times, and on the nth randomization call these information indices $H(c,b,a,US)_4^n$ and $H(b,a,US)_4^n$, respectively. We know from Appendix A.1 that

$$H(b,a,US)_4^n = H_{\text{between}}(c,b,a,US) + \sum_{c=1}^{C} \frac{T(c,b,a,US)}{T(b,a,US)E(b,a,US)} H(c,b,a,US)_4^n$$

Let $H(b,a,US)_4^*$ be the randomized value of the information index for 3-digit sector b when workers are allocated using 4-digit industry gender shares as described above. By definition, $H(c,b,a,US)_4^* = \frac{1}{N}\sum_{n=1}^{N} H(c,b,a,US)_4^n$ and $H(b,a,US)_4^* = \frac{1}{N}\sum_{n=1}^{N} H(b,a,US)_4^n$. So,

$$\begin{aligned} H(b,a,US)_4^* &= \frac{1}{N}\sum_{n=1}^{N}\left(H_{\text{between}}(c,b,a,US) + \sum_{c=1}^{C} \frac{T(c,b,a,US)}{T(b,a,US)E(b,a,US)} H(c,b,a,US)_4^n \right) \\ &= H_{\text{between}}(c,b,a,US) + \sum_{c=1}^{C} \frac{T(c,b,a,US)}{T(b,a,US)E(b,a,US)} H(c,b,a,US)_4^* \end{aligned} \tag{13}$$

It can similarly be shown that $H(a,US)_4^*$ – the randomized value of the information index for sector a when workers are randomly allocated using 4-digit industry gender shares – can be decomposed into between sector segregation and a weighted sum of randomized within sector segregation. Thus, we have that

$$H(a,US)_4^* = H_{\text{between}}(b,a,US) + \sum_{b=1}^{B} \frac{T(b,a,US)}{T(a,US)E(a,US)} H(b,a,US)_4^* \tag{14}$$

Substituting (13) into (14) gives

$$\begin{aligned} H(a,US)^* = {} & H_{\text{between}}(b,a,US) + \sum_{b=1}^{B} \frac{T(b,a,US)}{T(a,US)E(a,US)} H_{\text{between}}(c,b,a,US) \\ & + \sum_{b=1}^{B} \frac{T(b,a,US)}{T(a,US)E(a,US)} \sum_{c=1}^{C} \frac{T(c,b,a,US)}{T(b,a,US)E(b,a,US)} H(c,b,a,US)_4^* \end{aligned}$$

Iterating this process one more time gives (11). An analogous argument shows that each of (9) and (10) hold.

B Additional graphs

B.1 Segregation trend fixing the employment distribution across sectors at 1992 levels

Figure 14: Information index reweighted to preserve 1992 industry sector distribution

