Decomposing Changes in the Gender Wage Gap over Worker Careers*

Keyon Vafa† Susan Athey ‡ David M. Blei§

July 26, 2023

Abstract

A large literature in labor economics seeks to decompose observed gender wage gaps (GWGs) into different sources, including portions explained by cross-gender differences in education, occupation, and experience. This paper provides new methods for decomposing GWGs and applies them to data from the Panel Study of Income Dynamics (PSID). We build upon a recent technique that develops a "foundation model" of career transitions estimated using a large dataset of resumes, creating low-dimensional representations of worker histories. First, we develop a method for fine-tuning the foundation model to predict wages while ensuring that the representations do not omit features of history whose exclusion would bias decompositions. Second, when predicting wages in the PSID, we show that this method better predicts wages relative to regression-based approaches that use hand-crafted summary statistics of worker history. Third, we study GWG decompositions in the PSID, showing that relative to prior approaches, our method for including worker history explains a larger share of the GWG. Fourth, we introduce a novel decomposition of the change in GWGs between two periods, one early in careers and one later, for workers in the workforce in both periods. Fifth, we apply this decomposition, decomposing changes in GWGs in the PSID over 12-year intervals into two sources. The first source is predictable changes in GWGs associated with gender gaps in initial characteristics; we estimate that this source closes the GWG over time. The second source is differences in worker transitions; we estimate that this increases the GWG over time. We show that when analyzing changes in the GWG over 12-year intervals that begin early in workers' careers, the second effect dominates, and differences in worker transitions lead the GWG to widen over time; in contrast, for intervals that begin later in worker careers, the first effect dominates, and the GWG narrows over time.

*The authors thank Patrick Kline, Suresh Naidu, and Matthew Salganik for helpful comments while preparing this draft. Susan Athey thanks Karthik Rajkumar, Lilia Chang and Lisa Simon with whom she first began empirically exploring the problem of modeling labor market transitions. The authors also thank Zippia for generously sharing the dataset of resumes used in this paper.

†Department of Computer Science, Columbia University

‡Graduate School of Business, Stanford University

§Department of Computer Science and Department of Statistics, Columbia University

1 Introduction

The gender wage gap (GWG) is the difference in average wages earned by male and female workers. In the United States, the average female hourly wage is about 80% of the male hourly wage, a fact that has been the subject of an enormous body of research (Blau & Kahn, 2017). Some strands of literature focus on female labor force participation (Heckman & MaCurdy, 1980; Killingsworth & Heckman, 1986), while others focus on sources of differences between men and women, such as child care responsibilities (Goldin, 2006, 2014) or differences in geographic mobility (Benson, 2014).

This paper builds on another part of the literature, which seeks to explain the GWG itself through observable factors of the labor force. This literature is motivated by the recognition that the male and female labor forces differ in many observable ways, including in factors such as education, occupation, and industry. This literature seeks to *decompose* the GWG into (i) a part that can be explained by these factors, and (ii) the unexplained gap, which arises when men and women with the same observable characteristics receive different wages (Blinder, 1973; Oaxaca, 1973; Blau & Kahn, 2017). These decompositions can provide guidance to policy. If the unexplained part of the GWG is large, then interventions designed to address problems of bargaining or fairness in wage setting may be considered. On the other hand, if an important component of the gender wage gap is due to gender differences in worker characteristics, then attempts to close the gap might involve interventions that target education and career choices.

This paper contributes to both the methodological and empirical literatures surrounding decompositions of the GWG. There are two parts to the methodological contribution. First, we develop new methods based on machine learning to more fully account for full worker history when predicting wages, and thus explain a greater portion of the overall GWG. In accounting for each individual's full history of jobs, we introduce a new method that reduces the scope for omitted variables to bias the decompositions. Second, we develop novel decompositions of the GWG that exploit our richer model of how wages depend on history, and can be used to further refine the sources of the GWG into the impact of workers' early career characteristics, and the impact of

differences in career transitions over time. These decompositions can guide policy in new ways; for example, if the GWG is exacerbated over time in workers' careers by differential career transitions, then interventions that relate to mentoring and promotions (on the firm side) and women's choices and family support (on the worker side) may be impactful.

On the empirical side, we apply our methods to survey data from the Panel Study of Income Dynamics (Panel Study of Income Dynamics, public use dataset, 2023). On held-out data, the machine learning model predicts wages well, outperforming classical econometric approaches. We further document novel empirical findings about the sources of the GWG over workers' careers. We study two populations over 12-year intervals: one that begins early in the workers' careers, and the other one beginning later. For the younger population, the GWG increases in magnitude over time, which we find to be driven by differences in career transitions during the interval. For the older population, the GWG narrows over time, where differences in transitions widen the gap, but only modestly. This is more than offset by a second effect, where the characteristics and histories of women at the start of the interval are associated with greater predicted wage growth than men's holding fixed transitions.

In more detail, consider the problem of modeling the impact of worker history on wages. Perhaps due to limited data, much of the literature to date has focused on a relatively limited set of worker characteristics and hand-crafted summary statistics that capture key features of worker history (Blau & Kahn, 2017). However, it has long been understood that existing models may have omitted variables that are correlated with both wage and gender, and which have the potential to explain some of the observed GWG. Ideally, to account for such missing variables, a better wage model would account for the full worker history, rather than simple summary statistics. But it has been difficult to address this problem or assess its magnitude with the most commonly used survey datasets in the U.S., which are too small to fit such complex models.

In this paper, we propose a novel method to incorporate more detail about full worker histories into wage prediction models, and one that is tailored to studying GWGs. Our method builds upon Vafa et al. (2022), which proposes the CAREER model. Inspired by the foundation models used for language modeling (Large Language Models, or LLMs), CAREER is a foundation model based

on 23.7 million resumes that incorporates information about the sequences of jobs in a worker's career (but that does not incorporate individual wage data). Like LLMs, which are designed to predict sequences of words, CAREER's foundation model is based on a transformer neural network (Vaswani et al., 2017) and is designed to predict sequences of jobs. CAREER takes as input a set of jobs, and makes a prediction about the next job. Although the resume data is neither representative nor complete (just as LLM training data may be non-representative), the large quantity of resumes enables the estimation of a low-dimensional representation (formally, a mapping from history to a lower-dimensional vector of latent factors) that captures the elements of history that are useful to predicting career transitions. The representations of worker history learned by CAREER can serve as the "foundation" on which we can build analyses of smaller datasets.

How do we use CAREER to help decompose the GWG? In the literature on modeling language, it is common to "fine-tune" an LLM to a smaller dataset, in order to optimize performance for a particular type of text or for a prediction problem that is slightly different from the one the LLM is designed to solve. For example, LLMs can be tuned to have conversations with users (Ouyang et al., 2022), to classify text documents (Devlin et al., 2019), or to predict sentiment (Wang et al., 2023). In Vafa et al. (2022), CAREER is fine-tuned to predict occupational transitions, using data from the Panel Survey of Income Dynamics (the PSID) and the National Longitudinal Survey of Youth (NLSY). Vafa et al. (2022) show that a fine-tuned CAREER model significantly improves prediction quality on held-out data over standard regression-based models that use the same underlying data.

Thus the first methodological contribution of this paper is a new method of fine-tuning CAREER that is designed to predict wages directly and to accurately decompose the GWG. To this end, it is not sufficient to fine-tune CAREER to only improve the prediction of wages. Rather, we also require worker history representations that can predict mean male wages over the distribution of female histories. If the CAREER representation omits elements of history that help explain the GWG, the magnitude of the explained and unexplained parts of the GWG will be mis-estimated. We describe a necessary and sufficient condition for representations to not omit important variables, which is met when CAREER's representation includes all aspects of history that are correlated with gender. Thus we propose a fine-tuning algorithm that retains elements of history that are correlated with

gender into its objective. With a semi-synthetic study, we show that our method performs well. Our fine-tuned CAREER model enables us to decompose gender wage gaps with representations of history.

Our second methodological contribution is deriving and estimating a novel decomposition of the change in GWGs over a span of a worker's career. In detail, a history-adjusted GWG is one that depends on the distribution of history and covariates, i.e., of job sequences and other factors, for males and females. These gender-specific sequential distributions naturally factor into an initial distribution of the first job and covariates, and then subsequent "transitions" to the next job and next covariates. With this factorization, we will show how to interpretably decompose how the change in GWGs can be explained by gender differences in the initial distribution and gender differences in the transitions. Such a decomposition can help inform policy interventions, whether in providing more opportunities in the beginning of a career or whether to mitigate inequality in transition opportunities.

