ECONOMIC GRAPH RESEARCH INSTITUTE

WORKING PAPER NO. 6 | JULY 2026

Generalized Occupational Segregation Index: International Comparisons Over Time by Gender and Education

Matthew Baird, LinkedIn Economic Graph Research Institute

Nadine Fares, LinkedIn Economic Graph Research Institute*

Kory Kantenga, LinkedIn Economic Graph Research Institute

Danielle Kavanagh-Smith, LinkedIn Economic Graph Research Institute

Generalized Occupational Segregation Index: Global Comparisons Over Time by Gender and Education*

Matthew Baird† Nadine Fares‡ Kory Kantenga§
Danielle Kavanagh-Smith¶

July 23, 2026

Abstract

Occupational segregation, or the unequal likelihoods of members of one group to work in given occupations compared to other groups, has been studied for decades as a metric of equity. However, prior work has focused on the binary case (e.g., gender) and disregarded sorting across firms. In this paper, we develop a generalized occupational segregation index which can take any number of groups and is first estimated at the firm level. We show how the national average can be decomposed into within- and between-firm segregation. We also show how to normalize these metrics, and to test the null hypothesis of the firm employing workers in ways representative to the target populations. We estimate these indices of gender and educational attainment using LinkedIn data across twelve countries and ten years. We find the majority of firms are occupationally segregated, both for education and gender. Gender segregation is more pronounced and common than education. While some industries are more segregated than others, including Construction and Healthcare, the variation in firms' occupational segregation within any industry is substantial. Firms with higher occupational segregation tend to have lower median pay in job postings, greater pay dispersion, and fewer applicants per vacancy, consistent with predictions from crowding models of labor market discrimination.

Keywords: Labor, equity, occupational segregation
JEL Codes: J16, J24, D63

---

*We would like to thank participants at the Association for Public Policy Analysis and Management annual conference as well as LinkedIn Economics Workshop Series for valuable feedback, as well as Silvia Lara for review and feedback.

†Senior Staff Economist, LinkedIn Economic Graph Research Institute. Email: mdbaird@linkedin.com
‡Staff Data Scientist, Figma. All work done while at LinkedIn
§Staff Data Scientist, LinkedIn Economic Graph Research Institute
¶Senior Data Scientist, LinkedIn Economic Graph Research Institute

1 Introduction

Equality in the workforce can take on many facets. These may include outcomes such as equal pay and benefits, representation in leadership positions (Lara et al., 2023), and access to opportunities to develop skills and advance in careers, among others. An additional consideration is the occupation in which people work. Not only is this a potential direct outcome of interest–occupational representation can increase exposure to people outside one's own group and increase productive exchange of ideas–it is an important driver of many other outcomes. For example, the gender pay gap has repeatedly been shown to be partially due to differences in the occupations men and women work in (Clarke, 2020). So too people's pathway into management and leadership can be driven by the upstream choices of which occupations people work in.

Which occupation a person works in is a complicated function of personal preferences, environmental situations and opportunities, and the existence or absence of barriers to entry and progression. Each of these may differ depending on the groups a worker is part of, including gender, race, and educational attainment. Thus, while economists often frame occupational sorting as the result of the worker's choice, there is also a clear understanding that environment and barriers may shift workers away from the outcome they would have had in absence of those factors. One common example of this is examining gender differences in Science, Technology, Engineering, and Mathematics (STEM). Women are consistently underrepresented in STEM across the globe (Baird et al., 2023). It has been shown that barriers all along the educational and employment process have reinforced that gap. Girls in school may receive less encouragement to pursue STEM education than boys; they also face an educational and work environment wherein there are fewer people like themselves, leading to fewer networking opportunities and a work culture that is less friendly. Societal stereotypes may also serve as a barrier. There also is evidence for implicit and explicit biases in the hiring and promotion processes.

There may also be occupational sorting along educational attainment lines. Some of this certainly will arise from the educational pipeline providing workers with the right tools to do a certain job. There are also licensure requirements which are often tied to different education bars. However, there may also be barriers that workers without college degrees face even if they have the correct skill sets and there are no licensure requirements. Skills-first hiring may help mitigate some of these barriers and increase equity across educational lines, but it is unlikely that current economies are fully skills-first.

The way in which different groups are sorted into occupations–a phenomenon called occupational segregation–has largely followed the seminal work of Duncan and Duncan, 1955. They built a measure still used today. However, while that index contains valuable information about the extent to which two different groups are sorted across occupations, it misses important information about how workers are sorted within and between firms. Consider a hypothetical economy where men and women were equally represented in every occupation at the national level, but never worked at the same firms. This would represent a different kind of occupational segregation, between firms, that could easily reinforce discrimination and impede the productive exchange of ideas; equity and efficiency both could be damaged, despite having no sorting by gender within any individual firm (since each firm employs only one gender, within-firm occupational sorting is precluded). Segregation within firm could likewise be damaging for an economy. The Duncan Index, estimated at the aggregate level without firm identifiers, cannot capture either of these dimensions.

In this paper, we build a new model of occupational segregation which has several advantages over Duncan and Duncan, 1955. First, by being estimated at the firm level, we are able to better identify underlying sources of inequity and inefficiency that would be hidden in aggregate. This further allows us to examine averages and variances across firms, between countries and industries, and over time. Second, our model allows for examination not only of the overall segregation index for any firm or aggregation of firms, but the within-firm and between-firm segregation indices as well, which helps illuminate the nature of the sorting with more nuance. Third, our model allows us to examine more than two groups, such that we can naturally examine not only binary cases like gender, but multinomial categories like educational attainment. In fact, we show that Duncan and Duncan, 1955 is a special case of our model.

We find the majority of firms are occupationally segregated: 88% for gender and 73% for education. Gender segregation is larger and more frequent than education-based segregation. While consistent across countries and industries, there is substantial variation in the level of occupational segregation. The United States and Canada exhibit the highest rates of gender-based occupational segregation and India the highest for education. We additionally find that firm-level occupational segregation is associated with lower median pay, greater pay dispersion, and fewer applicants per vacancy, consistent with theoretical predictions from crowding models.

1.1 Related literature

This paper is related to three areas of the literature: the methodology of occupational segregation, empirical evidence on segregation trends, and firm-level segregation and its labor market consequences.

Measurement of occupational segregation

The most widely used measure of occupational segregation is the dissimilarity index, adapted from Duncan and Duncan, 1955's residential segregation measure, which captures the share of one group (out of two) that would need to change occupations to achieve parity with the other group. Subsequent work has proposed alternative indices. Karmel and MacLachlan, 1988 define their index (IP) treating both groups symmetrically rather than deviations from one reference group. The IP index takes the overall occupational counts and group shares and measures the share that would have to change occupations to reach zero segregation. Silber, 1989 introduces a segregation index connected to the Gini coefficient, linking segregation to the income inequality literature. Cortese et al., 1976 provide a critique of the Duncan index's limitations when applied to multi-group settings. Boisso et al., 1994 extend occupational segregation measures to the case with more than two groups, including decompositions of segregation across multiple groups simultaneously and developing tests of statistical significance, a precursor to both our multi-group generalization and our hypothesis-testing framework. Watts, 1998 creates an axiomatic characterization of segregation measures, identifying the properties that distinguish different indices. Frankel and Volij, 2011 extend this axiomatic approach, characterizing multigroup segregation measures, including the Mutual Information index, and providing normative foundations for choosing among them.

Mora and Ruiz-Castillo, 2003 develop entropy-based measures of multigroup segregation with within- and between-group decompositions (see also Mora and Ruiz-Castillo, 2009). Our approach differs in that we begin at the firm level and decompose into within-firm and between-firm components, whereas the entropy-based literature decomposes across organizational units at the aggregate level.

Our simulation-based normalization approach is related to each of these, but with extensions into new areas including the focus of segregation at the firm level. Carrington and Troske, 1997 show that naive segregation indices are biased when organizational units are small. They propose a normalization that compares realized segregation to the distribution expected under random assignment. This is the same conceptual strategy we employ, though we implement it via machine learning prediction of simulation moments to achieve computational tractability at scale (each firm with their own index distribution based on their number of occupations and workers).

Empirical evidence on occupational segregation trends

A large empirical literature has documented trends in occupational segregation. For the United States, Blau et al., 2013 calculated the evolution of occupational segregation by gender from 1970 to 2009, finding declines that slowed substantially after 1990. Charles and Grusky, 2004 provide cross-national analysis, documenting both universal patterns (e.g., female underrepresentation in manual occupations) and high levels of country-level variation in segregation levels. Blackburn et al., 2000 analyze cross-national patterns of gender segregation and inequality, including compositional differences in occupational structures across countries (a concern that is relevant to our cross-country analysis). Cortés and Pan, 2023 examine how children and parenthood drive gender gaps in earnings and occupational outcomes, and Petrongolo and Ronchi, 2020 examine how gender gaps vary with local labor market structure across countries. For developing countries, Borrowman and Klasen, 2020 provide evidence on gendered occupational segregation patterns that contextualize our findings for Kenya and India.

Firm-level segregation and labor market outcomes

Our firm-level approach connects to a smaller literature on within-establishment segregation. Carrington and Troske, 1998 measure workplace segregation using matched employer-employee data and find that establishment-level segregation is substantial even after accounting for occupational and industrial composition. Hellerstein and Neumark, 2008 use this in the United States, documenting significant within-establishment racial and ethnic segregation. Tomaskovic-Devey et al., 2006 document trends in workplace segregation by race, ethnicity, and sex from 1966 to 2003, analyzing the organizational processes that produce and reproduce these patterns.

Our Section 4 analysis linking segregation to firm outcomes draws on foundational work in the economics of discrimination. Becker, 1957 models employer discrimination and its consequences for wages and employment. Bergmann, 1974 develops the crowding hypothesis, predicting that exclusion of a group from certain occupations depresses wages in the occupations to which they are crowded, the theoretical mechanism underlying our hypotheses about pay and pay dispersion. Blau and Kahn, 2017 provide a comprehensive decomposition of the gender wage gap that highlights the role of occupational segregation. Our finding that educational segregation is associated with fewer applicants per vacancy connects to work on degree inflation and credential barriers, including Fuller and Raman, 2017, who document that many employers impose degree requirements that screen out qualified workers.

2 Index methodology

Consider the goal of understanding the extent to which a firm has a composition of employees that reflect the workforce along given dimensions such as gender or education. Proportional representation of groups likely has important equity considerations. However, in doing this exercise, we want to consider not only representation within the firm overall, but how members of each group are allocated across occupations. Put another way, even if a firm had equal men and women, if the men held all the executive positions only, and the women held only administrative support positions, we would not consider this an equitable composition of employees.

The goal then is to use a measurement that measures the extent of occupational segregation within the firm. There are two primary ways that a firm can create occupational segregation: in terms of disproportionate overall compositions (e.g., if a firm hired 90% men), and in terms of allocation within the firm by occupation. The first creates between-firm occupational segregation: firms that disproportionately hire from one group concentrate workers of that group across occupations at the national level.

For illustrative purposes, consider an economy with only two occupations and six firms (Table 1). We designed this example so that there were exactly the same number of men and women in the workforce (12 of each). Thus, the target representation of women within firm and occupation is 0.5. Firm 1 has exactly equality, and an occupational segregation index score would return a value of zero. The other firms all have different versions of occupational segregation, which we explore below.

Table 1: Hypothetical example: number of employees per firm

	Occ. 1		Occ. 2	
Firm	Men	Women	Men	Women
1	1	1	1	1
2	2	0	2	0
3	0	2	0	2
4	2	0	0	2
5	3	0	0	1
6	0	3	1	0

2.1 Generalized Occupational Segregation Index

The metrics are built on $n_{g|jk}$, the count of employees in firm j and occupation k who belong to group g (e.g., men vs women). From these we construct proportions of workers in group g for different samples. For example, $p_{g|jk}$ is the proportion of workers in group g from the sample of workers at firm j in occupation k, and is calculated by

$$p_{g|jk} = \frac{n_{g|jk}}{\sum_{g \in G} n_{g|jk}} = \frac{n_{g|jk}}{n_{jk}} \tag{1}$$

Here we also use the notation $n_{jk} = \sum_{g \in G} n_{g|jk}$ which represents the number of workers at firm j which are in occupation k. Throughout, j and k are treated as non-stochastic: the total number of workers that firm j has in occupation k is taken as given, while the allocation of those workers across g groups is stochastic.