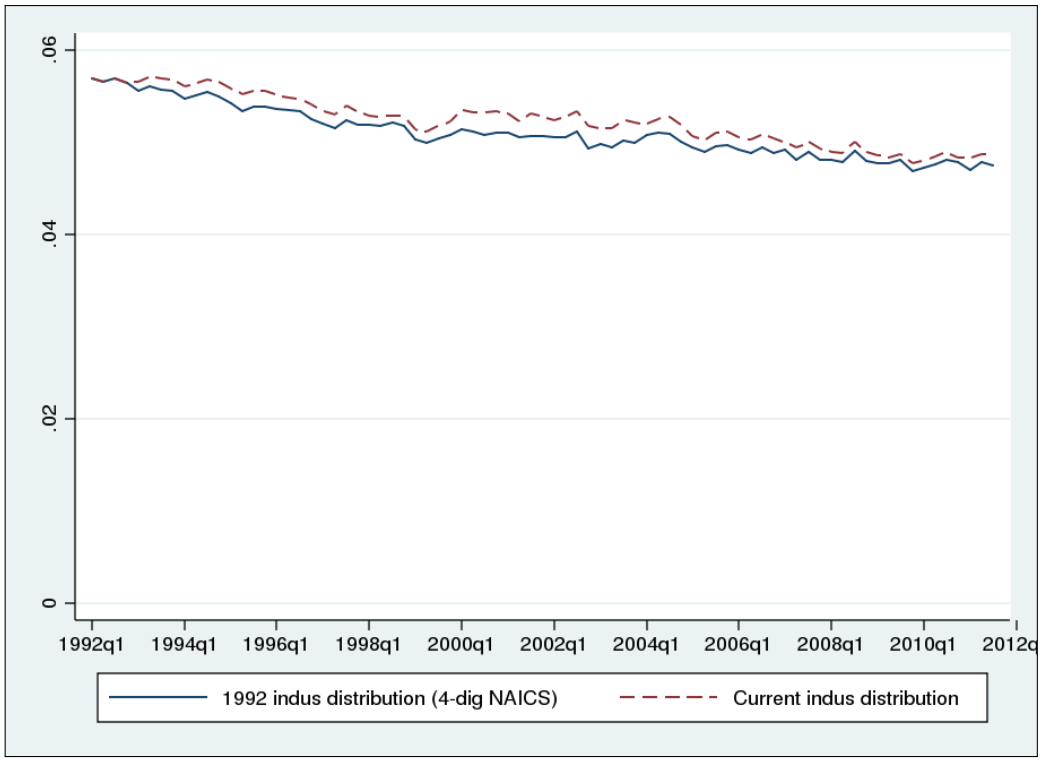


Figure 14 plots two versions of the normalized information index from 1992 to 2012. The dashed red line plots our estimates using the current employment distribution in each quarter (equal to solid blue line in Figure 10) and the solid blue line reweights each firm by multiplying it by the ratio of 1992 to 2010 employment for the firm's 4-digit industry. When industry weights are fixed at 1992 levels the decline in the normalized information index is more substantial than when industry weights reflect the current situation, indicating that the decline in segregation within industries has been somewhat offset by a shift in employment towards more segregated industries.