Finally, we apply our methodology to the PSID. We show that our method improves wage predictions, increasing the R^2 of wage regression in held-out test data from 0.456 to 0.527. We further show that our wage model is well-calibrated for both men and women. We investigate the learned representations of history to understand why including job histories improves predictive performance. We demonstrate that occupation labels omit important elements of history that refine an individual's current occupation; for example, our method groups together managers who were previously engineers. Next, we introduce a heuristic for reassigning occupation labels based on CAREER's representation of history. We show that baseline models trained on the reassigned labels make better predictions on held-out data.

We then apply our model to study the GWG. We first use our model to decompose cross-sectional wage gaps. This decomposition shows that existing wage models suffer from omitted variable bias, and that about 25% of the gender wage gap that is unexplained with a simpler regression model can be explained when using a richer model of history.

We then document how gender gaps change over workers' careers. We focus on studying the change in the GWG for a consistent set of workers who are active at two endpoints of a specified

time interval. We consider two populations of individuals over 12-year intervals: one that begins early in the workers' careers (ages 25-35), and the other later in careers (ages 40-50).

For the younger population, the GWG widens by 0.049 log points over 12 years. Changes in history explain 0.037 log points of this increase. Our decomposition divides the change explained by history into two components: one explained by differences in male and female initial histories, the other explained by differences in male and female history transitions over the 12-year interval. We find that the change in the explained wage gap is being driven by differences in transitions; although the explained gap would narrow for the same transitions, females transition to substantially lower-value histories than males. Comparisons to baseline models that do not consider full history suggest that these transitions have lower value not only due to workforce participation spells and final occupations, but due to the specific occupations held between periods.

For the older population, our decomposition reveals a different effect. The GWG decreases in magnitude after 12 years, as does the wage gap explained by history. However, the explained gap decrease is not attributed to females making higher-value transitions; on average, they make slightly lower-value transitions than males, costing them 0.023 log points relative to male transitions. Rather, the explained gap decrease is attributed to differences in female and male starting histories, with the same transitions earning 0.038 more log points for initial female histories than male histories.

1.1 Related Work

The gender wage gap has been the subject of an enormous literature (Altonji & Blank, 1999; Blau & Kahn, 2017). Some strands of the literature have studied female labor force participation (Heckman & MaCurdy, 1980; Killingsworth & Heckman, 1986). Others have focused on the sources of differences (Goldin, 2014), such as child care responsibilities and maternity leave (Waldfogel, 1998; Schönberg & Ludsteck, 2007; Miller, 2011). This paper follows a line of work that aims to decompose the gender wage gap, attributing differences in male and female wages to differences in observed factors (Blinder, 1973; Oaxaca, 1973; Fortin et al., 2011). Our paper does not attempt to model the underlying forces that lead to gender wage gaps. Instead, we propose a

method for decomposing the change in the explained wage gap into differences in male and female starting characteristics and transitions.

The seminal works of Blinder (1973) and Oaxaca (1973) provide methods for decomposing the gender wage gap into explained and unexplained components. While these methods were developed to analyze a single cross-section of workers, extensions proposed by Smith & Welch (1989) and Juhn et al. (1993) decompose changes in general gaps into changes in explained and unexplained components. These extensions have been applied to study changing gender wage gaps in a wide range of labor markets, including that of the United States between 1980-2010 (Blau & Kahn, 2017), South Africa in the post-apartheid period (Shepherd et al., 2008), and Japan, Russia, and the United States between 1993-2000 (Johnes & Tanaka, 2008). While these papers study the change in wage gaps over time, they do not study the change in the gender wage gap over the careers of a fixed population of individuals, as is the goal of our paper.

Our analysis studies the early-career gender wage gap, attributing the increase over time in the wage gap explained by history to differential male and female transitions. Early-career wages are important for life-time wage growth (Murphy & Welch, 1990; Rubinstein & Weiss, 2006; Guvenen et al., 2021), and our work complements other decompositions of the gender wage gap in this period. Loprest (1992) decomposes wage growth differences among men and women over their first four years of working full-time, finding that women who change jobs have less wage growth than men who change jobs in that period. Manning & Swaffield (2008) use a panel survey to study the early-career gender pay gap in the UK, finding a large early-career unexplained gap after adjusting for human capital factors, job-shopping, and psychological theories. Menzel & Woodruff (2021) analyze the changing wage gap over the careers of garment workers in Bangladesh, focusing on gender differences in promotion rates. Much of this work has focused on the highly educated (Black et al., 2008; Goldin, 2014): Wood et al. (1993) study the widening earnings gap for law school graduates, while Bertrand et al. (2010) analyze the early-career divergence of male and female MBA graduate earnings, attributing the divergence to differences in pre-MBA training, career interruptions, and weekly working hours. Meanwhile, Goldin et al. (2017) study the change in the gender earnings gap after schooling to understand the role of employment differences across

establishments versus differences in pay within establishments in contributing to the widening gap. Similar to our paper, these methods study how the wage gap changes over time for a fixed group of individuals. However, unlike these papers, our goal is to further decompose the change in the explained gap into effects of differential starting characteristics and differential transitions. The methodology we develop here can complement the analyses in these papers.

The decomposition in this paper has a similar motivation to one described by Loprest (1992), who breaks down the difference in wage growth for males and females who change jobs into two terms: the difference in growth for individuals who transition from full to part-time work and the difference for those who do not. This decomposition does not aim to isolate the gap explained by differential transitions, so the wage functions are not held fixed. This means that the wage gap difference for the transition subpopulation can be nonzero even if males and females transition at the same rates. In contrast, the decomposition we propose breaks up the difference in explained gaps into two terms, one explained solely by differential starting characteristics, the other explained solely by differential transitions.

One strand of the gender wage gap literature has used detailed labor force trajectory information from matched worker-firm data to decompose earnings gaps and changes in gaps over time. Monti et al. (2020) demonstrate that characteristics of past employers help explain the gender wage gap. Other approaches examine individuals transitioning between firms in order to isolate the impact of individual characteristics versus firm characteristics on wage, adapting the model developed by Abowd et al. (1999) (hereafter AKM). For example, Card et al. (2016) decompose gender differences in firm pay among Portuguese workers into sorting and bargaining effects. Barth et al. (2017) apply the AKM model to identify an establishment component of earnings and then decompose changes in gender earnings gaps between two ages into changes in individual and establishment components. While these studies use detailed longitudinal data in their analyses, they do not aim to decompose changes in the explained gender wage gaps over careers into effects of differential starting characteristics and differential transitions. The dataset we use in our analysis does not have detailed firm data, so we only model occupational trajectories. However, the methodology we develop can also incorporate firm trajectories when this data is available.

This paper develops a model that relates occupational trajectories to wage. We find that detailed representations of labor experience explain more of the wage gap than coarse-grained summary statistics of labor experience, such as total years worked. This corroborates previous findings that incomplete measures of experience can discard factors that help explain the wage gap. For example, prior studies have found that potential experience (an inexact measure of experience that does not measure workforce interruptions) explains less of the wage gap than years of actual work experience (Regan & Oaxaca, 2009; Blau & Kahn, 2013). Light & Ureta (1995) estimate a wage model with detailed measures of year-by-year experience, finding that the timing of work experience explains a substantial portion of the wage gap. We develop machine learning methodology to condition on an individual's entire occupational trajectory. This procedure provides measures of experience that explain more of the wage gap than traditional measures.

The decomposition developed in this paper does not rely on linearity or other functional forms. In this vein, it is similar to other nonparametric decompositions that do not rely on linearity (Sinning et al., 2008; Breunig & Rospabe, 2004; Ulrick, 2005). The novel aspect of our decomposition is not in the functional form of the underlying wage models, but rather the decomposition of explained wage gap differentials into transition differentials and starting characteristic differentials.

This paper contributes to a line of work that uses machine learning methods to model labor market transitions (Li et al., 2017; Rajkumar, 2021; Vafa et al., 2022). Specifically, we develop a transformer neural network to model wages from job histories. While transformers were initially developed for natural language processing (Vaswani et al., 2017), to model sequences of words in a sentence, they have since been extended to model non-textual data, such as images (Dosovitskiy et al., 2021), music (Huang et al., 2019), and molecular chemistry (Schwaller et al., 2019). The model developed in this paper extends the model proposed by Vafa et al. (2022), which developed a transformer to predict occupational transitions. However, our methodology differs from the methodology in Vafa et al. (2022) in a few ways, due to our focus on using transformers to predict wages from history rather than transitions; see Section 3.1 for more discussion of these differences.