Likewise, the overall population proportion of workers in group g is given by P_g (capital P distinguishes population-level proportions from subgroup proportions such as $p_{g|jk}$)

$$P_g = \frac{\sum_{j \in J} \sum_{k \in K} n_{g|jk}}{\sum_{j \in J} \sum_{k \in K} n_{jk}} = \frac{n_g}{n} \tag{2}$$

P_g and n_g are the only segmentations by g that are non-stochastic–we assume that the number of workers in the economy belonging to each group is deterministic. While this is a limiting assumption, it is outside the scope of this paper to consider and model group-level differences in labor force participation decisions.

The occupational segregation index for a firm is an estimate of, on average, how close the proportion of workers in each group (e.g., men and women) within each occupation at the firm are to the population proportions in the group (the overall national proportions of men and women). The closer the numbers, the smaller the occupational segregation index. The overall occupation segregation index ($OS_j^{overall}$) for firm j is given by equation 3.

$$OS_j^{overall} = \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left| p_{g|jk} - P_g \right| \tag{3}$$

The inner element $\left| p_{g|jk} - P_g \right|$ represents, for a given firm (e.g., LinkedIn) and occupation k (e.g., software engineers), how closely the group proportions within that occupation (e.g., the share of male and female software engineers) align with the national workforce proportions of each group.

These differences are averaged across k occupations present in firm j. Each occupation receives more weight–or importance–if there are more total workers in that occupation within the firm. Thus, if a firm is composed 80% by occupation 1 and 20% in occupation 2, then the metric upweights the group disparities represented in occupation 1. This weight is given by

$$\theta_{jk} = \frac{n_{jk}}{\sum_{k \in K} n_{jk}} \tag{4}$$

Finally, ω is a scaling parameter given by equation 5, included to ensure the metric ranges from 0 (no segregation) to 1 (maximum segregation) by dividing by the theoretical maximum the index could take under maximal occupational segregation

$$\omega = \frac{1}{2(1 - \min_g P_g)} \tag{5}$$

2.2 Connection with the Duncan Index

The Duncan Index is the most common index used to measure occupational segregation. The Duncan Index is given by

$$DI = \frac{1}{2} \sum_{k} \left| \frac{\sum_{j \in J} n_{M|k}}{\sum_{k \in K} \sum_{j \in J} n_{M|k}} - \frac{\sum_{j \in J} n_{W|k}}{\sum_{k \in K} \sum_{j \in J} n_{W|k}} \right| \tag{6}$$

Intuitively, this captures the fraction of men in occupation k across all firms minus the fraction of women in occupation k across all firms, summed up across all occupations.

In the appendix, we show that the Duncan Index is a special case of our Occupational Segregation index under the following conditions:

1. There are only two groups (e.g., men and women)

2. There is only one firm (or in other words, it is calculated economy-wide)

3. The two groups are identical in magnitude (this is relevant for the scaling parameter ω)

Our more flexible model allows for estimates at the firm level and across more than one groups (i.e., education levels, such as high school graduates, associate degree holders, bachelor's degree holders, and graduate degree holders). Estimation at the firm level has many benefits, including the ability to detect whether there is occupational segregation within firms.

2.3 Components of Occupational Segregation Index

We can decompose the overall segregation index using the triangle inequality in two ways: first, by breaking it down by within and between firm segregation, and second by breaking it down by within and between occupation segregation.

First, we demonstrate the calculation for within and between firm occupational segregation.

$$
\begin{aligned}
OS_j^{overall} &= \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left| p_{g|jk} - P_g \right| \\
&= \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left| p_{g|jk} - p_{g|j} + p_{g|j} - P_g \right| \\
&\leq \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left(\left| p_{g|jk} - p_{g|j} \right| + \left| p_{g|j} - P_g \right| \right) \\
&= \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left| p_{g|jk} - p_{g|j} \right| + \omega \sum_{g \in G} \sum_{k \in K} \theta_{jk} \left| p_{g|j} - P_g \right| \\
&= \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left| p_{g|jk} - p_{g|j} \right| + \omega \sum_{g \in G} \left| p_{g|j} - P_g \right| \\
&= OS_j^{within-firm} + OS_j^{between-firm} \qquad (7)
\end{aligned}
$$

Consider $OS_j^{within-firm} = \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left| p_{g|jk} - p_{g|j} \right|$. Intuitively, instead of measuring the distance between the fraction in each group within an occupation and the target of the population proportion of the same group, we use the target of the fraction of workers from each group at the firm. This thus takes as given the overall proportion of workers within the firm, and only considers how those employees are allocated across occupations within the firm. Imagine a firm with only 10% of all employees which are women, but for every occupation, 10% of workers within the firm are women. This firm would not have a good overall occupational segregation score given they are far off from the approximately 50-50 split of gender, but they would have a perfect score for the within-firm segregation index given women have the same representation in every occupation equally.

On the other hand, $OS_j^{between-firm} = \omega \sum_{g \in G} \left| p_{g|j} - P_g \right|$ measures the overall representation of each group at the firm, irrespective of occupation. The same hypothetical firm discussed in the prior paragraph which had a very good score for within-firm occupational segregation would have a poor score for between-firm occupational segregation given they hire so few women overall (Table 2).

Note firm 1 represents a firm with no occupational segregation, which is reflected for overall, within, and between. Firms 2 and 3 are fully occupationally segregated, driven only by between-firm segregation. This is shown by them only hiring one gender. Thus it is equal for each occupation (e.g., for firm 2, at 0% women, 100% men) and have zero within-firm occupational segregation. But this is only due to the fact that they hire no women.

Table 2: Hypothetical example: occupational segregation scores

	Occ. 1		Occ. 2				
Firm	Men	Women	Men	Women	OS	OS within	OS between
1	1	1	1	1	0	0	0
2	2	0	2	0	1	0	1
3	0	2	0	2	1	0	1
4	2	0	0	2	1	1	0
5	3	0	0	1	1	0.75	0.5
6	0	3	1	0	1	0.75	0.5

Firm 4 on the other hand also has a score for occupational segregation of one, but this arises from them having the highest value of the within-firm occupational segregation. This is shown by the fact that they have equal numbers of men and women overall at the firm (and thus no between-firm segregation), but men are all in occupation 1 while women are all in occupation 2. Firms 5 and 6 have combinations of different occupational segregation. They have the same scores as each other because both men and women can be segregated.

We can also do a decomposition based on within-occupation and between-occupation segregation:

$$\begin{aligned} OS_j^{overall} &= \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left| p_{g|jk} - P_g \right| \\ &= \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left| p_{g|jk} - P_{g|k} + P_{g|k} - P_g \right| \\ &\leq \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left(\left| p_{g|jk} - P_{g|k} \right| + \left| P_{g|k} - P_g \right| \right) \\ &= \omega \sum_{k \in K} \theta_{jk} \sum_{g \in G} \left| p_{g|jk} - P_{g|k} \right| + \omega \sum_{g \in G} \sum_{k \in K} \theta_{jk} \left| P_{g|k} - P_g \right| \\ &= OS_j^{within\ occupation} + OS_j^{between\ occupation} \end{aligned} \tag{8}$$

Within-occupation segregation asks the extent to which a firm's hiring in each occupation deviates from that occupation's national-level group composition. In other words, if women comprise 20% of software engineers nationally, then a firm's hiring in software engineering would be non-segregated (in the within-occupation sense) if it matched that same

20% rate. Between-occupation segregation captures the extent to which the national-level participation rate of each group in a given occupation differs from that group's overall share of the workforce. The inner absolute value term is identical for all firms and reflects the mismatch between economy-wide occupation participation rates and population shares. In this respect it is the component most closely aligned with the Duncan Index. However, for any given firm these population-level mismatches are weighted by the firm's occupational employment shares. Firms concentrated in nationally imbalanced occupations will therefore have a higher between-occupation segregation score, independent of their own hiring decisions. This decomposition thus separates a firm's overall occupational segregation into the portion the firm controls directly (within-occupation, which treats hiring as non-segregated as long as the firm matches the population rates within each occupation) and the portion driven by economy-wide occupational composition that the firm does not control.

2.4 Hypothesis testing and normalization

If the goal is to make comparisons between firms, industries, geographies, years, or other factors, then using the occupational segregation index without any adjustment will not offer a fair comparison. Intuitively, a firm with fewer employees and many occupations will have a harder time having an equitable distribution of employees than a firm with more employees and few occupations, even if both are hiring within each occupation in complete alignment with population-level group proportions. Both of these are due to a small number of employees-per-occupation–if a firm hires only two people in a given occupation and does so randomly with respect to a population with an equal number of men and women, there is only a 25% probability that they will hire one man and one woman. Thus, what may look like bias against a group may just be due to random chance.

In order to account for this randomness and make fair comparisons, we simulate the distribution of the occupational segregation index for a firm with the same distribution of occupations and target populations, but which was hiring completely at random with respect to the target population proportions. We call this representative hiring, noting that due to small firm sizes and large numbers of occupations, a firm hiring representatively could still end up with high occupational segregation indices in some draws.

We use these simulated at-random hiring distributions of the occupational segregation indices with the firm's realized values in two ways. First, we divide the realized occupational segregation index of the firm by the mean of the at-random simulated distribution of the index. A value of 1.2 for example would imply that the first has an occupational segregation index 1.2 times (or 20%) higher than would happen if hiring was random. Second, we calculate the 95th percentile of the at-random simulated metric distribution. This allows us to do a one-tailed hypothesis test of whether the realized occupational segregation indices are statistically different from the null hypothesis of hiring at random. Given the millions of firms across countries and years, simulating the representative hiring distribution moments for each firm individually is computationally intractable. Instead, we train a two-layered machine learning model for a representative sample of firms for which we simulate the moments. We use this trained model to predict the two target moments (mean under at-random hiring and the 95th percentile) for the whole sample of firms as functions of the firm's count of workers in each occupation and the target population ratios. This procedure has excellent out of sample prediction (given the distribution is a direct function of only the observed inputs), and allows us to scale the sample significantly. Appendix B provides the details on the sampling, training, prediction, and validation of these models.

3 Results for Index Calculations

In order to demonstrate the application of this methodology, we now show estimates across twelve countries and ten years. Future applications will extend coverage further and examine relationships with additional outcomes.

3.1 Summary Statistics for Occupational Segregation Indices

Table 3 shows the mean statistics for gender. We walk through this table to explain interpretation. The first column shows the mean of the unadjusted metric. For example, the overall occupational segregation index metric for gender across these ten years and 12 countries is 0.517. Alone, this does not yield meaningful insights. However, the second column demonstrates that, if each of the over fifty thousand firms evaluated had been hiring in a representative way (i.e., with probabilities of hiring equal to the target population shares), then the occupational segregation index would have a mean of 0.358. Thus, the unadjusted mean is higher than what we would expect under representative hiring. We can use the metric for each firm under representative hiring to see how much it differs by. This normed mean is the ratio of the unadjusted metric and the representative hiring 'target' metric. If we average this ratio, we find that on average firms have an overall segregation index that is 1.754 times as large as the metric we would expect under representative hiring. Additionally, as discussed above we calculate the 95th percentile of the distribution for each firm's occupational segregation index under representative hiring. We use this to do a single-tailed hypothesis test where the null distribution is that the firm is employing workers in a way consistent with representative hiring. 85.3% of firms overall reject this null in favor of occupationally-segregated hiring with respect to gender: their occupational segregation index exceeds the 95th percentile of the distribution of the metric if they had been hiring representatively. Thus, there is evidence for a substantial amount of occupational segregation at the firm level. For comparison, the Duncan index for these countries and years averages to 0.306 (standard deviation of 0.03), smaller than the unadjusted mean in the new index.