B.2 Randomized segregation for different industry partitions

Figure 15: Randomized segregation for different industry partitions 1992

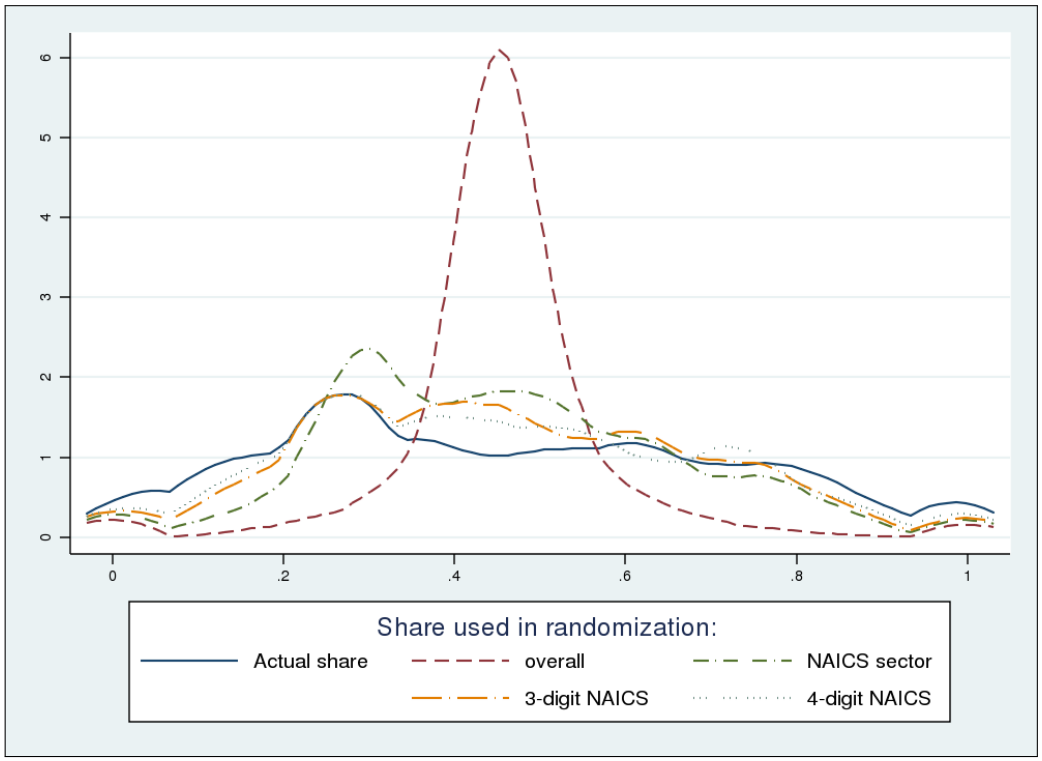


Figure 16: Randomized segregation for different industry partitions 2002

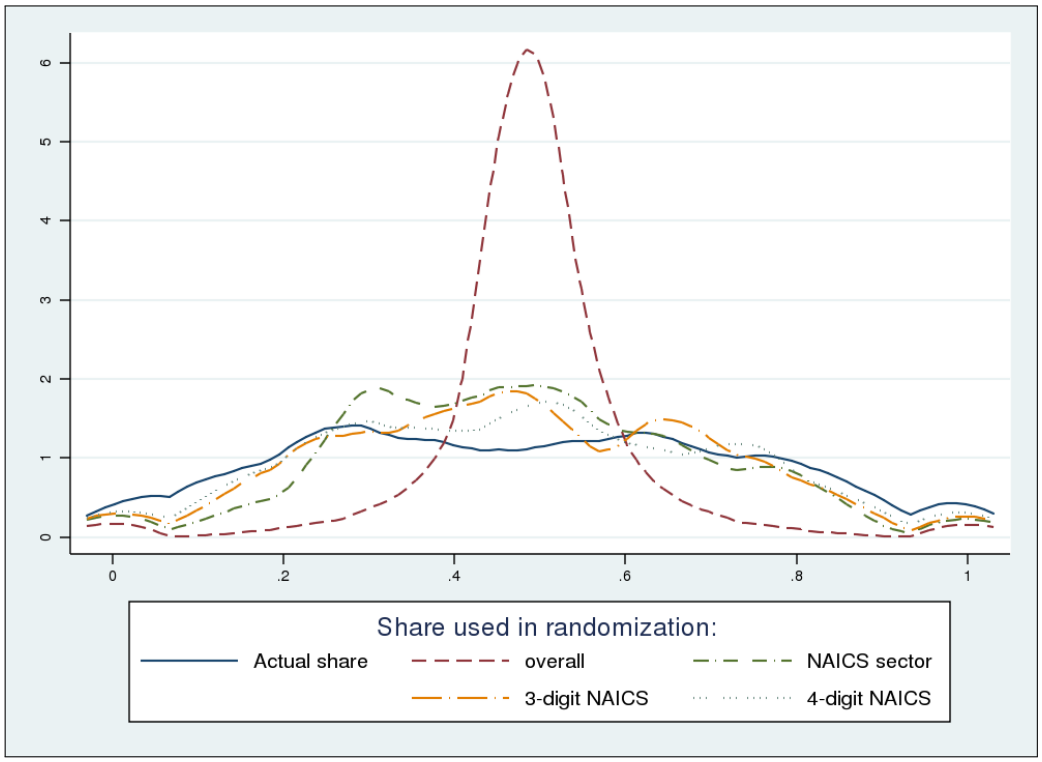
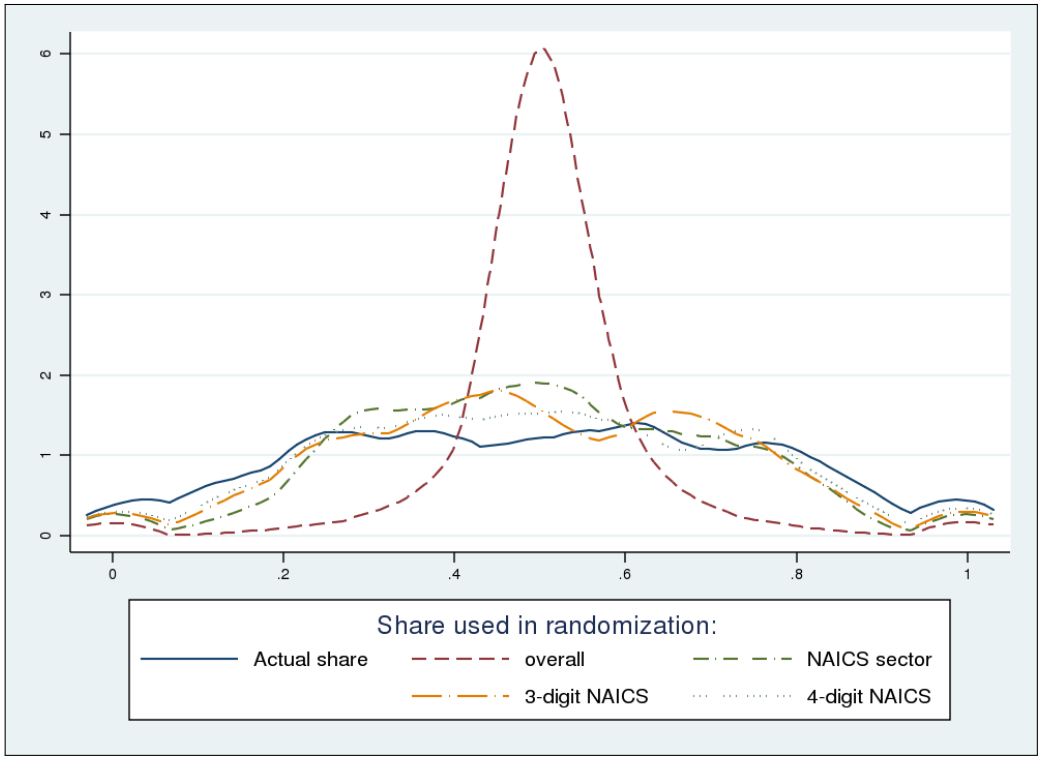
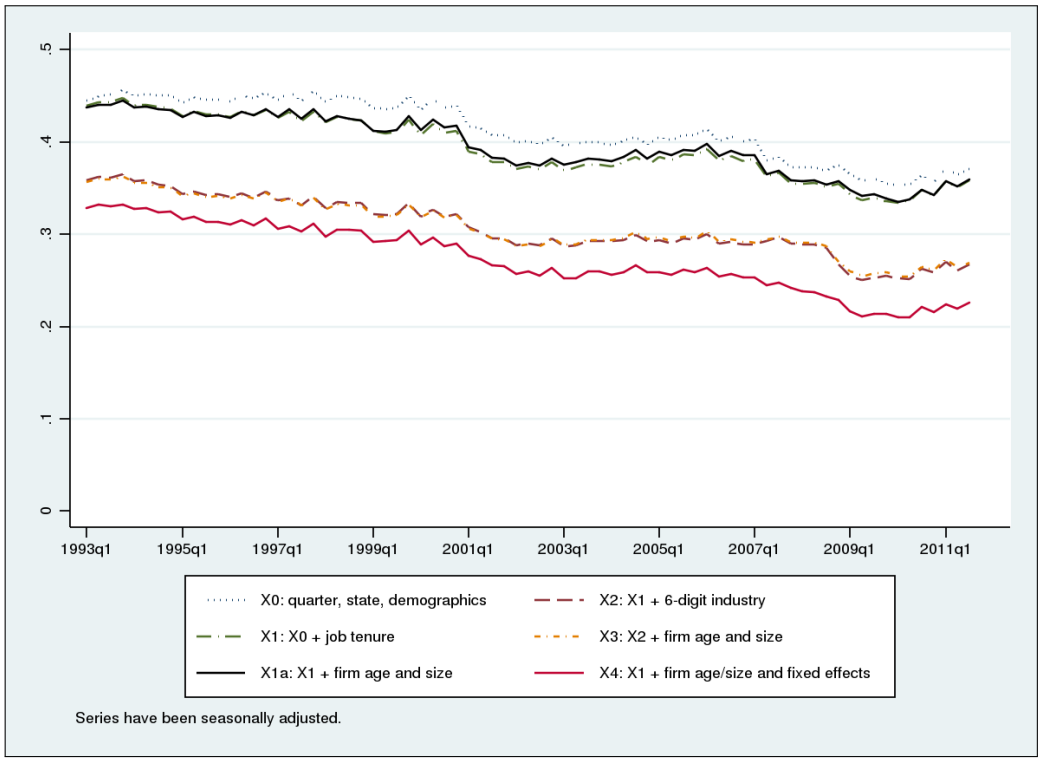


Figure 17: Randomized segregation for different industry partitions 2010



B.3 Alternate gender earnings graph

Figure 18: Gender gap over time – alternate ordering of control variables



References

John M. Abowd, Bryce E. Stephens, Lars Vilhuber, Fredrik Andersson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock. The LEHD infrastructure files and the creation of the quarterly workforce indicators. In *Producer Dynamics: New Evidence from Micro Data*, pages 149–230. University of Chicago Press, 2009.