Limitations. Our paper models expected wage functions in order to estimate decompositions of the gender wage gap. However, this is not the only way to decompose wage gaps. Other approaches, such as those based on propensity weighting, can be used to estimate the same quantities; Goraus et al. (2017) and Huber & Solovyeva (2020) find wage gap decompositions to be sensitive to the approach used. Our decision to decompose gender wage gaps by estimating a wage model follows the majority of the gender wage gap literature (Altonji & Blank, 1999; Blau & Kahn, 2017). We plan to compare this approach to other methods in future work.

Wage gap decompositions typically select one group as a reference group. Thus, they are sensitive to the "index problem," as the explained and unexplained wage gaps can vary with the selected reference group (Suh, 2010; Kim, 2010). Solutions have been proposed to avoid this problem, such as by using Shapley values in decompositions (Devicienti, 2010; Kimhi & Hanuka-Taflia, 2019). We do not try to solve this problem. Instead, we follow Blau & Kahn (2017) in using the male wage model as the reference group for the explained gap. We plan to re-run our analyses with the female wage model as the reference group as a robustness check.

2 Decomposing Wage Gap Changes over Careers

Our goal is to investigate how changes in the gender wage gap over careers are related to male and female characteristics evolving in different ways. An important decision for this analysis is the set of characteristics to study. There are some characteristics, such as educational attainment, that contribute to the gender wage gap (Blau & Kahn, 2017) yet may not evolve much over an individual's career. We include these characteristics in our analysis.

Since we are studying the effects of transitions, it is also important to include characteristics that evolve substantially over an individual's career such as labor market experience. Experience can be measured in many ways, and it is frequently summarized with coarse-grained summary statistics like years of experience (Blau & Kahn, 2017). However, broad summaries can discard important aspects about career pathways; two salespeople who transition to managers over a 12-year span may have differences in their work trajectories that account for wage differences. If important aspects of

career trajectories are omitted from a decomposition, the full effect of differential transitions will not be revealed. To obtain a richer decomposition, we need to consider more detailed descriptions of career trajectories. Thus, we include a variable in our decompositions that encodes an individual's entire occupational history, the sequence of occupations over their careers.

Below, we propose a decomposition of the change in gender wage gaps based on history and other covariates. We first follow standard techniques to decompose the change in wage gaps into two components, the change in the amount explained by observable characteristics and the change that is unexplained. We then introduce a decomposition that further breaks up the change in explained wage gaps into two components: a component that captures the effect of differential initial characteristics, and a component that captures the effect of differential transitions.

2.1 Notation

We use capital letters to denote random variables and lower-case letters to denote specific realizations.

For each individual i, we observe log wage $Y_i \in \mathbb{R}$, covariates $X_i \in \mathbb{R}^P$, and gender G_i, which can take on the values "f" or "m". We also observe labor history, $H_i \in \mathscr{H}$, which is a sequence of discrete occupations and years, $H_i = ((H_{i1}, D_{i1}), \ldots, (H_{iL}, D_{iL}))$, and where each occupation label $H_{il} \in \{1, \ldots, J\}$ encodes the occupation an individual worked in during year D_l or their labor status, e.g. "unemployed" or "student."[1] We refer to the observed covariates and history together as "characteristics." We will sometimes use the shorthand $Z_i = (X_i, H_i)$ to refer to histories and covariates together. We use $\nu(x, h|g)$ to denote the distribution of covariates and histories conditional on gender, $X_i, H_i|G_i = g$. Similarly, we will use $\nu(x, h)$ to denote the distribution of covariates and histories over the entire population. We will also use the notation $\nu(G = f|x, h)$ to denote the probability an individual is female given covariates x and histories h.

Some of our analysis involves comparing individuals at two time periods, which for convenience we refer to as $t = 0$ and $t = 1$ (where in our application, these will correspond to two distinct ages of workers, but in other applications the time periods could be calendar years or years of experience).

[1] We use the terms "occupation" and "job" synonymously.

When this is the case, we superscript each variable by the time period. For example, X_i^0 would indicate an individual's covariates at period 0, with X_i^1 denoting their covariates at period 1. We then define the distribution of covariates and histories at each period to be $\nu(x^t, h^t|g)$. We let $\nu(x^1, h^1|x^0, h^0, g)$ denote the conditional distribution of X_i^1 and H_i^1 at period 1 given $G_i = g$ and observed variables $X_i^0 = x^0$ and $H_i^0 = h^0$ at period 0. We refer to $\nu(x^1, h^1|x^0, h^0, g)$ as the **transition distribution** of characteristics for gender g. For any analysis that does not involve changing variables over time, we omit t from the notation.

Define the conditional mean function:

$$\mu_g(x, h) = \mathbb{E}[Y|G = g, X = x, H = h]. \tag{1}$$

We assume the conditional mean function is invariant across periods. We use $\hat{\mu}$ to denote estimates of the conditional mean function.

2.2 Decomposing wage gap changes into differential starting characteristics and transitions

Males and females enter the labor workforce with different characteristics, which can contribute to an early-career gender wage gap (Manning & Swaffield, 2008). However, these characteristics evolve over an individual's career, as does the gender wage gap. Male and female characteristics do not evolve in the same way; in a population of sales representatives, all the same age and five years out of college, the distribution of career pathways for males will differ from that of females. How does the difference in the evolution of these characteristics from fixed starting positions affect the gender wage gap? To study this question, we will propose a decomposition of how wage gaps change over the course of careers.

The raw gender wage gap. We consider a population of male and female salary workers at two time periods, denoted by $t = 0$ and $t = 1$. The time periods are fixed in relation to a worker's career; for example, the two time periods may correspond to a 12-year interval in an individual's career.

We begin by defining the raw gender wage gap, the average difference in expected wages for females and males. Define the raw gender wage gap at period t (WG_t) to be

$$\text{WG}_t = \mathbb{E}_{\nu(z^t|f)}\left[\mu_f(Z^t)\right] - \mathbb{E}_{\nu(z^t|m)}\left[\mu_m(Z^t)\right], \tag{2}$$

where $Z^t = (X^t, H^t)$ refers to both covariates and histories.

The change in the gender wage gap between these two periods is then given by:

$$\begin{aligned}\text{WG}_1 - \text{WG}_0 = &\left(\mathbb{E}_{\nu(z^1|f)}\left[\mu_f(Z^1)\right] - \mathbb{E}_{\nu(z^1|m)}\left[\mu_m(Z^1)\right]\right) &(3)\\ &- \left(\mathbb{E}_{\nu(z^0|f)}\left[\mu_f(Z^0)\right] - \mathbb{E}_{\nu(z^0|m)}\left[\mu_m(Z^0)\right]\right). &(4)\end{aligned}$$

Male and female characteristics evolve between periods 0 and 1. Male characteristics evolve according to $\nu(z^1|z^0, m)$, while female characteristics evolve according to $\nu(z^1|z^0, f)$. We would like to measure how much of the change in the gender wage gap can be attributed to differences in these distributions.