Table 3: Summary Statistics of Segregation Indices for Gender

	Unadjusted mean	Representative target mean	Normed mean	Share rejecting representative hiring
Between firm	0.259	0.088	6.019	0.752
	(0.170)	(0.050)	(5.430)	(0.432)
Between occupation	0.286	0.008	76.138	1.000
	(0.089)	(0.010)	(47.552)	(0.000)
Overall	0.517	0.358	1.754	0.853
	(0.146)	(0.128)	(0.873)	(0.354)
Within firm	0.455	0.337	1.556	0.863
	(0.149)	(0.126)	(0.463)	(0.344)
Within occupation	0.439	0.329	1.556	0.867
	(0.137)	(0.117)	(0.591)	(0.340)

Note: Means, with standard deviations in parentheses.

The next two rows of Table 3 show the decomposition with respect to within and between firm occupational segregation. If we just examine the unadjusted means, within-firm occupational segregation appears to account for substantially more of the overall segregation than the between-firm occupational segregation. That is, it is not that the results is primarily driven by all-male firms with men working across all occupations and all-female firms with women working across all occupations (which would be captured under between-firm segregation). Instead, it is more driven by within-firm, women and men being sorted to different occupations, holding constant the number of men and women the firm hired.

This is also reflected in a higher share of firms having employment that does not resemble representative hiring (86.3% vs. 75.2% for between-firm hiring). The between-firm representative target mean for representative hiring is substantially smaller than that of within-firm, 0.088 versus 0.337. This is because, while it may be difficult to hire proportionally for three positions in occupation a and five positions in occupation b for example, it is easier to hire proportionally across all 200 workers among all 20 occupations at the firm. As a result, firms on average are 6.019 times as high as the representative index mean for between-firm occupational segregation.

The final two rows decompose the effect to within-occupation and between-occupation segregation. Again, within-firm segregation is higher when unadjusted. This metric takes as given the economy-wide shares of men and women in each occupation, and compares the firm's level to those targets. Between occupation reflects the extent to which they have workers in highly segregated occupations at the economy level. We again find high rates of rejecting the null hypothesis, at 86.7% for within-occupation segregation and 100% for between-occupation, since representative hiring treats the occupational distribution of jobs at the firm as fixed, and the only source of random deviation arises from economy-wide occupational hiring shares, which under representative hiring will be very close to their targets given the very large number of workers at the economy level. This is not necessarily as useful of a hypothesis test given this setup.

Table 4 presents the means for education occupational segregation. The same overall patterns hold, where a majority of firms occupationally segregate, and within firm or occupation segregation is larger than between when unadjusted, but smaller when normed. Interestingly, the mean and normed means, as well as shares rejecting representative hiring, are all smaller than observed for gender. For example, for overall segregation, the mean is 1.455 times larger than the representative index for education compared to 1.754 times larger for gender. For education, 70.2% of firms have an index which is higher than the 95th percentile of the representative index distribution compared to 85.3% of firms for gender. This is particularly surprising for two reasons: discrimination based on gender is prohibited by law, whereas educational requirements in hiring are generally not. Indeed, degree requirements are common and often legally permitted. Second, many occupations are licensed, which requires certain educational attainment.

One of the reasons why education may have lower occupational segregation than gender is because there are five categories in the data compared to two for gender. More categories may increase the target means and thus scale less. Appendix Table 6 recalculates the index while making education a binary variable: 4-year degree or higher versus less than four year degrees (AA or no college). Interestingly, while the representative targets are lower, so too are the unadjusted means. The normed means are similar to those for the five-category education measure, and remain below the normed means and share rejecting representative hiring observed for gender. Thus, it is not driven just by the number of segments in the index.

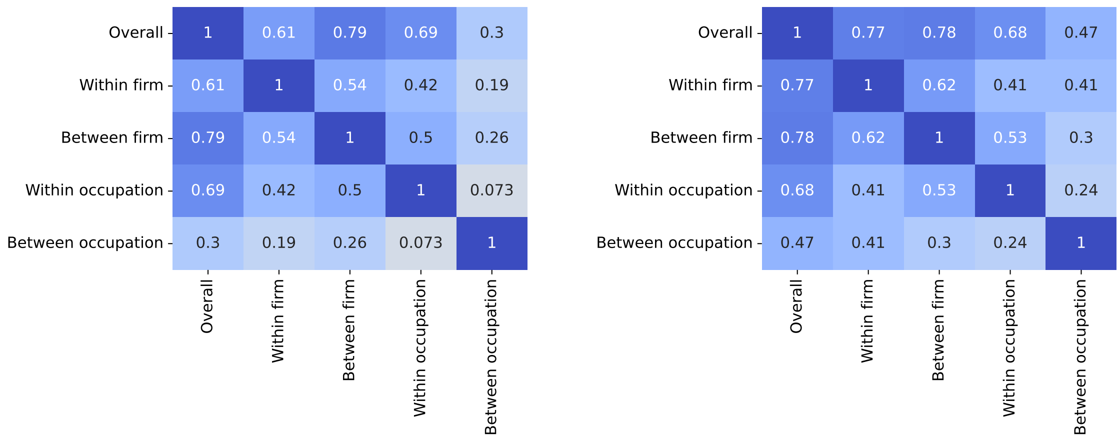
Table 4: Summary Statistics of Segregation Indices for Education

	Unadjusted mean	Representative target mean	Normed mean	Share rejecting representative hiring
Between firm	0.182	0.079	3.899	0.801
	(0.097)	(0.038)	(3.291)	(0.399)
Between occupation	0.191	0.008	48.239	1.000
	(0.070)	(0.010)	(31.687)	(0.002)
Overall	0.396	0.318	1.455	0.702
	(0.101)	(0.107)	(0.636)	(0.458)
Within firm	0.320	0.286	1.255	0.503
	(0.111)	(0.104)	(0.438)	(0.500)
Within occupation	0.310	0.280	1.251	0.517
	(0.091)	(0.094)	(0.424)	(0.500)

Note: Means, with standard deviations in parentheses.

We can also examine correlations across the five occupational segregation metrics (each normalized by the mean index value for a similar firm under representative hiring). There is positive correlation between each of the metrics, with the highest correlation coefficients being between the overall metric and the between-firm segregation metric, which measures the overall representation of each group at the firm irrespective of occupations, compared to the workforce representation rates. However, within-occupation and within-firm are likewise very strongly correlated with the overall occupational segregation metric, at between a value of 0.6 and 0.7. Between-occupation segregation–that driven by whether the firm concentrates in occupations that are nationally segregated–has the lowest correlation with other metrics. Overall, there is some level of information contained in each of these metrics not fully explained by another.

Figure 1: Correlation Coefficients in Occupational Segregation Normed Indices



(a) Gender

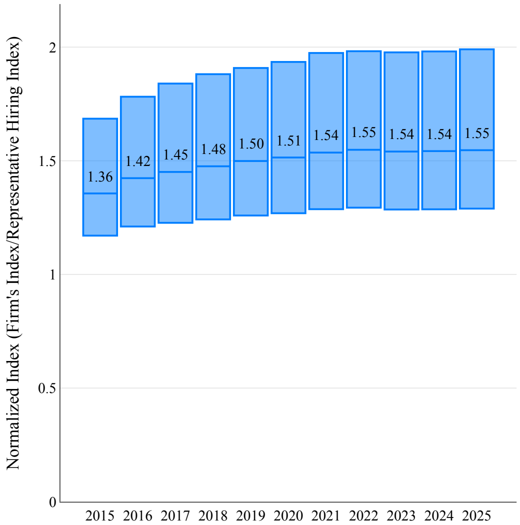
(b) Education

3.2 Occupational Segregation Indices Across Time

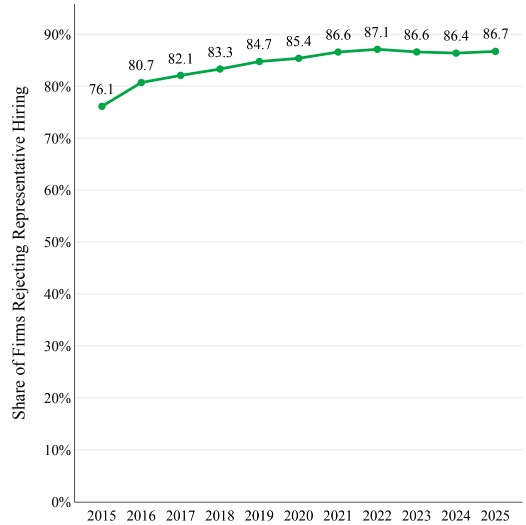
We examine trends over time for the overall occupational segregation index first by gender, shown in Figure 2. We present the two primary metrics for the overall occupational segregation index: the normalized index and the share of firms whose index rejects the null hypothesis of representative hiring. We see in both cases that occupational segregation has been increasing over the last decade. For the normalized index, the overall occupational segregation index went from 36% larger than the firm's index under representative hiring in 2015 (1.36) to 55% higher in 2025. However, most of this expansion in occupational segregation occurred between 2015 and 2021, with subsequent stability. Interestingly, this mirrors the trend in hiring of women into leadership, which increased over the 2015-2022 period and then plateaued. The interquartile range also grows slowly over this time period. Additionally, 76.1% of firms were statistically distinguishable from representative hiring in 2015; that increased to 86.7% in 2025.

Next, we examine the same trends for occupational segregation by education, as shown in Figure 3. As found in the summary statistics, we see somewhat lower levels of occupational segregation by education than by gender for each year. We also see slower increase in occupational segregation over time by education than we did for gender–20% higher in 2015 than the representative index mean, and 27% higher in 2025. The share of firms for which we reject representative hiring has also increased from 65% in 2015 to 72% in 2025–still

Figure 2: Occupational Segregation Indices for Gender Over Time



(a) Normalized Index (Firm's Index / Representative Hiring Index)



(b) Share of Firms Rejecting Representative Hiring

Note: Subplot (a) shows the interquartile range and median. Subplot (b) shows the percent of firms whose occupational segregation index is statistically differentiable from the distribution of the index under representative hiring.

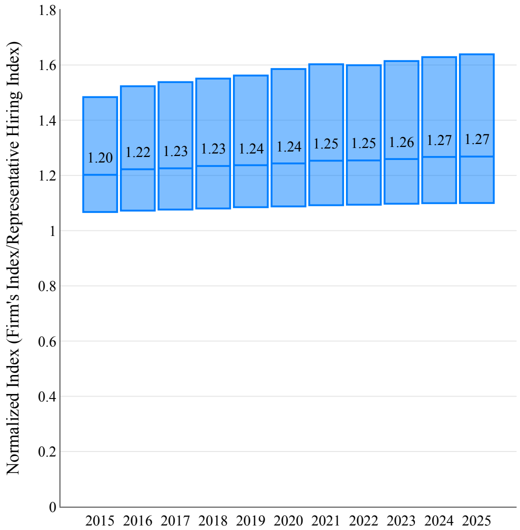
lower than the rates for gender.

In the Appendix, we report the over-time breakdowns for the decomposition of the metrics. For gender (Figure 13), all show increases over time, with between-occupation segregation increasing the most over time. For education-based occupational segregation (Figure 14), we actually find more stable over-time measures for within-firm sorting, and to a lesser extent within-occupation. However, overall we still see high levels of education-based occupational segregation over time.

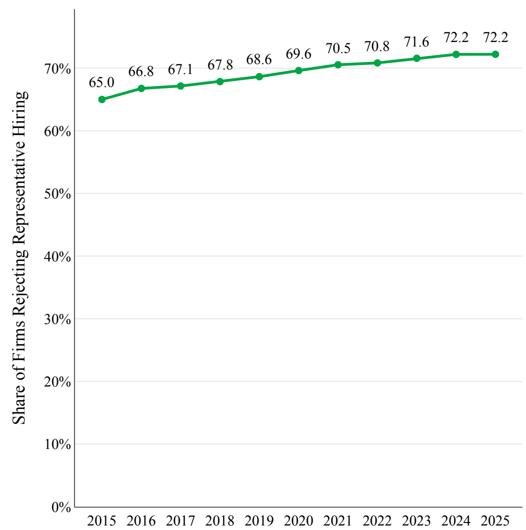
3.3 Occupational Segregation Index Across Countries

We next turn to occupational segregation across the twelve countries we evaluate: Australia, Brazil, Canada, France, Germany, India, Kenya, Mexico, Nigeria, Singapore, the

Figure 3: Occupational Segregation Indices for Education Over Time



(a) Normalized Index (Firm's Index / Representative Hiring Index)



(b) Share of Firms Rejecting Representative Hiring

Note: Subplot (a) shows the interquartile range and median. Subplot (b) shows the percent of firms whose occupational segregation index is statistically differentiable from the distribution of the index under representative hiring.