Jérôme Adda, Christian Dustmann, and Katrien Stevens. The career costs of children. December 2011.

Kimberly Bayard, Judith Hellerstein, David Neumark, and Kenneth Troske. New evidence on sex segregation and sex differences in wages from matched employee-employer data. *Journal of Labor Economics*, 21(4), 2003.

Marianne Bertrand, Claudia Goldin, and Lawrence F. Katz. Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American Economic Journal: Applied Economics*, 2(3): 228–255, July 2010.

Francine D. Blau and Lawrence M. Kahn. Gender differences in pay. Unpublished, June 2000.

Francine D. Blau and Lawrence M. Kahn. The US gender pay gap in the 1990s: Slowing convergence. *Industrial and Labor Relations Review*, 60(1):45–66, October 2006.

Francine D. Blau and Lawrence M. Kahn. Female labor supply: Why is the US falling behind? Unpublished, January 2013.

Dale Boisso, Kathy Hayes, Joseph Hirschberg, and Jacques Silber. Occupational segregation in the multidimensional case. *Journal of Econometrics*, 61:161–171, 1994.

United States Census Bureau. More working women than men have colleage degrees, April 2011. URL `https://www.census.gov/newsroom/releases/archives/education/cb11-72.html`.

Magnus Bygren. Unpacking the causes of segregation across workplaces. *Acta Sociologica*, 56(1):3–19, 2013.

David Card, Ana Rute Cardoso, and Patrick Kline. Bargaining and the gender wage gap: A direct assessment. August 2013.

William J. Carrington and Kenneth R. Troske. Gender segregation in small firms. *The Journal of Human Resources*, 30(3):503–533, Summer 1995.

William J. Carrington and Kenneth R. Troske. On measuring segregation in samples with small units. *Journal of Business & Economic Statistics*, 15(4):402–409, 1997.

William J. Carrington and Kenneth R. Troske. Interfirm segregation and the black/white wage gap. *Journal of Labor Economics*, 16(2):231–260, April 1998a.

William J. Carrington and Kenneth R. Troske. Sex segregation in U.S. manufacturing. *Industrial and Labor Relations Review*, 51(3):445–464, April 1998b.

Otis Dudley Duncan and Beverly Duncan. A methodological analysis of segregation indexes. *American Sociological Review*, 20(2):210–217, 1955.

Claudia Goldin. A grand gender convergence: Its last chapter. *The American Economic Review*, 104(4): 1091–1119, 2014.

Claudia Goldin and Lawrence F. Katz. The cost of workplace flexibility for high-powered professionals. *The Annals of the American Academy*, 638:45–67, November 2011.

Claudia Goldin and Lawrence F. Katz. The most egalitarian of all professions: Pharmacy and the evolution of a family-friendly occupation. September 2012.

Robert Hutchens. One measure of segregation. *International Economic Review*, 45(2):555–578, 2004.

Robert M. Hutchens. Segregation curves, lorenz curves, and inequality in the distribution of people across occupations. *Mathematical Social Sciences*, 21:31–51, 1991.

Abraham Mosisa and Steven Hipple. Trends in labor force participation in the united states. *Monthly Labor Review*, pages 35–57, October 2006.

Sean F. Reardon and Glenn Firebaugh. Measures of multigroup segregation. *Sociological Methodology*, 32 (1):33–67, 2002.

Jacques Silber. Factor components, population subgroups and the computation of the gini index of inequality. *The Review of Economics and Statistics*, 71(1):107–115, February 1989a.

Jacques Silber. On the measurement of employment segregation. *Economics Letters*, 30:237–243, 1989b.

Jacques Silber. Occupational segregation indices in the multidimensional case: A note. *The Economic Record*, 68(2):276–277, September 1992.

Henri Theil. *Statistical Decomposition Analysis. With applications in the social and administrative sciences.* North-Holland Publishing Company, 1972.