Decomposing the wage gap into explained and unexplained components. We first decompose the gender wage gap at each period separately. Using a classical decomposition (Blinder, 1973; Oaxaca, 1973), the gender wage gap at period t can be divided into explained and unexplained components. By adding and subtracting $\mathbb{E}_{\nu(z^t|f)}\left[\mu_m(Z^t)\right]$ to (2), we arrive at

$$\begin{aligned}\text{WG}_t &= \mathbb{E}_{\nu(z^t|f)}\left[\mu_f(Z^t)\right] - \mathbb{E}_{\nu(z^t|m)}\left[\mu_m(Z^t)\right] &(5)\\ &= \mathbb{E}_{\nu(z^t|f)}\left[\mu_f(Z^t) - \mu_m(Z^t)\right] &(6)\\ &\quad + \mathbb{E}_{\nu(z^t|f)}\left[\mu_m(Z^t)\right] - \mathbb{E}_{\nu(z^t|m)}\left[\mu_m(Z^t)\right]. &(7)\end{aligned}$$

The portion of the gap in (6) is the unexplained wage gap. It measures the average difference in expected wage for females and males with the same characteristics. Meanwhile, (7) is the explained wage gap. It is the portion of the gap attributable to differences in distributions of male and female characteristics. For example, if the unexplained wage gap is zero, then groups of males and females

with similar covariates X and histories H will have no difference in their wages.[2] Instead, the wage gap would be attributed solely to differences in distributions of male and female characteristics.[3]

Decomposing changes in wage gaps. Denoting by UWG_t the unexplained wage gap at period t (6) and by EWG_t the explained wage gap at period t (7), rewrite (5) to (7) as

$$\text{WG}_t = \text{EWG}_t + \text{UWG}_t. \tag{8}$$

The change in the gender wage gap (3) can then be broken up into changes in explained and unexplained wage gaps (Smith & Welch, 1989):

$$\text{WG}_1 - \text{WG}_0 = (\text{EWG}_1 - \text{EWG}_0) + (\text{UWG}_1 - \text{UWG}_0). \tag{9}$$

This decomposition can help relate the change in the gender wage gap to differences in male and female transition distributions. The difference in the unexplained wage gap only involves the distribution of female characteristics, and so is unrelated to a difference in transitions. Thus, we focus on the change in explained wage gaps:

$$\begin{aligned} \text{EWG}_1 - \text{EWG}_0 = \mathbb{E}_{\nu(z^1|f)}\left[\mu_m(Z^1)\right] - \mathbb{E}_{\nu(z^1|m)}\left[\mu_m(Z^1)\right] & \tag{10} \\ - \left(\mathbb{E}_{\nu(z^0|f)}\left[\mu_m(Z^0)\right] - \mathbb{E}_{\nu(z^0|m)}\left[\mu_m(Z^0)\right]\right). & \tag{11} \end{aligned}$$

Our goal is to decompose the difference in explained wage gaps into a portion due to differences in transitions, $\nu(z^1|z^0, f)$ vs. $\nu(z^1|z^0, m)$, and a portion due to differences in initial distributions, $\nu(z^0|m)$ vs. $\nu(z^0|f)$. The difference in explained gaps measures the change in the wage gap due to differences in male and female characteristics between periods. But it doesn't address the role of

[2] The unexplained gender wage gap is not a direct measure of gender discrimination. If the unexplained wage gap for a set of characteristics is zero, it does not imply that there is no wage discrimination; rather, that conditional on the same set of characteristics, males and females do not have pay differences. In this case, some characteristics, such as industry or occupation, may be the result of historic discrimination (Blau & Kahn, 2017). Conversely, a non-zero value of the unexplained wage gap does not imply discrimination, since there may be unmeasured variables that affect wage yet vary between genders.

[3] A similar decomposition could be performed by adding and subtracting the expected female wage for the male characteristics. We follow Blau & Kahn (2017) in focusing on the male model with female characteristics.

differences in *transition distributions*. There can be a large change in the explained wage gap even when the transitions distributions of males and females are identical.

To see why, consider an example where the only characteristic is occupation, and it can take on one of two values: low-paying and high-paying. In this example, 80% of males start in high-paying occupations, while 50% of females start in high-paying occupations. Consider that for both genders, all individuals in high-paying occupations will stay in high-paying occupations, while half of the individuals in low-paying occupations will transition to high-paying occupations. Thus, at the end of the interval, 90% of males are in high-paying occupations, compared to 75% of females. The gap in the percent of females and males in high-paying occupations decreases between time periods, from 30% to 15%. But this is not because females make more valuable transitions; rather, females have more to gain from the same transitions as males because they start out in lower-paying positions on average.

Thus, we further decompose (10) and (11) to clarify the role of transitions. Re-write the change in explained gaps ((10) and (11)) using iterated expectations:

$$\begin{aligned}\text{EWG}_1 - \text{EWG}_0 = {} & \mathbb{E}_{\nu(z^0|f)}\left[\mathbb{E}_{\nu(z^1|z^0,f)}\left[\mu_m(Z^1)|Z^0\right]\right] - \mathbb{E}_{\nu(z^0|m)}\left[\mathbb{E}_{\nu(z^1|z^0,m)}\left[\mu_m(Z^1)|Z^0\right]\right] && (12)\\ & - \left(\mathbb{E}_{\nu(z^0|f)}\left[\mu_m(Z^0)\right] - \mathbb{E}_{\nu(z^0|m)}\left[\mu_m(Z^0)\right]\right). && (13)\end{aligned}$$

Now subtract and add $\mathbb{E}_{\nu(z^0|f)}\left[\mathbb{E}_{\nu(z^1|z^0,m)}\left[\mu_m(Z^1)|Z^0\right]\right]$:

$$\begin{aligned}\text{EWG}_1 - \text{EWG}_0 = {} & \mathbb{E}_{\nu(z^0|f)}\left[\mathbb{E}_{\nu(z^1|z^0,f)}\left[\mu_m(Z^1)|Z^0\right]\right] && (14)\\ & - \mathbb{E}_{\nu(z^0|f)}\left[\mathbb{E}_{\nu(z^1|z^0,m)}\left[\mu_m(Z^1)|Z^0\right]\right] && (15)\\ & + \mathbb{E}_{\nu(z^0|f)}\left[\mathbb{E}_{\nu(z^1|z^0,m)}\left[\mu_m(Z^1)|Z^0\right]\right] && (16)\\ & - \mathbb{E}_{\nu(z^0|m)}\left[\mathbb{E}_{\nu(z^1|z^0,m)}\left[\mu_m(Z^1)|Z^0\right]\right] && (17)\\ & - \left(\mathbb{E}_{\nu(z^0|f)}\left[\mu_m(Z^0)\right] - \mathbb{E}_{\nu(z^0|m)}\left[\mu_m(Z^0)\right]\right). && (18)\end{aligned}$$

This decomposition sheds light on the role of transition distributions in the difference of explained gaps. (14) and (15) measure the value of male transitions relative to those of females, given the

same (female) starting characteristics:

$$\underbrace{\mathbb{E}_{\nu(z^0|f)}\left[\mathbb{E}_{\nu(z^1|z^0,f)}\left[\mu_m(Z^1)|Z^0\right]\right]}_{\substack{\text{expected period 1 wage for female transitions} \\ \text{and female initial characteristics}}} - \underbrace{\mathbb{E}_{\nu(z^0|f)}\left[\mathbb{E}_{\nu(z^1|z^0,m)}\left[\mu_m(Z^1)|Z^0\right]\right]}_{\substack{\text{expected period 1 wage for male transitions} \\ \text{and female initial characteristics}}} \tag{19}$$

We refer to (19) as the **effect of differential transitions**; it keeps the starting characteristics fixed and varies the transitions.

Meanwhile, (16) to (18) capture the expected increase in the gender gap if wages grow according to the average male transition path for both genders:

$$\overbrace{\mathbb{E}_{\nu(z^0|f)}\left[\mathbb{E}_{\nu(z^1|z^0,m)}\left[\mu_m(Z^1)|Z^0\right]\right] - \mathbb{E}_{\nu(z^0|m)}\left[\mathbb{E}_{\nu(z^1|z^0,m)}\left[\mu_m(Z^1)|Z^0\right]\right]}^{\text{expectation of period 1 explained gap if both genders have male transitions}} \tag{20}$$

$$- \underbrace{\left(\mathbb{E}_{\nu(z^0|f)}\left[\mu_m(Z^0)\right] - \mathbb{E}_{\nu(z^0|m)}\left[\mu_m(Z^0)\right]\right)}_{\text{period 0 explained gap}}. \tag{21}$$

We refer to (20) and (21) as the **effect of differential starting characteristics**; it keeps transitions fixed and varies the starting characteristics.

Now, the role of transitions is clarified. In the simplified example from above involving a single characteristic, (19) will be zero since males and females have the same transition distributions. Instead, the change in the explained gap will be attributed to (20) and (21): the explained gap decreases because females have more to gain from making the same transitions as males.

We note that for the components of this decomposition to be non-parametrically identifiable, overlap must hold:

$$\nu(G = f|x^t, h^t) < 1. \tag{22}$$

Overlap is necessary for the non-parametric identifiability of two terms: the first term of the explained wage gap, $\mathbb{E}_{\nu(z^t|f)}[\mu_m(Z^t)]$, and the expected male wage for female initial characteristics and male transitions, $\mathbb{E}_{\nu(z^0|f)}[\mathbb{E}_{\nu(z^1|z^0,m)}[\mu_m(Z^1)|Z^0]]$. The former term requires averaging the male wage function over the distribution of female characteristics, while the latter term averages the male transition distribution from starting female characteristics. In practice, even limited

overlap may make components of this decomposition difficult to estimate. Section 4 discusses a trimming strategy to address limited overlap.