United Kingdom, and the United States.[1] This analysis in particular highlights the reliance our estimates have on our sample, and how representative it is. Thus, for gender, as shown in Figure 4, we find the United States and Canada have the highest rates of overall occupational segregation (around 90% of firms, and indices around 60% higher (1.6) than the representative hiring mean), while the lowest rates are for Kenya (45.4% of firms) and India (53.7% of firms). On the other hand, occupational segregation by education is the highest in India of the twelve countries, followed by the United States. Australia and Singapore have the lowest levels.

The finding that Kenya and India exhibit the *lowest* gender-based occupational segregation is counterintuitive given prior research of these countries' gender gaps in labor force participation. This may reflect LinkedIn sample composition: LinkedIn penetration is higher in the United States, Canada, and the United Kingdom than in Kenya and India, where users

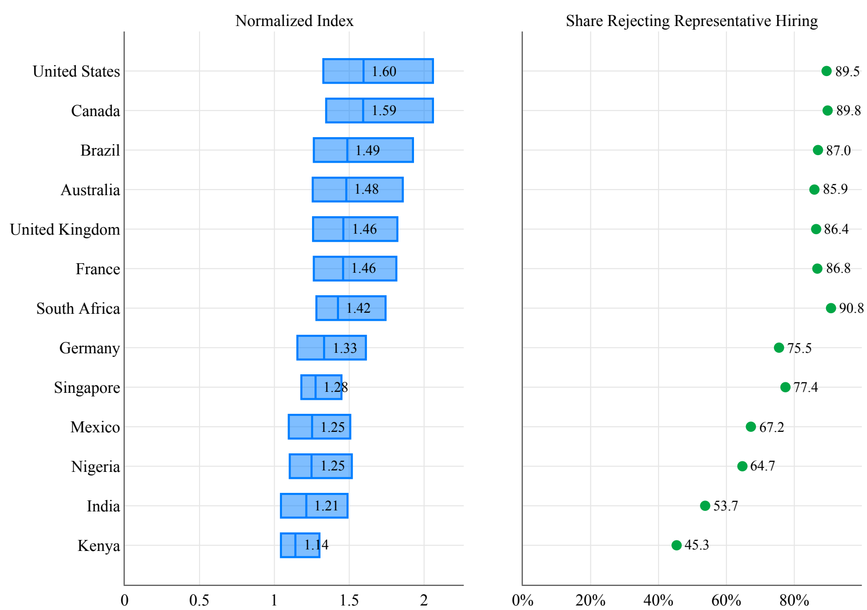
[1] Countries were selected based on geographic spread and sufficient LinkedIn sample sizes to produce stable firm-level estimates across the ten-year window.

tend to be drawn from a more professional, urban, and broader formally educated segment of the workforce. The firms observed on LinkedIn in these countries may therefore represent a segment of the national labor market more likely to have low occupational segregation, compressing our metrics relative to the economy as a whole. Cross-country comparisons should be interpreted with this sample selection in mind.

The reversal between gender and education dimensions is also noteworthy. India exhibits the lowest gender-based segregation but the *highest* education-based segregation among the twelve countries. This may reflect the strong role of educational credentials and professional licensure in India's formal sector, which creates occupational barriers along educational lines even as gender composition within those occupations is more balanced among LinkedIn-observed firms and workers.

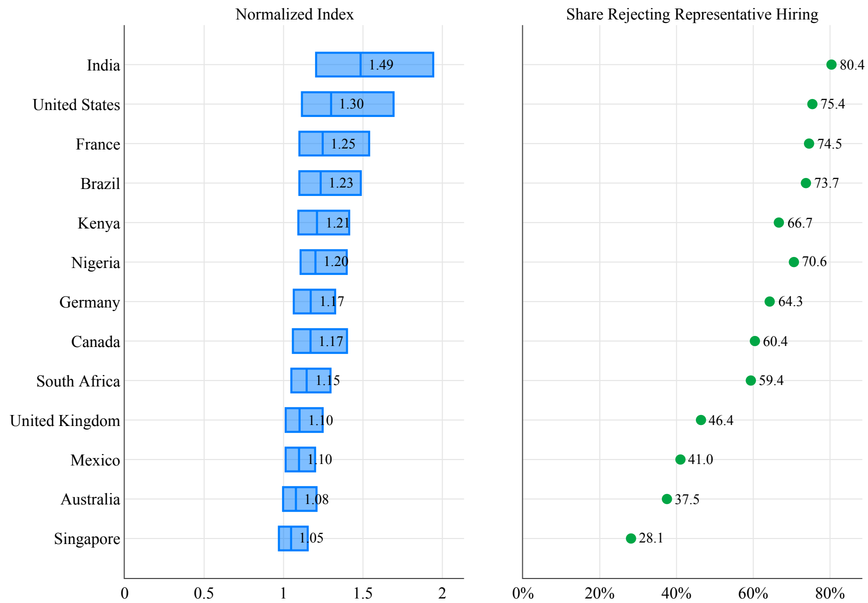
More broadly, cross-country differences likely reflect some combination of labor market institutions (anti-discrimination legislation, parental leave policies), occupational licensing policies, educational pipeline differences, and variation in LinkedIn sample composition. Disentangling these factors is beyond the scope of this paper.

Figure 4: Occupational Segregation Indices for Gender, by Country



Note: Chart presents box plots for firm-level estimates with the given country and year.

Figure 5: Occupational Segregation Indices for Education, by Country



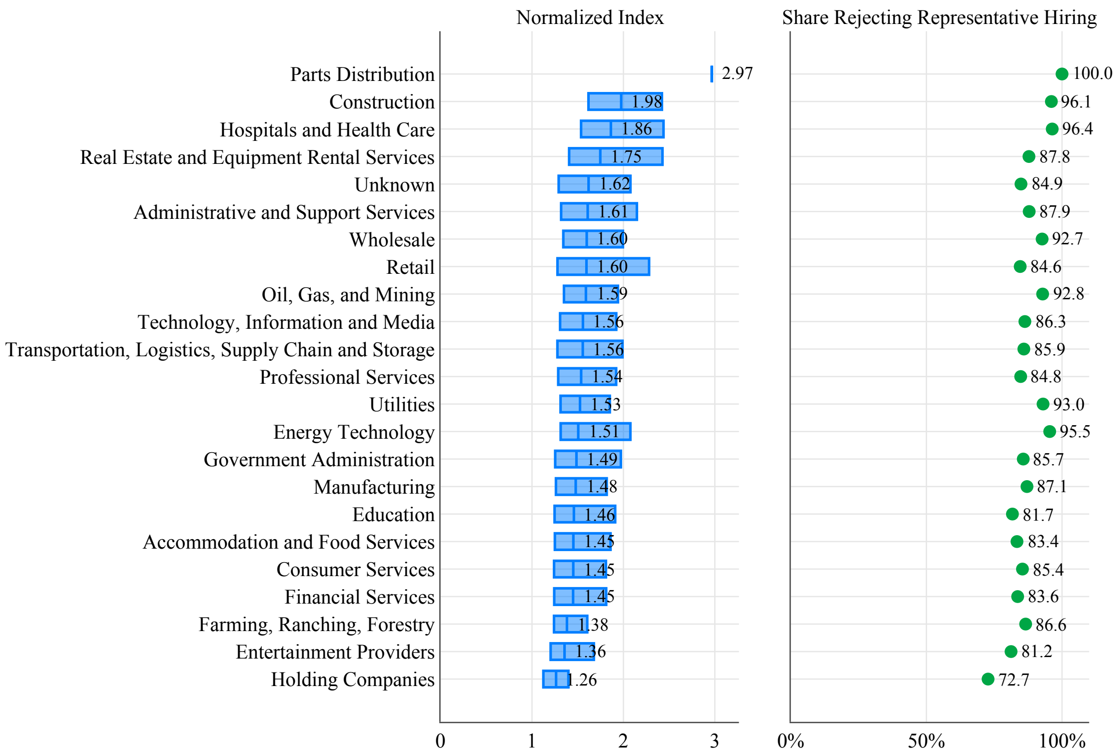
Note: Chart presents box plots for firm-level estimates with the given country and year.

3.4 Occupational Segregation Indices Across Industries

There is also meaningful variation across industries in the extent to which there is occupational segregation. For gender, the construction and hospitals-and-healthcare industries have the highest rates of occupational segregation. These are fields with differences in gender representation overall (low in construction, high in healthcare), but with highly stratified traditional gender roles reinforced by licensure barriers for many of the higher-paying occupations within the industry. Meanwhile, holding companies, entertainment providers, and farming have lower rates, although still quite high by gender.

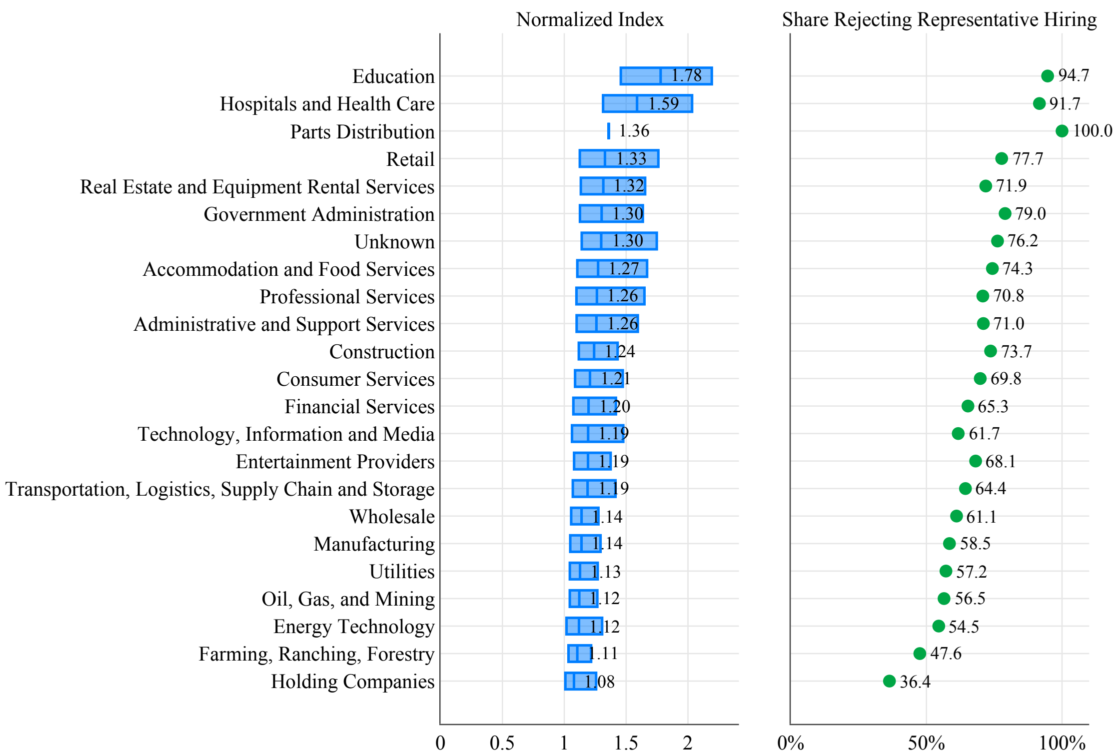
The across-industry variation in typical occupational segregation is much higher by industry than it is for gender. Here, education and hospitals/healthcare have the highest rates, with over 90% of firms rejecting representative hiring and indices that are 78% and 59% higher than the mean unbiased value. These are industries again with strict licensing connected directly to educational attainment. Meanwhile, farming, energy, oil and gas, and utilities all have much lower rates of occupational segregation by education. These are fields with fewer occupational barriers in terms of licenses, allowing for across-educational-line transitions.

Figure 6: Occupational Segregation Indices for Gender, by Industry



Note: Chart presents box plots for firm-level estimates with the given country and year.

Figure 7: Occupational Segregation Indices for Education, by Industry



Note: Chart presents box plots for firm-level estimates with the given country and year.

4 Occupational Segregation and Firm Outcomes

Beyond documenting the prevalence and distribution of occupational segregation, we examine whether segregation is associated with downstream firm-level outcomes. We consider four outcomes motivated by a simple theoretical model of two-occupations and two-groups (see Appendix D for the full derivation): median pay, pay dispersion, hiring intensity, and applicant volume per vacancy. All estimates are correlational; their interpretation is guided by the theoretical mechanisms outlined below.

4.1 Hypotheses and Mechanisms

We model occupational segregation as restricting access to higher paying occupations for at least one group (e.g., women). Under no segregation, both groups supply labor to all occupations. Under full segregation, the exclusion of some groups reduces labor supply to higher paying occupations and increases supply to lower paying occupations. Given downward-sloping labor demand, wages in the gated, higher paying occupations rise further, while wages in the other lower occupations fall. This generates four testable predictions. The model in the appendix uses a conceptual framework of a firm with a higher-paying occupation H and a lower-paying occupation L, and two groups of workers.

H1 (median pay). Segregation has two effects on average wages: a *composition effect* (workers shift from higher- to the lower-paying occupations, directly depressing average pay) and a *price-shift effect* (wages at the top rise, wages at the bottom fall, with ambiguous net impact). Because the composition effect is unambiguously negative and is the dominant mechanism, we predict lower median pay at segregated firms.

H2 (pay dispersion). Excluding one group from higher-paying occupations raises wages in those occupations (due to reduced labor supply) and lowers wages in lower-paying occupations (due to increased labor supply), increasing the 90th-to-10th percentile pay ratio. This prediction is unambiguous under downward-sloping labor demand. An alternative mechanism operates in models with worker heterogeneity (Roy, 1951; Heckman and Honoré, 1990): if segregation prevents high-ability workers from sorting into high-wage occupations, wages may fall at *both* ends of the distribution: the top, because high-ability workers are crowded out of high-wage roles, and the bottom, because they are crowded into low-wage ones. Whether dispersion rises or falls under this mechanism depends on the relative magnitudes. We treat these as competing predictions and examine both in the results.

H3 (applications per post). Excluding one group from higher-paying occupations directly thins the applicant pool for the firm's most competitive roles. While the applicant pool for lower-paying occupations expands, vacancies also rise there and match quality may deteriorate as workers from the constrained segment (e.g., women) who are suited for higher-paying occupations are instead crowded into lower-paying ones. The net effect on applications per vacancy is negative because the higher-paying occupation is typically a minority occupation (most workers are in lower-paying roles), so halving its applicant pool represents a large absolute loss in applications. The offsetting expansion of the lower-paying applicant pool is proportionally smaller and is partly absorbed by the simultaneous rise in vacancies there.

H4 (new-hires-to-employees ratio). Higher wages in the higher paying occupations reduce labor demand and hiring there; lower wages in lower-paying occupations increase labor demand but the influx of workers may raise or lower net hiring depending on demand conditions. The aggregate effect on firm hiring intensity depends on the relative demand elasticities of the two occupations and is *theoretically ambiguous*.

4.2 Empirical Strategy

We estimate a simple correlational model through regression analysis. For each hypothesis, we use the log outcome for firm i in period t. This is regressed on an indicator equal to one if the firm's overall occupational segregation index exceeds the 95th percentile of its simulated representative-hiring distribution (the hypothesis test in Section 2). The regression additionally controls for country, year, industry, and primary-occupation fixed effects. The coefficient on occupational segregation captures the average log difference in outcomes between occupationally segregated and non-segregated firms within the same country-year-industry-occupation cell.

These estimates are correlational, not causal. The fixed effects control for industry- and country-level confounds, but within-cell endogeneity remains a limitation. Results are reported separately for the two segments (occupational segregation by education and by gender). The appendix additionally reports the findings for the binary education approach (bachelors degree or higher versus lower).

4.3 Regression Results

Table 5 presents the estimates. H1, H2, and H3 are strongly supported across all three segments; H4 is mixed, consistent with theoretical ambiguity.

Table 5: Regressions of Outcomes and Occupational Segregation (Education and Gender)

Outcome	Education	Gender	N
log median pay	-0.0110* (-0.0207, -0.0013)	-0.0180* (-0.0325, -0.0035)	65168
log 90th percentile pay	0.0445*** (0.0339, 0.0551)	0.0338*** (0.0180, 0.0496)	65168
log 10th percentile pay	-0.0589*** (-0.0704, -0.0473)	-0.0693*** (-0.0865, -0.0521)	65168
log ratio of 90th and 10th pay	0.1034*** (0.0916, 0.1152)	0.1031*** (0.0855, 0.1207)	65168
log average applicants	-0.1028*** (-0.1118, -0.0939)	-0.0341*** (-0.0461, -0.0221)	277662
log hires to employees ratio	0.0161*** (0.0126, 0.0195)	-0.0038 (-0.0082, 0.0005)	403713

Note: Each cell reports estimates from a separate regression. Confidence intervals in parentheses.

H1 (median pay). Segregated firms have significantly lower median pay across all segments. The effect is largest in the gender specification (-0.018 log points).

H2 (pay dispersion). Segregated firms exhibit approximately 10 log points higher 90–10 pay ratios, statistically significant across both segments. This pattern is driven by increases at the 90th percentile and larger (in log terms) decreases at the 10th percentile, consistent with the model's prediction that segregation raises wages at the top while compressing them at the bottom.

The appendix specification that recodes education as a binary measure offers a direct test of the two mechanisms. There we observe declines at both the 10th *and* 90th percentiles, consistent with the Roy model prediction that high-ability workers are crowded out of high-wage occupations, depressing wages at the top even as they compress wages at the bottom. Despite the decrease at the 90th percentile, the 90–10 pay ratio still increases by 7.8 log points, indicating that the overall dispersion effect remains positive even under this alternative mechanism.

H3 (applicants per post). Segregated firms attract significantly fewer applicants per vacancy across all segments, with the largest effect in the education specification (-0.103 log points). This is consistent with a substantially thinned applicant pool in the higher-paying occupation, not fully offset by expanded applicant flows to lower-paying roles.

H4 (new hires/employees). For gender, the coefficient is small and not statistically significant (-0.004), consistent with theoretical ambiguity. For educational segregation, the coefficient is positive and significant. A plausible interpretation beyond the static model is that educational segregation correlates with elevated turnover in lower-paying occupations (workers placed in roles below their credential level may be less engaged and more likely to separate), generating more replacement hiring.

5 Discussion

In this paper, we develop new firm-level measures of occupational segregation which can be applied across more than binary classifications such as gender. We also demonstrate how to decompose these into within vs between firm as well as within vs between occupation segregation. We normalize these metrics and conduct hypothesis tests. We apply this to LinkedIn firm data across twelve countries and ten years.

Several findings are of interest in data we apply these metrics to, some surprising. First, we find firms have, on average, an overall occupational segregation index about 1.8 times what would be expected under representative hiring for gender, and 1.5 times for education. Even after accounting for firm size and occupational diversity, which can make proportional hiring difficult, the majority of firms exhibit occupational segregation (88% for gender and 73% for education). Across all metrics, gender occupational segregation is consistently higher than education-based occupational segregation, despite the legal and licensure barriers that might be expected to produce the opposite result.

Additionally, although there are some industries which are lower or higher in their measures of occupational segregation, with Construction as well as Hospitals and Health Care being particularly segregated, the within-industry spread of firms' segregation is much higher than between-industries as shown by the large box plot spreads for each industry compared to the differences in the medians. Thus, no industry is immune from having a large segment of firms with either gender or educational occupational segregation.

The application of this data has limitations. Insofar as certain firms or types of workers are not represented on the LinkedIn platform, this may affect the measured values of the metrics. This is at least partly mitigated by the fact that the target proportions from which we measure deviations are also drawn from LinkedIn. Section 4 presents correlational evidence linking occupational segregation to median pay, pay dispersion, applicant volume, and hiring intensity, finding consistent associations across gender and educational dimensions. Future analysis may examine additional countries and explore causal mechanisms.

References

Baird, M., Lara, S., Hood, R., Gahlawat, N., & Ko, P. (2023). International gender representation in STEM employment and skills. *LinkedIn Economic Graph White Paper*.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago Press.

Bergmann, B. R. (1974). Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal*, *1*(2), 103–110.

Blackburn, R. M., Jarman, J., & Brooks, B. (2000). The puzzle of gender segregation and inequality: A cross-national analysis. *European Sociological Review*, *16*(2), 119–135. https://doi.org/10.1093/esr/16.2.119

Blau, F. D., Brummund, P., & Liu, A. Y.-H. (2013). Trends in occupational segregation by gender 1970–2009: Adjusting for the impact of changes in the occupational coding system. *Demography*, *50*(2), 471–492. https://doi.org/10.1007/s13524-012-0151-7

Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, *55*(3), 789–865. https://doi.org/10.1257/jel.20160995

Boisso, D., Hayes, K., Hirschberg, J., & Silber, J. (1994). Occupational segregation in the multidimensional case: Decomposition and tests of significance. *Journal of Econometrics*, *61*(1), 161–171. https://doi.org/10.1016/0304-4076(94)90082-5

Borrowman, M., & Klasen, S. (2020). Drivers of gendered sectoral and occupational segregation in developing countries. *Feminist Economics*, *26*(2), 62–94. https://doi.org/10.1080/13545701.2019.1649708

Carrington, W. J., & Troske, K. R. (1997). On measuring segregation in samples with small units. *Journal of Business & Economic Statistics*, *15*(4), 402–409. https://doi.org/10.1080/07350015.1997.10524718

Carrington, W. J., & Troske, K. R. (1998). Sex segregation in U.S. manufacturing. *Industrial and Labor Relations Review*, *51*(3), 445–464. https://doi.org/10.1177/001979399805100305

Charles, M., & Grusky, D. B. (2004). *Occupational ghettos: The worldwide segregation of women and men*. Stanford University Press.

Clarke, H. M. (2020). Gender stereotypes and gender-typed work. *Handbook of labor, human resources and population economics.*

Cortés, P., & Pan, J. (2023). Children and the remaining gender gaps in the labor market. *Journal of Economic Literature*, *61*(4), 1359–1409. https://doi.org/10.1257/jel.20221549

Cortese, C. F., Falk, R. F., & Cohen, J. K. (1976). Further considerations on the methodological analysis of segregation indices. *American Sociological Review*, *41*(4), 630–637.

Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, *20*(2), 210–217.

Frankel, D. M., & Volij, O. (2011). Measuring school segregation. *Journal of Economic Theory*, *146*(1), 1–38. https://doi.org/10.1016/j.jet.2010.10.008

Fuller, J. B., & Raman, M. (2017). *Dismissed by degrees: How degree inflation is undermining U.S. competitiveness and hurting America's middle class* (Report). Accenture, Grads of Life, Harvard Business School.

Heckman, J. J., & Honoré, B. E. (1990). The empirical content of the roy model. *Econometrica*, *58*(5), 1121–1149. https://doi.org/10.2307/2938338

Hellerstein, J. K., & Neumark, D. (2008). Workplace segregation in the United States: Race, ethnicity, and skill. *Review of Economics and Statistics*, *90*(3), 459–477. https://doi.org/10.1162/rest.90.3.459

Karmel, T., & MacLachlan, M. (1988). Occupational sex segregation—increasing or decreasing? *Economic Record*, *64*(3), 187–195. https://doi.org/10.1111/j.1475-4932.1988.tb02057.x

Lara, S., Baird, M., & Hood, R. (2023). Progress and barriers in global gender leadership. *LinkedIn Economic Graph White Paper*.