3 Estimation

The decomposition in Section 2 attributes changes in the gender wage gap to differences in male and female transitions and starting characteristics. To estimate these terms from data, we turn to longitudinal survey datasets. These datasets follow individuals over their careers, interviewing them regularly about their wage, occupation, and other characteristics. They are constructed to be nationally representative of the general population, ensuring that analyses made from these smaller datasets generalize to larger groups. A large literature has used these datasets to estimate and decompose gender wage gaps in the United States population (Blau & Kahn, 2017).

Decomposing the change in explained wage gaps involves estimating the average of wage functions with respect to distributions of characteristics. To estimate these terms, we proceed by fitting a wage model, $\hat{\mu}_g(x, h)$, to longitudinal survey data. This model should approximate the conditional average wage function for the survey population, $\mu_g(x, h)$. With a wage model in tow, each component of the decomposition in Section 2 can be estimated by replacing the true mean wage function μ_g with the wage model $\hat{\mu}_g$, averaging over the empirical distribution of characteristics.

What is the best way to model wages from covariates and career trajectories? In principle, $\mu_g(x, h)$ can be estimated with a regression, using an indicator for each possible work history h. But this approach is statistically untenable: longitudinal survey datasets are small, so there are many more possible work histories than data points available. Thus, classical econometric approaches for decomposing gender wage gaps have summarized histories with coarse-grained summary statistics, such as years of experience and current occupation (Blau & Kahn, 2017). While these approaches allow for efficient estimation, they can discard important aspects about career pathways.

To address this challenge, we develop a machine learning solution. Our method centers on learning a *representation* of the job history $\lambda_\theta(H) : \mathscr{H} \to \mathbb{R}^D$, a parameterized function that summarizes a full job history with a lower-dimensional vector. Then, to model wages as a function

of history, we replace histories with their representations.

A good representation will help sidestep statistical estimation issues by summarizing complex histories with low-dimensional vectors. However, by definition, a low-dimensional representation discards aspects of job history. A representation that omits important aspects of job history will not explain the same amount of the gender wage gap as the full history; this is an example of omitted variable bias (OVB) (Chernozhukov et al., 2022a). We derive a necessary and sufficient condition for low-dimensional representations to not omit important variables, which is met when the representations are predictive of both gender and wage.

Learning a low-dimensional representation of job history from survey data can suffer from the same issue as trying to model wages from high-dimensional histories: the survey datasets are small. To address this problem, we additionally analyze a large-scale, passively collected dataset of millions of job histories to learn an initial representation $\lambda_\theta(H)$. As this dataset does not contain wage or gender information, we fit the initial representation to model individual job sequences. The principle is that by learning important features for predicting job sequences, the representation also acquires features that predict wage and gender.

Our method puts these steps together:

1. Learn an initial representation from a massive databank of job histories.

2. Adjust the representation so it is effective at modeling wages on longitudinal survey data and incorporate it into a wage model, $\hat{\mu}_g(x, \lambda_\theta(h))$.

3. Use the estimated model $\hat{\mu}_g(x, \lambda_\theta(h))$ to decompose the change in the explained gender wage gap into differential transitions and starting characteristics.

Below, we describe our methodology in more detail.

3.1 Modeling job histories with representations

Our methodology for decomposing gender wage gaps centers on finding a low-dimensional representation of job history that makes accurate wage predictions on survey data. We adapt CAREER,

a machine learning technique for learning representations from labor data (Vafa et al., 2022). CAREER was initially developed to predict an individual's future occupations.[4] We modify CAREER to learn representations that are targeted to predict wages rather than future occupations, enabling its use for wage gap estimation.

Modeling wage with CAREER. CAREER parameterizes a representation of an individual's career using transformer neural networks, the same mathematical ideas behind large language models (Vaswani et al., 2017). Transformers were originally developed to represent sequences of words in a body of text; we adapt them to represent sequences of jobs in a career. Formally, recall that a history H is defined as a sequence of jobs and years, $H = ((H_1, D_1), \ldots, (H_L, D_L))$, where each occupation $H_l \in \{1, \ldots, J\}$ at timestep l corresponds to the occupation an individual worked in during year D_l. CAREER uses neural networks to summarize this sequence with a representation, $\lambda_\theta(H) \in \mathbb{R}^D$. See Vafa et al. (2022) for more details about the parameterization of the neural network.

Our wage model combines history and the other covariates assuming additive separability:

$$\mu_g(x, h) = \mu_g(x, \lambda_\theta(h)) = \mu_g^X(x) + \mu_g^H(\lambda_\theta(h)), \tag{23}$$

where $\mu_g^X : \mathbb{R}^P \to \mathbb{R}$ predicts wages from non-history covariates and $\mu_g^H : \mathbb{R}^D \to \mathbb{R}$ predicts wages from representations of history.[5] The history term is modeled as

$$\mu_g^H(\lambda_\theta(h)) = \rho_g(\lambda_\theta(h)), \tag{24}$$

[4] Vafa et al. (2022) include preliminary experiments that demonstrate that these representations, which are fit to model job transitions, can be incorporated into wage regressions. Our methodology here extends this approach in a few ways. First, we fit representations directly to model wage rather than fitting them to model transitions and holding them fixed for wage regressions. Second, the estimation procedure in Section 3.1 differs, both by incorporating covariates and encouraging sufficient representations; these modifications improve predictions. Finally, we modify the objective for leveraging large-scale resume data Section 3.2 so that it is more appropriate for the downstream task of predicting wages rather than job transitions.

[5] We employ this additive separability assumption because it makes the decompositions more interpretable. In Section 6, we will show that even with this assumption, our model makes better predictions of wage than models that do not use histories.

where $\rho_g : \mathbb{R}^D \to \mathbb{R}$ is a two-layer feedforward neural network. Meanwhile, the covariate term $\mu_g^X(x)$ is modeled linearly,

$$\mu_g^X(x) = \alpha_g + \beta_g \cdot x, \tag{25}$$

where $\alpha_g \in \mathbb{R}$ is a gender-specific intercept and $\beta_g \in \mathbb{R}^P$ is a vector of regression coefficients.[6]

Representations can omit variables. This model replaces high-dimensional job histories H with low-dimensional representations $\lambda_\theta(H)$. However, a low-dimensional representation of history will by definition discard components of job history. A representation that omits important variables will not explain the same amount of the wage gap as full history; it will be explaining the wage gap for a smaller set of history variables. This is an example of omitted variable bias (OVB).

Define by $\text{EWG}_t(\lambda_\theta)$ the portion of the gender wage gap explained by covariates and a representation λ_θ:

$$\text{EWG}_t(\lambda_\theta) = \mathbb{E}_{\nu(x^t,h^t|f)}[\mu_m(X^t, \lambda_\theta(H^t))] - \mathbb{E}_{\nu(x^t,h^t|m)}[\mu_m(X^t, \lambda_\theta(H^t))]. \tag{26}$$

We call $\text{EWG}_t(\lambda_\theta)$ the *representation-based explained wage gap*. We would like to learn a representation that does not omit important variables, so that $\text{EWG}_t(\lambda_\theta)$ equals the true explained wage gap EWG_t.

Here we ask, for what representations λ_θ is the representation-based explained wage gap $\text{EWG}_t(\lambda_\theta)$ equal to the true explained wage gap EWG_t?

Proposition 1. *$EWG_t(\lambda_\theta) = EWG_t$ if and only if the following zero-omitted variable bias (ZOVB) condition holds:*

$$\text{Cov}_{\nu(x^t,h^t|m)}\left(\mu_m(X^t, H^t) - \mu_m(X^t, \lambda_\theta(H^t)), \frac{e(X^t, H^t)}{1 - e(X^t, H^t)} - \frac{e(X^t, \lambda_\theta(H^t))}{1 - e(X^t, \lambda_\theta(H^t))}\right) = 0, \tag{27}$$

[6]The decomposition in Section 2 requires averaging the male wage function over the distribution of female characteristics, but it does not require averaging the female wage function over the distribution of male characteristics. Thus, in principle, the female wage function does not need to be estimated. However, since the representation of history is estimated from the population of males and females, we will gain statistical power from finding representations that predict women's wages if the wage functions use similar aspects of history.

for $e(X^t, H^t) = \nu(G = f|X^t, H^t)$, $p = \nu(G = f)$, *and, slightly abusing notation,* $e(X^t, \lambda_\theta(H^t)) = \nu(G = f|X^t, \lambda_\theta(H^t))$ *and* $\mu_m(x^t, \lambda_\theta(h^t)) = \mathbb{E}[Y|X = x^t, \lambda_\theta(H) = \lambda_\theta(h^t), G = m]$. *It then follows immediately that the difference in explained wage gaps is preserved by the representation,* $EWG_1(\lambda_\theta) - EWG_0(\lambda_\theta) = EWG_1 - EWG_0$.