Mora, R., & Ruiz-Castillo, J. (2003). Additively decomposable segregation indexes: The case of gender segregation by occupations and human capital levels in Spain. *Journal of Economic Inequality*, *1*(2), 147–179. https://doi.org/10.1023/A:1026198429377

Mora, R., & Ruiz-Castillo, J. (2009). The invariance properties of the mutual information index of multigroup segregation. In *Research on economic inequality* (pp. 33–53). Emerald Group Publishing. https://doi.org/10.1108/S1049-2585(2009)0000017005

Petrongolo, B., & Ronchi, M. (2020). Gender gaps and the structure of local labor markets. *Labour Economics*, *64*, 101819. https://doi.org/10.1016/j.labeco.2020.101819

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*(2), 135–146. https://doi.org/10.1093/oxfordjournals.oep.a041049

Silber, J. G. (1989). On the measurement of employment segregation. *Economics Letters*, *30*(3), 237–243.

Tomaskovic-Devey, D., Zimmer, C., Stainback, K., Robinson, C., Taylor, T., & McTague, T. (2006). Documenting desegregation: Segregation in American workplaces by race, ethnicity, and sex, 1966–2003. *American Sociological Review*, *71*(4), 565–588. https://doi.org/10.1177/000312240607100403

Watts, M. (1998). Occupational gender segregation: Index measurement and econometric modeling. *Demography*, *35*(4), 489–496. https://doi.org/10.2307/3004016

APPENDIX

Appendix A: Duncan Index as Special Case of Occupational Segregation Index

Here we demonstrate that the Duncan Index is a special case of our occupational segregation index. We adjust the notation slightly from above to simplify the derivation.

Notation:

- n_{kM}: number of workers in occupation k who are men
- n_{kW}: number of workers in occupation k who are women
- $N_M = \sum_k n_{kM}$: total number of men across all occupations
- $N_W = \sum_k n_{kW}$: total number of women across all occupations
- $P_M = \frac{N_M}{N_M+N_W}$
- $P_W = \frac{N_W}{N_M+N_W}$
- $p_{kM} = \frac{n_{kM}}{n_{kM}+n_{kW}}$
- $p_{kW} = \frac{n_{kW}}{n_{kM}+n_{kW}}$

The Duncan index of segregation is given by

$$DI = \frac{1}{2}\sum_k \left| \frac{n_{kM}}{N_M} - \frac{n_{kW}}{N_W} \right| \tag{9}$$

In this same situation (not looking across firms, only one group), our index is given by

$$OS = \omega \sum_k \theta_k (|P_M - p_{kM}| + |P_W - p_{kW}|) \tag{10}$$

First, remember that

$$\omega = \frac{1}{2(1 - \min_g P_g)} \tag{11}$$

In the case of two groups, this can be simplified. Assume that there are more men than women in the labor force (the final result is generalizable to the reverse). Then we have

$$\omega = \frac{1}{2\max(P_W, P_M)} = \frac{N_W + N_M}{2N_M} \tag{12}$$

Additionally,

$$\theta_k = \frac{n_{kM} + n_{kW}}{N_M + N_W} \tag{13}$$

Substituting equations 4 and 5, as well as the notation definitions at the start, into equation 2, we can rewrite our index as

$$GI = \frac{N_W + N_M}{2N_M} \sum_k \frac{n_{kM} + n_{kW}}{N_M + N_W} \left(\left| \frac{N_M}{N_M + N_W} - \frac{n_{kM}}{n_{kM} + n_{kW}} \right| + \left| \frac{N_W}{N_M + N_W} - \frac{n_{kW}}{n_{kM} + n_{kW}} \right| \right) \tag{14}$$

Now note that

$$\begin{aligned} \left| \frac{N_W}{N_M + N_W} - \frac{n_{kW}}{n_{kM} + n_{kW}} \right| &= \left| 1 - \frac{N_M}{N_M + N_W} - (1 - \frac{n_{kM}}{n_{kM} + n_{kW}}) \right| && (15) \\ &= \left| \frac{N_M}{N_M + N_W} - \frac{n_{kM}}{n_{kM} + n_{kW}} \right| && (16) \end{aligned}$$

We substitute equation 8 into equation 6, in addition to some simplification, and have

$$GI = \frac{1}{N_M} \sum_k (n_{kM} + n_{kW}) \left| \frac{N_M}{N_M + N_W} - \frac{n_{kM}}{n_{kM} + n_{kW}} \right| \tag{17}$$

Bringing the always-positive $(n_{kM} + n_{kW})$ into the absolute value function and distributing, we have

$$GI = \frac{1}{N_M} \sum_k \left| \frac{N_M(n_{kM} + n_{kW})}{N_M + N_W} - n_{kM} \right| \tag{18}$$

$$= \frac{1}{N_M} \sum_k \left| \frac{N_M n_{kM} + N_M n_{kW} - N_M n_{kM} - N_M n_{kM}}{N_M + N_W} \right| \tag{19}$$

$$= \frac{1}{N_M(N_M + N_W)} \sum_k |N_M n_{kW} - N_M n_{kM}| \tag{20}$$

$$= \frac{N_W}{N_M + N_W} \sum_k \left| \frac{n_{kW}}{N_W} - \frac{n_{kM}}{N_M} \right| \tag{21}$$

If men and women are equally numbered in the economy, then the first ratio is equal to 1/2, and the entire index is identical to the Duncan index shown in equation 1.

Ultimately then, our index is a generalization of the Duncan Index in three ways:

- Our index allows for more than just two groups
- Our index allows for estimation at the firm level within an economy
- Our index allows for a decomposition of within-firm and between-firm occupational segregation

Which is to say, the Duncan index can be considered a special case of our index where there are only two groups of equal size in the economy and the index is calculated across the entire economy, and not for subgroups (such as firms).

Appendix B: Imputation of Mean and 95th Percentile of Indices Under Employment-at-Random

Employment-at-random assumes that the count of workers in any occupation has probability of belonging to a group (e.g., male or female) equal to the population proportion of that group. For tractability, we leverage a two-stage trained random forest model on the simulated distributions of the indices.

Note, each of the five occupational segregation metrics can be expressed as a weighted mean of demeaned multinomial random variables:

$$
\begin{aligned}
OS_j^{overall} &= \omega \sum_{k\in K} \theta_{jk} \sum_{g\in G} \left|p_{g|jk} - P_g\right| \\
&= \omega \sum_{k\in K} \theta_{jk} \sum_{g\in G} \left|\frac{n_{g|jk}}{n_{jk}} - P_g\right| \\
&= \omega \sum_{k\in K} \theta_{jk} \left[\frac{1}{n_{jk}} \sum_{g\in G} \left|n_{g|jk} - n_{jk}P_g\right|\right]
\end{aligned}
$$

We can do so for all five metrics:

$$
\begin{aligned}
OS_j^{overall} &= \omega \sum_{k\in K} \theta_{jk} \left[\frac{1}{n_{jk}} \sum_{g\in G} \left|n_{g|jk} - n_{jk}P_g\right|\right] \\
OS_j^{within-firm} &= \omega \sum_{k\in K} \theta_{jk} \left[\frac{1}{n_{jk}} \sum_{g\in G} \left|n_{g|jk} - n_{jk}p_{g|j}\right|\right] \\
OS_j^{between-firm} &= \omega \sum_{k\in K} \theta_{jk} \left[\frac{1}{n_{j}} \sum_{g\in G} \left|n_{g|j} - n_{j}P_g\right|\right] \\
OS_j^{within-occupation} &= \omega \sum_{k\in K} \theta_{jk} \left[\frac{1}{n_{jk}} \sum_{g\in G} \left|n_{g|jk} - n_{jk}P_{g|k}\right|\right] \\
OS_j^{between-occupation} &= \omega \sum_{k\in K} \theta_{jk} \left[\frac{1}{n_{k}} \sum_{g\in G} \left|n_{g|k} - n_{k}P_g\right|\right]
\end{aligned}
$$

In other words, we can generalize the random variable underlying each index model m as

$$
os_j^m = \omega \sum_{k\in K} \theta_{jk} \left[\frac{1}{n_m} \sum_{g\in G} L1(n_m, P_m)\right] = \omega \sum_{k\in K} \theta_{jk} \left[\frac{1}{n_m} \sum_{g\in G} \left|X_{g|m} - n_m P_{g|m}\right|\right]
$$

That is, we have the normed (ω) and weighted mean (θ_{jk}) of the mean absolute deviation ($L1$ distance metric) between the random variable under representative employment of the number of people in each group (based on a binomial distribution with $X_m =$

$(x_{1|m}, ..., x_{G|m}) \sim Multinomial(n_m, P_m))$ and the expected number of workers given the target probabilities $n_m * P_m$.

It is computationally intractable for us to simulate the distribution of os_j^m for every firm. Instead, we simulate it for a representative sample of firms and then predict the relevant moments of the distribution using a two-layered random forest model. This allows us to predict the moments using attributes of n_m, P_m as the predictive features instead of simulating for the remaining firms.

Specifically, we take the following steps

Step 1: Create random sample

We create a random sample of firms using Latin hypercube sampling.

Step 2: Estimate prediction models of $os_m(n_m, P_m)$

Step 2.1: Simulate $L1(n_m, P_m)$

For the training set, we next simulate $L1(n_m, P_m)$. We do so by taking R draws from the target multinomial distribution for each firm.

Step 2.2: Compute moments of $L1(n_m, P_m)$

Using the simulated distributions, we next calculate several target empirical moments of the resulting distribution of $L1(n_m, P_m)$:

- Mean
- Standard deviation
- The following quantiles: 1st, 25th, 50th, 75th, 95th, and 99th

We use as feature predictors of these moments the two inputs n_m and P_m (where P_m is a G-length vector), as these fully determine $L1(N_m, P_m)$. We tested versions where we additionally train on $\ln(n_m)$ and $1(n_m = 1)$ to help with the model training, but improvements to the model prediction were minimal.

Some of the models have stochastic target proportions, specifically the within-firm model (where $p_{g|j}$) is non-deterministic for a given firm and the within-occupation metric ($P_{g|k}$, even though it is at the national level, is stochastic with respect to this set-up). We also random draw the underlying n's in the simulations in order to calculate those target probabilities (the proportion of workers at firm j that belong to group g and the proportion of workers in the economy that belong to group g, respectively). The moments are still calculated in the same way, we just introduce this new source of randomness into the simulation, and base the left-side distribution based on the draw's relevant n.

Step 2.3: Fit random forest models of inner elements $moment(L1) = f_{moment}(n_m, P_m)$

Next, we fit a random forest model for each of the moments of $L1$ described above (the mean, standard deviation, and quantiles).

Step 2.4: Fit random forest model for mean$(os_m(n_m, P_m))$ and $p95(os_m(n_m, P_m))$ as functions of several metrics

We next fit a random forest model based on the following moments

- For each simulated (not predicted) moment from Step 2.3
  - The weighted mean (with weights given by θ_j)
  - The weighted standard deviation (with weights given by θ_j)
- Several attributes of n_m across all occupations, including
  - The sum (total number of workers)
  - The mean (average number of workers in each occupation
  - The number of occupations (count of $n_m > 0$)
  - The maximum (how many workers are in the most common occupation in the firm)
  - The minimum (how many workers are in the least common occupation in the firm)

We fit two models: mean$(os_m(n_m, P_m))$ and $p95(os_m(n_m, P_m))$

Importantly, to capture the full variances and structure, we fit the models based not on the predicted moments, but on the empirical moments (i.e. not using Step 2.3's inner random forest model for the outcomes, but only for the inputs).

Step 3: Predict $os_m(n_m, P_m)$

We next fit the predictions of each $os_m(n_m, P_m)$ for all firms in the subsample, using either the fitted random forests from a full-subsample trained models approach, or from a sample-split of a portion of the firms. The former is used for validation (step 4 below), while the latter is used for prediction on the full sample (step 5 below).