Proof. We start by shifting and rescaling each component of the second term of (27), writing it as:

$$\left(\frac{e(X^t, H^t)}{1 - e(X^t, H^t)}\frac{1-p}{p} - 1\right) - \left(\frac{e(X^t, \lambda_\theta(H^t))}{1 - e(X^t, \lambda_\theta(H^t))}\frac{1-p}{p} - 1\right) \tag{28}$$

Write out the covariance term involving the first term in (28):

$$\begin{aligned}
&\mathrm{Cov}_{\nu(x^t,h^t|m)}\left(\mu_m(X^t, H^t) - \mu_m(X^t, \lambda_\theta(H^t)), \frac{e(X^t, H^t)}{1 - e(X^t, H^t)}\frac{1-p}{p} - 1\right) \\
&= \mathbb{E}_{\nu(x^t,h^t|m)}\left[\left(\mu_m(X^t, H^t) - \mu_m(X^t, \lambda_\theta(H^t))\right)\left(\frac{e(X^t, H^t)}{1 - e(X^t, H^t)}\frac{1-p}{p} - 1\right)\right],
\end{aligned}$$

since $\mathbb{E}_{\nu(x^t,h^t|m)}\left[\frac{e(X^t,H^t)}{1-e(X^t,H^t)}\frac{1-p}{p}\right] = 1$. Then the above can be written as

$$\begin{aligned}
&\int \left(\mu_m(x^t, h^t) - \mu_m(x^t, \lambda_\theta(h^t))\right)\left(\frac{e(x^t, h^t)}{1 - e(x^t, h^t)}\frac{1-p}{p} - 1\right)\nu(x^t, h^t|m)dx^t dh^t \\
&= \int \left(\mu_m(x^t, h^t) - \mu_m(x^t, \lambda_\theta(h^t))\right)\left(\frac{\nu(x^t, h^t, f)}{\nu(x^t, h^t, m)}\frac{1-p}{p}\right)\frac{\nu(x^t, h^t, m)}{1-p}dx^t dh^t \\
&\quad - \int \left(\mu_m(x^t, h^t) - \mu_m(x^t, \lambda_\theta(h^t))\right)\nu(x^t, h^t|m)dx^t dh^t \\
&= \int \left(\mu_m(x^t, h^t) - \mu_m(x^t, \lambda_\theta(h^t))\right)\nu(x^t, h^t|f)dx^t dh^t \\
&\quad - \int \left(\mu_m(x^t, h^t) - \mu_m(x^t, \lambda_\theta(h^t))\right)\nu(x^t, h^t|m)dx^t dh^t \\
&= \mathbb{E}_{\nu(x^t,h^t|f)}\left[\mu_m(X^t, H^t) - \mu_m(X^t, \lambda_\theta(H^t))\right] - \mathbb{E}_{\nu(x^t,h^t|m)}\left[\mu_m(X^t, H^t) - \mu_m(X^t, \lambda_\theta(H^t))\right].
\end{aligned}$$

Now examine the part of the covariance involving the second part of the second term in (27).

$$\begin{aligned}
&\mathrm{Cov}_{\nu(x^t,h^t|m)}\left(\mu_m(X^t,H^t)-\mu_m(X^t,\lambda_\theta(H^t)),\frac{e(X^t,\lambda_\theta(H^t))}{1-e(X^t,\lambda_\theta(H^t))}\frac{1-p}{p}-1\right)\\
&=\mathbb{E}_{\nu(x^t,h^t|m)}\left[\left(\mu_m(X^t,H^t)-\mu_m(X^t,\lambda_\theta(H^t))\right)\left(\frac{e(X^t,\lambda_\theta(H^t))}{1-e(X^t,\lambda_\theta(H^t))}\frac{1-p}{p}-1\right)\right],
\end{aligned}$$

since $\mathbb{E}_{\nu(x^t,h^t|m)}\left[\frac{e(X^t,\lambda_\theta(H^t))}{1-e(X^t,\lambda_\theta(H^t))}\frac{1-p}{p}\right]=1$. Then the above can be written as

$$\begin{aligned}
&\mathbb{E}_{\nu(x^t,\lambda(h^t)|m)}\left[\mathbb{E}_{\nu(h^t|m,x^t,\lambda(h^t))}\left[\left(\mu_m(X^t,H^t)-\mu_m(X^t,\lambda_\theta(H^t))\right)\left(\frac{e(X^t,\lambda_\theta(H^t))}{1-e(X^t,\lambda_\theta(H^t))}\frac{1-p}{p}-1\right)\right]\right]\\
&=\mathbb{E}_{\nu(x^t,\lambda(h^t)|m)}\left[\mathbb{E}_{\nu(h^t|m,x^t,\lambda(h^t))}\left[\mu_m(X^t,H^t)-\mu_m(X^t,\lambda_\theta(H^t))\right]\cdot\left(\frac{e(X^t,\lambda_\theta(H^t))}{1-e(X^t,\lambda_\theta(H^t))}\frac{1-p}{p}-1\right)\right]\\
&=0.
\end{aligned}$$

■

Typically, parameters underlying wage models are inferred by minimizing the predictive error of wage (Blau & Kahn, 2017). However, Proposition 1 shows that predictive accuracy isn't the only important quality when summarizing data with a representation λ_θ. It is possible to learn representations that induce relatively low predictive errors yet invalidate estimates of the explained wage gap.

Between these two goals — predictive accuracy and zero omitted variable bias — lies a tension. Representations of history that lead to accurate predictions of wage will invalidate decompositions when $\lambda_\theta(h)$ omits variables. On the other hand, if $\lambda_\theta(h)$ satisfies the ZOVB criterion but is not predictive of wage, the expected wage function can lead to poor predictions for small datasets; for example, keeping $\lambda_\theta(h)=h$ satisfies ZOVB, but will lead to a poor model of expected wage when histories are high-dimensional.

Training representations without omitted variables. Instead, we cast the problem of finding a representation of history that is both predictive of wage and satisfies ZOVB as a constrained

optimization:

$$\hat{\theta} = \arg\min_{\theta} \mathbb{E}_{\nu(x,h,g,y)}[(Y - \mu_G(X, \lambda_\theta(H)))^2] \tag{29}$$

$$\text{s.t. } G \perp\!\!\!\perp H | \lambda_\theta(H), X, \tag{30}$$

where the form of the representation function λ_θ is fixed. The objective in (29) encourages low-dimensional representations that are predictive of history, while the constraint in (30) enforces the representations to satisfy ZOVB. We refer to the constraint in (30) as **sufficiency**.

For most representations λ_θ, this objective is intractable. We propose an iterative procedure for approximating a solution to this objective. Our procedure is based on projected gradient descent (Calamai & Moré, 1987). At the beginning of the procedure, θ and all other parameters are randomly initialized. At a high level, each iteration of the procedure has two steps. In the first step, θ and the other parameters are updated with gradient descent to minimize the predictive error of wage in (29). In the second step, θ is then projected to the space of sufficient representations, i.e. those satisfying (30). Our procedure involves repeating these two steps; when the expected wage function is non-convex in θ, the solution of the minimization step depends on the output of the projection step.

In the first step, all parameters are updated to minimize the predictive error of wage. There are three kinds of parameters: representation parameters θ, feedforward network parameters ρ_g (24), and non-history covariate regression coefficients $\alpha_g \in \mathbb{R}$ and $\beta_g \in \mathbb{R}^p$. In principle, all parameters can be updated with gradient descent. However, we find that our algorithm converges faster when the regression coefficients α_g and β_g are updated with ordinary least squares. Denote the estimated parameters at iteration τ as $\hat{\theta}^{(\tau)}$, $\hat{\alpha}_g^{(\tau)}$, $\hat{\beta}_g^{(\tau)}$, and $\hat{\rho}_g^{(\tau)}$. Then at each iteration, $\hat{\theta}^{(\tau)}$ and $\hat{\rho}^{(\tau)}$ are first updated with gradient descent, keeping $\hat{\alpha}_g^{(\tau)}$ and $\hat{\beta}_g^{(\tau)}$ fixed:

$$\hat{\theta}^{(\tau+1)} = \hat{\theta}^{(\tau)} - \nabla_{\hat{\theta}^{(\tau)}} \tfrac{1}{N} \textstyle\sum_i (y_i - \hat{\mu}_{g_i}(x_i, \lambda_{\hat{\theta}^{(\tau)}}(h_i); \hat{\alpha}_g^{(\tau)}, \hat{\beta}_g^{(\tau)}, \hat{\rho}_g^{(\tau)})) \tag{31}$$

$$\hat{\rho}_g^{(\tau+1)} = \hat{\rho}_g^{(\tau)} - \nabla_{\hat{\rho}_g^{(\tau)}} \tfrac{1}{N} \textstyle\sum_i (y_i - \hat{\mu}_{g_i}(x_i, \lambda_{\hat{\theta}^{(\tau)}}(h_i); \hat{\alpha}_g^{(\tau)}, \hat{\beta}_g^{(\tau)}, \hat{\rho}_g^{(\tau)})), \tag{32}$$

where we have re-written the conditional mean function $\hat{\mu}_{g_i}(x_i, \lambda_{\hat{\theta}^{(\tau)}}(h_i); \hat{\alpha}_g^{(\tau)}, \hat{\beta}_g^{(\tau)}, \hat{\rho}_g^{(\tau)}))$ to make its dependence on parameters explicit. Then, the regression coefficients α_g and β_g are updated with ordinary least squares:

$$\hat{\alpha}_g^{(\tau+1)}, \hat{\beta}_g^{(\tau+1)} = \underset{\alpha_g, \beta_g}{\arg\min} \frac{1}{N} \sum_i (y_i - \hat{\mu}_{g_i}(x_i, \lambda_{\hat{\theta}^{(\tau+1)}}(h_i); \alpha_g, \beta_g, \hat{\rho}_g^{(\tau+1)}))^2. \tag{33}$$

We use a held-out sample of the training data for validation, and stop updating parameters when the validation loss begins to worsen. This technique is known as early stopping, and it is a common method for regularizing transformers to prevent them from overfitting (Dodge et al., 2020).

It is challenging to characterize the space of sufficient representations. We develop a heuristic to encourage sufficient representations. Our heuristic is based on the fact that sufficient representations will carry all aspects of history that are predictive of gender. Thus, a sufficient representation should make accurate predictions of an individual's gender from their characteristics. To encourage sufficiency from a representation $\lambda_\theta(H)$, we update the parameters θ to improve gender predictions. Introducing logistic regression coefficients $\gamma \in \mathbb{R}^D, \eta \in \mathbb{R}^P$, we update the parameters θ, γ, and η with gradient descent:

$$\begin{aligned}\hat{\theta}^{(\tau+1)} = \hat{\theta}^{(\tau)} + \nabla_{\hat{\theta}^{(\tau)}} \frac{1}{N} \sum_i \big(& 1(g_i = f) * \log(\hat{\pi}(x_i, h_i; \hat{\theta}^{(\tau)}, \hat{\gamma}^{(\tau)}, \hat{\eta}^{(\tau)})) \\ & + 1(g_i = m) * \log(1 - \hat{\pi}(x_i, h_i; \hat{\theta}^{(\tau)}, \hat{\gamma}^{(\tau)}, \hat{\eta}^{(\tau)}))\big),\end{aligned}$$

where $\hat{\pi}(x_i, h_i; \theta, \gamma, \eta) = \sigma\left(\frac{1}{1+\exp(-(\gamma \cdot \lambda_\theta(h_i) + \eta \cdot x_i))}\right)$ for the logistic function σ. The analogous updates are performed for $\hat{\gamma}^{(\tau+1)}$ and $\hat{\eta}^{(\tau+1)}$. Since this objective is non-convex in the representation parameters, the goal is to encourage a final representation that is sufficient for gender while retaining the aspects from history that are important for wage.

The full algorithm and training details are presented in Appendix A. Appendix B includes semi-synthetic experiments that demonstrate the effectiveness of this approach. Appendix C shows that projecting does not diminish the predictive performance of the wage model; in fact, it often improves it.

We estimate the model parameters above using all available observations for full-time workers

from longitudinal survey data.[7] When estimating parameters, we use cross-fitting to ensure that model predictions are not made using the same samples used for training (Chernozhukov et al., 2018). We randomly divide individuals into five folds. When an individual has multiple observations over different years, all of these observations are included in the same fold. We then train five different models on longitudinal survey data, each one holding out a different fold. All the predictions used in the analysis of this paper use a model trained without the fold we are predicting on.

3.2 Leveraging large-scale data to learn representations

We have described a method that fits a wage model to survey data. The longitudinal survey datasets used for estimating wage gaps in the United States are small, containing only thousands of individuals. However, the transformer neural networks that underpin our proposed wage model are unlikely to learn useful representations of job histories from these small datasets (Kaplan et al., 2020). Here, we present an approach that learns effective representations via *transfer learning*, a technique from machine learning that trains predictive models by augmenting smaller datasets with larger, related datasets.

Specifically, we leverage large-scale, passively-collected resume data alongside the smaller longitudinal data to learn representations of job history. Our approach is inspired by the method developed in Vafa et al. (2022), which learned representations from resume data to help predict job transitions on smaller survey data. However, in contrast to Vafa et al. (2022), our goal is to predict a quantity that does not appear in resume data: an individual's wage. Thus, we cannot use wage predictions from resume data to form the basis for wage predictions on survey data.

Instead, we develop a strategy that leverages resume data to learn important features without directly predicting wage. While resume datasets do not contain wages, they can contain millions of samples of job sequences. These sequences encode information about the relationships between

[7] Although our decompositions will focus on smaller subpopulations (e.g. individuals younger than 35 in a certain year), we find that models fit to only the subpopulations of interest form worse predictions than models fit to the full sample. In Section 3.3, we describe an additional training step that adjusts the fitted model so its predictions are targeted to the subpopulations in each decomposition.

jobs. If there are features that are relevant both to whether jobs occur in a sequence together and wage, they can be gleaned from resume data. Our strategy is to first fit a representation that is useful for predicting which jobs co-occur in careers; the representation is then adjusted to predict wages on survey data.

Specifically, we begin by initializing the representation function λ_θ randomly. We then train this representation to predict the occupation trajectories described in resumes. Our training objective encourages the representation to encode features that help predict the other jobs in a career; this process is known as "pretraining". Our objective modifies the one described in Vafa et al. (2022), and it is described in more detail in Appendix A. Then, to fit a wage model and sufficient representation to survey data, we initialize the projection procedure described above with the "pretrained" representation.

3.3 Estimating decomposition terms

Given a fitted wage model $\hat{\mu}_g(x, h)$, the decompositions in Section 2 can be estimated by replacing the true wage function with model predictions. First, a subpopulation must be defined for the decomposition (e.g. all working individuals aged 25-35 working between 1990-2007). Then, averages over distributions of characteristics can be taken over empirical samples of the subpopulation. For example, the term $\mathbb{E}_{\nu(x^0,h^0|f)}[\mu_m(x^0, h^0)]$ can be estimated with

$$\frac{1}{|C_f|}\sum_{i\in C_f}\hat{\mu}_m(x_i^0, \lambda_\theta(h_i^0)), \tag{34}$$

where C_f is the index set of female samples in the subpopulation of interest.

Because our model assumes linear separability in the covariates and history, we can decompose the change in each variable separately. That is, the effect of changing all characteristics can be broken up into the effects of changing covariates and the effects of changing histories. Then, we can decompose the change explained by covariates into effects of differential starting covariates and differential covariate transitions, and separately do the same for history.