The model works by creating one row per occupation per firm and using the fit inner random forest model to predict the moments of $L1$, and then using these and the other features listed in step 2.4 to predict the final two moments. These are then retained with one row per firm.

Step 4: Validation

For validation purposes, we randomly split the sampling frame into a training group and a testing group. We simulate the true distributions of representative hiring for all firms in the sampling frame, but only estimate the random forest models on the training group. We then predict the representative hiring metric moments for all firms in the sampling frame using the estimates of the random forest models from the training group. This provides in-sample and out-of-sample predictions.

The predictions are very good, including for the out-of-sample validation firms. The correlation coefficients between the actual metric and the predicted one from the two stage random forests is typically above 0.99, and is never below 0.94.

Step 5: Storing models for full-sample prediction

We estimate the random forest. models on the full sampling frame after validation, and store these for prediction of the representative hiring moments for all firms, not just those in the sampling frame.

Figure 8: Validation for overall occupational segregation

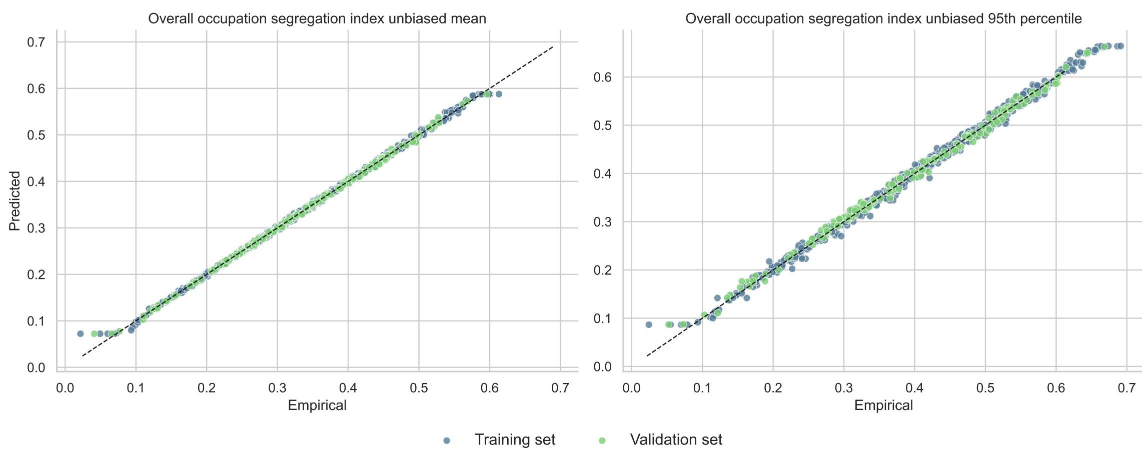


Figure 9: Validation for within-firm occupational segregation

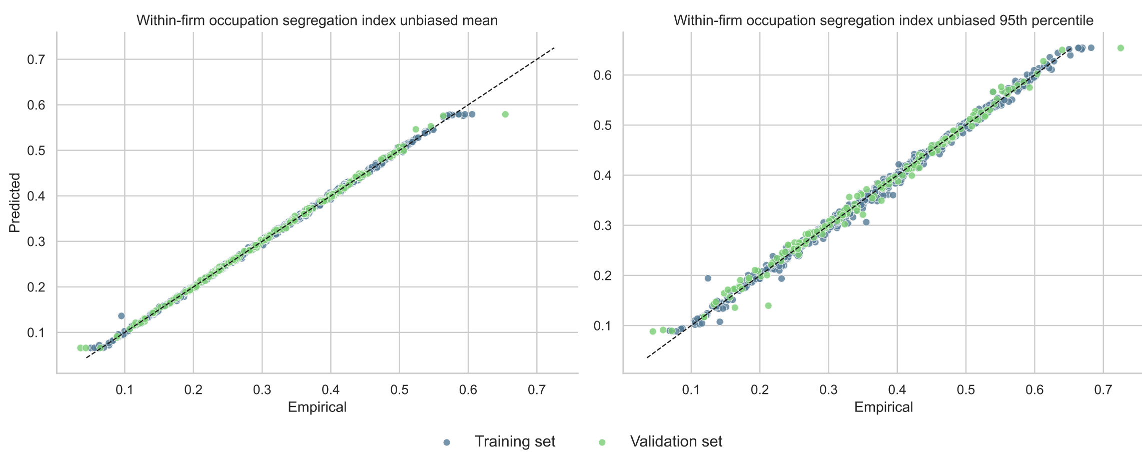


Figure 10: Validation for between-firm occupational segregation

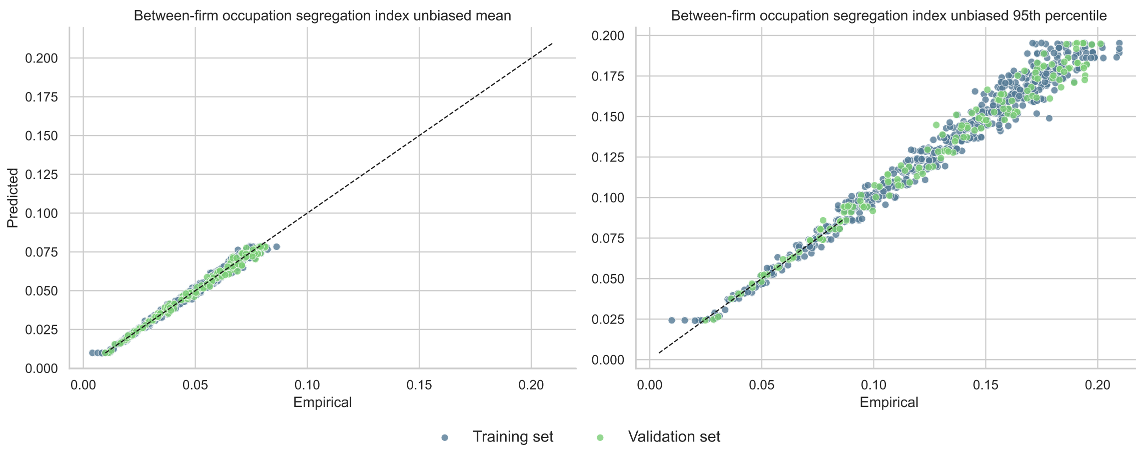


Figure 11: Validation for within-occupation occupational segregation

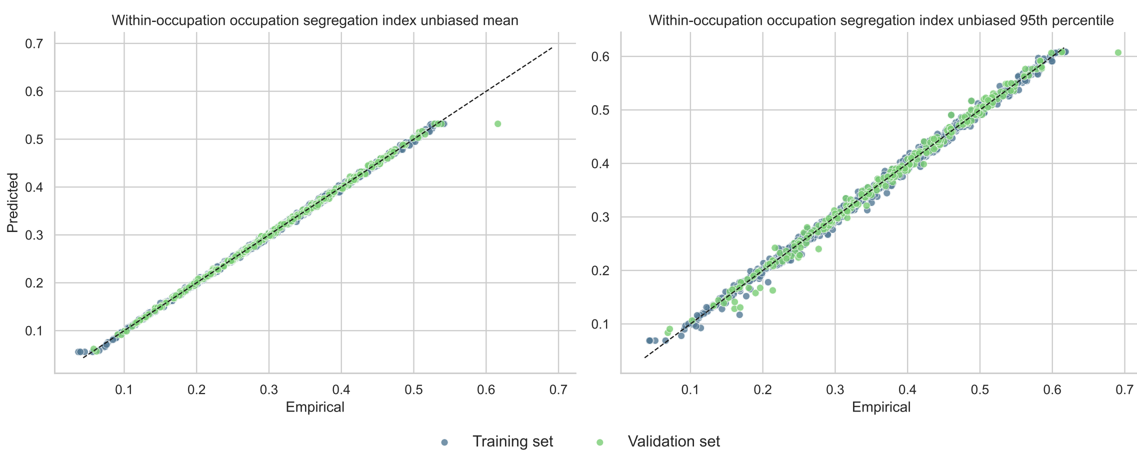


Figure 12: Validation for between-occupation occupational segregation

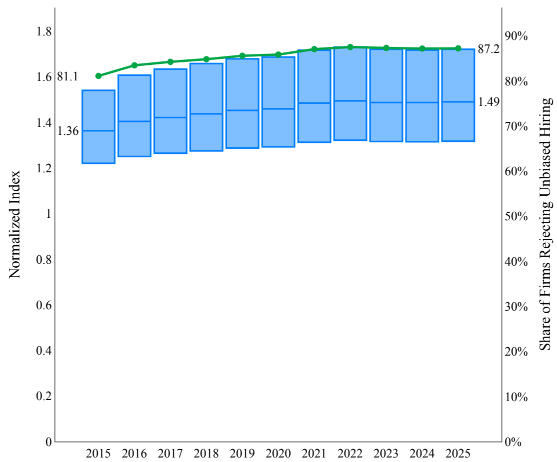
Appendix C: Additional Tables and Figures

Table 6: Summary Statistics of Segregation Indices for Binary Education (BA+ vs Less than BA)

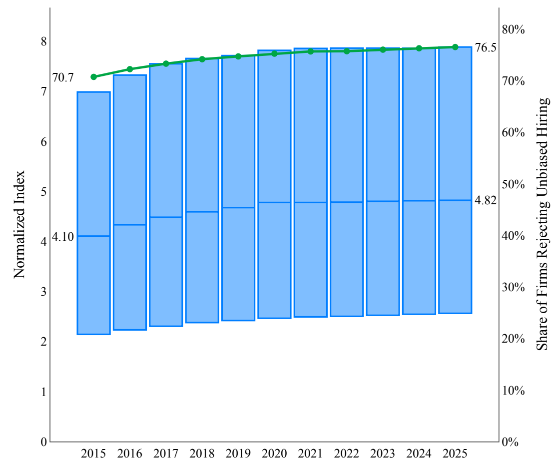
	Unadjusted mean	Representative target mean	Normed mean	Share rejecting representative hiring
Between firm	0.078	0.039	3.905	0.552
	(0.079)	(0.024)	(4.612)	(0.497)
Between occupation	0.086	0.003	51.489	0.998
	(0.043)	(0.004)	(33.166)	(0.039)
Overall	0.165	0.134	1.512	0.496
	(0.082)	(0.054)	(0.967)	(0.500)
Within firm	0.132	0.126	1.215	0.309
	(0.083)	(0.078)	(0.417)	(0.462)
Within occupation	0.126	0.118	1.247	0.316
	(0.070)	(0.050)	(0.677)	(0.465)

Note: Means, with standard deviations in parentheses.

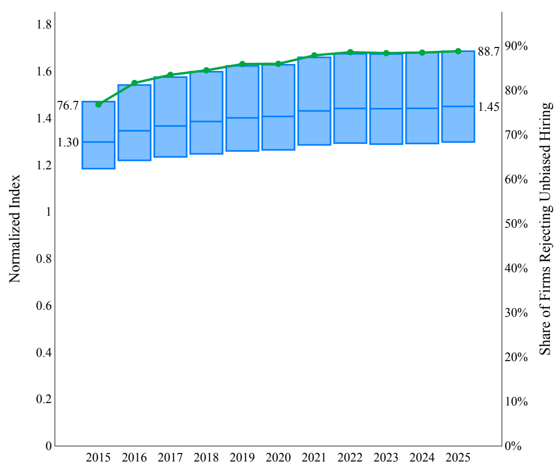
Figure 13: Occupational Segregation Indices: Within vs. Between Firm and Occupation (Gender)



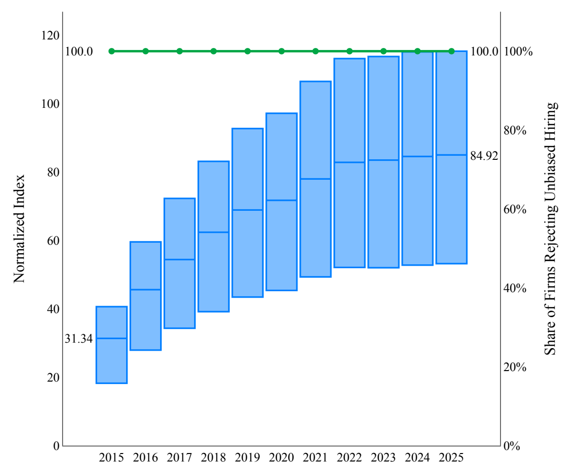
(a) Within-firm



(b) Between-firm



(c) Within-occupation



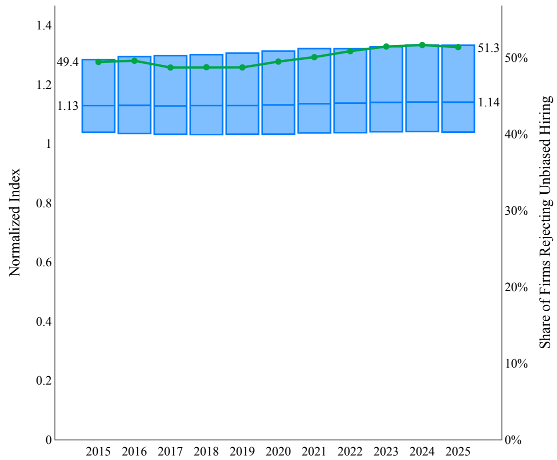
(d) Between-occupation

Note: Box plots represent the interquartile range and median. The scatter plots show the percent of firms whose occupational segregation index is statistically differentiable from the distribution of the index under representative hiring.

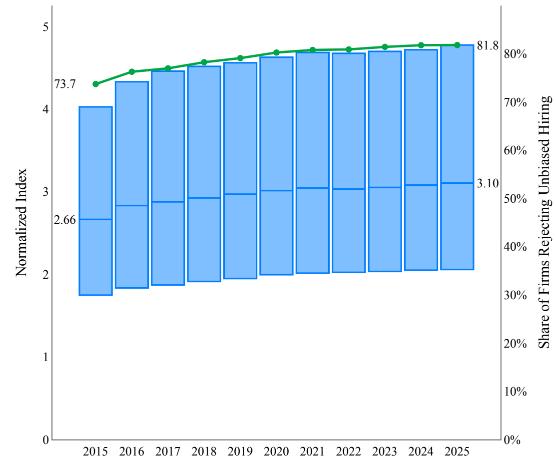
Table 7: Industry-level segregation measures

	Within firm	Between firm	Within occupation	B
Accommodation and Food Services	1.545	2.632	1.480	
Administrative and Support Services	1.491	4.725	1.513	
Construction	1.729	6.981	1.660	
Consumer Services	1.433	4.968	1.466	
Education	1.409	4.445	1.467	
Energy Technology	1.493	7.643	1.466	
Entertainment Providers	1.400	3.904	1.471	
Farming, Ranching, Forestry	1.441	3.581	1.406	
Financial Services	1.489	3.169	1.438	
Government Administration	1.442	4.025	1.416	
Holding Companies	1.400	3.249	1.305	
Hospitals and Health Care	1.599	7.632	1.435	
Manufacturing	1.482	5.165	1.403	
Oil, Gas, and Mining	1.579	6.707	1.416	
Parts Distribution	2.398	19.893	2.339	
Professional Services	1.498	4.510	1.456	
Real Estate and Equipment Rental Services	1.338	3.861	1.514	
Retail	1.481	5.205	1.666	
Technology, Information and Media	1.501	4.861	1.429	
Transportation, Logistics, Supply Chain and Storage	1.542	4.687	1.438	
Unknown	1.371	5.158	1.400	
Utilities	1.515	5.532	1.378	
Wholesale	1.521	5.714	1.499	

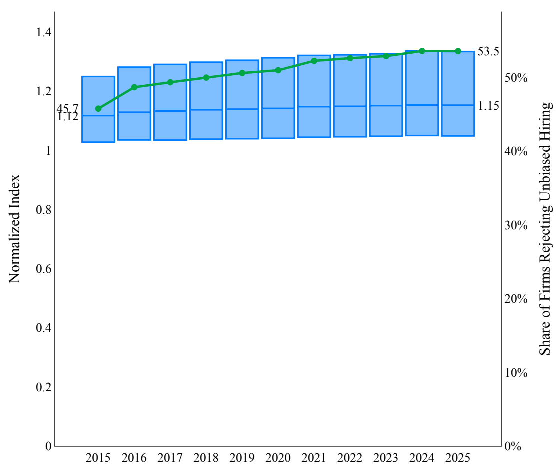
Figure 14: Occupational Segregation Indices: Within vs. Between Firm and Occupation (Education)



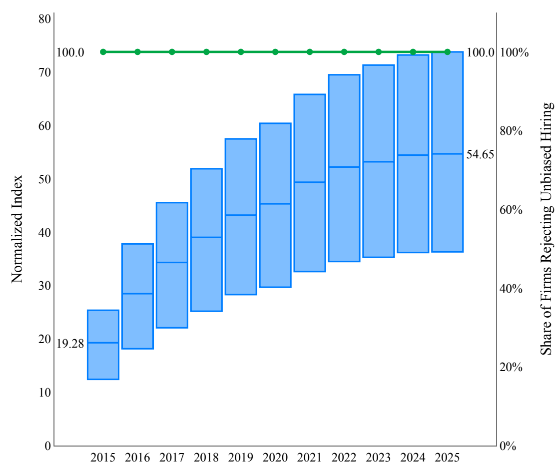
(a) Within-firm



(b) Between-firm



(c) Within-occupation



(d) Between-occupation

Note: Box plots represent the interquartile range and median. The scatter plots show the percent of firms whose occupational segregation index is statistically differentiable from the distribution of the index under representative hiring.

Appendix D: Theoretical Model of Occupational Segregation and Firm Outcomes

This appendix develops the two-occupation, two-group model underlying the four hypotheses in Section 4.

Setup. Consider a firm with two occupations, H (higher-paying) and L (lower-paying), and two equally-sized groups of workers, M and W (e.g., men and women), each with total labor supply normalized to 1. Wages $w_H > w_L$ are determined by labor market clearing: the firm faces downward-sloping labor demand curves $D_H(w_H)$ and $D_L(w_L)$, with $D'_k(\cdot) < 0$. Let $\ell^g_H \in [0, 1]$ denote the share of group $g \in \{M, W\}$ working in occupation H, with $\ell^g_L = 1 - \ell^g_H$. Total labor supply to each occupation equals the sum of allocations across groups, and wages adjust to clear each occupation's market.

Case 1: No Segregation. Both groups allocate labor freely. In the symmetric equilibrium, both groups allocate share $\ell^* \in (0, 1)$ to H. Total labor supply is $N^*_H = 2\ell^*$ to occupation H and $N^*_L = 2(1 - \ell^*)$ to occupation L, with equilibrium wages $w^*_H = w_H(2\ell^*)$ and $w^*_L = w_L(2(1 - \ell^*))$.

Case 2: Full Segregation. Group W is excluded from H and works entirely in L: $\ell^W_H = 0$, $\ell^W_L = 1$. Group M works entirely in H: $\ell^M_H = 1$, $\ell^M_L = 0$. Labor supply becomes $N^{\text{seg}}_H = 1$ and $N^{\text{seg}}_L = 1$.

The key sufficient condition for the following results is that W would have supplied some labor to H absent segregation ($\ell^* > 0$). In the empirically relevant case where H is a minority occupation ($\ell^* < 1/2$, e.g., management), full segregation reduces supply to H from $2\ell^*$ to 1 and increases supply to L from $2(1 - \ell^*)$ to 1.[2] Relative to Case 1:

$$\Delta w_H \equiv w^{\text{seg}}_H - w^*_H > 0 \qquad \text{(supply to } H \text{ falls, wage rises)} \tag{22}$$

$$\Delta w_L \equiv w^{\text{seg}}_L - w^*_L < 0 \qquad \text{(supply to } L \text{ rises, wage falls)} \tag{23}$$

Total employment is held fixed: $dn_H = -dn_L$.

[2] If $\ell^* > 1/2$, the directional wage effects reverse. The empirically natural case is $\ell^* < 1/2$, where most workers are in the lower-paying occupation and segregation sends the excluded group from a minority occupation into the majority one.

Hypothesis 1: Average Wages. The firm's average wage is $\bar{w} = \alpha_H w_H + \alpha_L w_L$, where $\alpha_k = n_k/n$ is occupation k's employment share. Taking a total derivative and imposing fixed total employment:

$$d\bar{w} = \underbrace{\alpha_H\, dw_H + \alpha_L\, dw_L}_{\text{price-shift effect}} + \underbrace{(w_H - w_L)\, d\alpha_H}_{\text{composition effect}} \tag{24}$$

The *price-shift effect* is ambiguous in sign: $dw_H > 0$ contributes positively and $dw_L < 0$ contributes negatively, with the net sign determined by relative demand elasticities and occupational employment shares.

The *composition effect* is unambiguously negative: as W is excluded from H, employment shifts toward the lower-paying occupation ($d\alpha_H < 0$), and since $w_H > w_L$, the product $(w_H - w_L)\, d\alpha_H < 0$. This direct reallocation of workers from higher- to lower-paying jobs depresses average pay. The prediction of lower average wages follows from the sum of an ambiguous price-shift and an unambiguous negative composition effect.

Hypothesis 1: Firms with occupational segregation should have lower paying jobs than firms with no occupational segregation, all else equal.

Hypothesis 2: Pay Dispersion. In this model, the 90th percentile of firm pay corresponds to w_H and the 10th percentile to w_L. The 90-10 ratio is $R = w_H/w_L$. Moving from Case 1 to Case 2 raises w_H and lowers w_L:

$$\frac{\partial R}{\partial\, \text{seg}} = \frac{\Delta w_H}{w_L^*} + \frac{w_H^*\, |\Delta w_L|}{(w_L^*)^2} > 0 \tag{25}$$

Both terms are strictly positive. This prediction is unambiguous under any downward-sloping demand and does not depend on relative elasticities.

Hypothesis 2: Firms with occupational segregation should have a larger ratio of the 90th to 10th percentile of pay (greater compensation variance).

Hypothesis 3: Applications Per Post. Let A_k and V_k denote applications and vacancies in occupation k. Under no segregation, vacancies in H attract applicants from both M and W. Under full segregation, W is excluded from H: the applicant pool for H contracts to M only, reducing A_H by W's contribution (approximately halving raw applicant volume if groups applied at equal rates).

For occupation L: the applicant pool expands as W is redirected there, but V_L also rises as the employment share of L increases. Whether A_L/V_L rises or falls is therefore uncertain. Moreover, workers better suited for H who are constrained to L represent lower-quality matches, increasing screening costs and reducing effective application quality.

The firm's aggregate applications-per-post ratio is:

$$\frac{A}{V} = \frac{A_H + A_L}{V_H + V_L} \tag{26}$$

The reduction in A_H is a large, direct effect on the firm's most contested roles. The change in A_L/V_L is uncertain in sign and smaller in magnitude. The dominant effect is the thinned pool in H, yielding a net reduction in applications per vacancy.

Hypothesis 3: Firms with occupational segregation should have fewer applications per job post.

Hypothesis 4: New-Hires-to-Employees Ratio. Let r_k be the hiring rate in occupation k (new hires divided by employment). The firm-level ratio is $r = (r_H n_H + r_L n_L)/n$. Under segregation, higher wages in H reduce labor demand there (by a magnitude proportional to η_H, the demand elasticity), while lower wages in L increase demand (proportional to η_L). The net effect:

$$dr \;\propto\; \eta_L \, |\Delta w_L| - \eta_H \, |\Delta w_H| \tag{27}$$

is indeterminate: if labor demand in L is more elastic, the hiring rate rises; if demand in H is more elastic, it falls. Since Δw_H and Δw_L are themselves determined by the demand-curve slopes, no further sign restrictions follow without additional parametric assumptions.

Hypothesis 4: Firms with occupational segregation have ambiguous differences in new-hires-to-employees ratios, depending on relative demand elasticities across occupations.