When decomposing the effect explained by changing covariates, the remaining terms that need

to be estimated are conditional expectations over possible transitions, which have the form:

$$\mathbb{E}_{\nu(x^1|x^0,g)}\left[\mu_m(X^1)|X^0\right]. \tag{35}$$

This is a function from period-0 covariates to period-1 predicted wages from covariates. Rather than modeling the transition distributions, we estimate (35) using empirical samples and estimates of the wage model. Specifically, we model

$$\hat{\mu}_m(x_i^1) = \psi_{g0} + \psi_{g1} \cdot x_i^0 + \epsilon_i, \qquad \forall i \in C_g \tag{36}$$

with regression coefficients $\psi_{g0} \in \mathbb{R}, \psi_{g1} \in \mathbb{R}^p$. This regresses the predicted male wage from period-1 covariates on the period-0 covariates for individuals with gender g. We fit ψ_{g0} and ψ_{g1} using ordinary least squares. The full conditional expectation terms are then estimated using the fitted coefficients $\hat{\psi}_{g0}$ and $\hat{\psi}_{g1}$. For example, (16) is estimated with

$$\frac{1}{|C_f|}\sum_{i\in C_f}\left(\hat{\psi}_{m0} + \hat{\psi}_{m1} \cdot x_i^0\right), \tag{37}$$

where C_f is the index set of female observations for the decomposition subpopulation.[8]

To model the conditional expectations of histories, we perform the analogous regression, albeit using the representations of history $\hat{\lambda}_\theta(h_i^0)$ as the predictors.

Debiasing decompositions. When decomposing the gender wage gap, it is important for a wage model's average predicted wage for each gender to equal the observed average wage. That is, the wage predictions should be unbiased for the decomposition subpopulation:

$$\frac{1}{C_f}\sum_{i\in C_f}\hat{\mu}_f(x_i^t, h_i^t) = \frac{1}{C_f}\sum_{i\in C_f} y_i^t \tag{38}$$

$$\frac{1}{C_m}\sum_{i\in C_m}\hat{\mu}_m(x_i^t, h_i^t) = \frac{1}{C_m}\sum_{i\in C_m} y_i^t, \tag{39}$$

[8] We can further decompose effects explained by changes in all covariates into effects explained by changes in each covariate separately, due to the linear separability assumption. When decomposing the change of a single covariate p, we use the predicted value from that covariate $\hat{\mu}_m(x_{ip}^1)$ on the left hand side of (36) and use all covariates x_i^0 on the right hand side.

where C_f and C_m are the index set of female and male samples in the decomposition subpopulation, respectively.

Classical applications of decompositions like the Blinder-Oaxaca decomposition (Blinder, 1973; Oaxaca, 1973) have estimated wage models by fitting linear regressions on the same populations used for decompositions. Since linear regression fit with ordinary least squares is unbiased on the data it is fit to, the models have been unbiased on the decomposition population. However, in our case, we are not using ordinary least squares to fit wage models, which may allow bias. Moreover, as described in Section 3.1, we are not fitting our model to only the subpopulation of interest. Rather, we are fitting our model to the complete survey sample and applying it to the subpopulation of interest in decompositions. This is a *distribution shift* (Koh et al., 2021); even a model that is unbiased for the general population may be biased on another subpopulation.

We develop a procedure to adjust wage models so that they are unbiased for a given decomposition subpopulation. Given a wage model $\hat{\mu}_g(x, h)$, we adjust the model separately for each period. Call the adjusted model for the period t subpopulation $\hat{\mu}_g^t(x, h)$. The adjusted model adds two terms to the original:

$$\hat{\mu}_g^t(x, h) = \delta_g^t * \hat{\mu}_g(x, h) + \psi_g^t, \tag{40}$$

where $\psi_g^t, \delta_g^t \in \mathbb{R}$ are gender-specific scale and location terms. We estimate both parameters using ordinary least squares:

$$\hat{\delta}_g^t, \hat{\psi}_g^t = \underset{\delta_g^t, \psi_g^t}{\arg\min} \frac{1}{C_g} \sum_{i \in C_g} (y_i - \hat{\mu}_{g_i}^t(x, h))^2. \tag{41}$$

guaranteeing the model will be unbiased for the subpopulation it is fit to. We use cross-fitting to fit these parameters, dividing the fold that $\mu_g(x, h)$ is not originally trained on into five subfolds, alternatively holding one out at a time. Although the average residual will not be exactly zero on the held-out split, this procedure encourages unbiasedness in expectation.

Because the adjusted models may differ for periods 0 and 1, we adjust the decomposition in Section 2 to account for model differences. We modify the difference in explained wage gaps term

((10) and (11)) so that the same model is used for each period:

$$\hat{\text{EWG}}_1 - \hat{\text{EWG}}_0 = \mathbb{E}_{\nu(x^1,h^1|f)}\left[\hat{\mu}_m^1(X^1,H^1)\right] - \mathbb{E}_{\nu(x^1,h^1|m)}\left[\hat{\mu}_m^1(X^1,H^1)\right] \tag{42}$$

$$- \left(\mathbb{E}_{\nu(x^0,h^0|f)}\left[\hat{\mu}_m^1(X^0,H^0)\right] - \mathbb{E}_{\nu(x^0,h^0|m)}\left[\hat{\mu}_m^1(X^0,H^0)\right]\right). \tag{43}$$

This keeps the interpretation of the explained difference component as the change due to changing characteristics over time with a fixed model. The further decompositions of this term into differential transitions and differential starting characteristics are then performed analogously, using the model $\hat{\mu}^1$ to estimate all terms.

We must then add a term to account for the models changing over time,

$$\hat{\text{M}}_1 - \hat{\text{M}}_0 = \mathbb{E}_{\nu(x^0,h^0|f)}\left[\hat{\mu}_m^1(X^0,H^0)\right] - \mathbb{E}_{\nu(x^0,h^0|m)}\left[\hat{\mu}_m^1(X^0,H^0)\right] \tag{44}$$

$$- \left(\mathbb{E}_{\nu(x^0,h^0|f)}\left[\hat{\mu}_m^0(X^0,H^0)\right] - \mathbb{E}_{\nu(x^0,h^0|m)}\left[\hat{\mu}_m^0(X^0,H^0)\right]\right) \tag{45}$$

When the period-0 and period-1 models are the same, this term is zero. If the above term is small, it means that the general wage model $\hat{\mu}_g(x,h)$ is effective at modeling the decomposition subpopulations at periods 0 and 1. When the change is large, it means that the adjustments have a large effect.

Finally, the estimated components due to changing unexplained wage gaps are analogous to their definition in Section 2, using $\hat{\mu}_g^t$ as the wage model at period t:

$$\hat{\text{UWG}}_1 - \hat{\text{UWG}}_0 = \mathbb{E}_{\nu(x^1,h^1|f)}\left[\hat{\mu}_f^1(X^1,H^1)\right] - \mathbb{E}_{\nu(x^1,h^1|f)}\left[\hat{\mu}_m^1(X^1,H^1)\right] \tag{46}$$

$$- \left(\mathbb{E}_{\nu(x^0,h^0|f)}\left[\hat{\mu}_f^0(X^0,H^0)\right] - \mathbb{E}_{\nu(x^0,h^0|f)}\left[\hat{\mu}_m^0(X^0,H^0)\right]\right) \tag{47}$$

Putting this all together, the change in gaps is decomposed into:

$$\text{WG}_1 - \text{WG}_0 = \hat{\text{EWG}}_1 - \hat{\text{EWG}}_0 + \hat{\text{UWG}}_1 - \hat{\text{UWG}}_0 + \hat{\text{M}}_1 - \hat{\text{M}}_0. \tag{48}$$

This decomposition has the same form as the decomposition proposed by Juhn et al. (1993), where

the model change over time is not due to the underlying wage function changing but rather to adjusting the estimated model to each period's subpopulation.

The decomposition above relates the difference in expected wage gaps to other expectations. In practice, when the expectations are evaluated on empirical samples, the left-hand side of (48) will not exactly equal the right-hand side. This will be the case even for unbiased estimation procedures due to cross-fitting. We refer to this term as the error due to cross-fitting. Table 14 in Appendix E performs our main decomposition without performing the debiasing step. The high-level results are the same as with the debiasing step, although the cross-fitting error terms are larger.

4 Data

In the United States, gender wage gaps are most commonly estimated using longitudinal survey data (Blau & Kahn, 2017). Here, we describe the survey dataset we use for wage gap estimation. We also discuss how we prepare a corpus of passively-collected worker profiles to aid our estimate of the wage gap.

Panel Study of Income Dynamics. We estimate gender wage gaps using the Panel Study of Income Dynamics (Panel Study of Income Dynamics, public use dataset, 2023), or PSID. PSID is a longitudinal survey that follows a cohort of American families.[9] Interviews were conducted annually from 1968 until 1997, and they have been conducted biannually since. Since the survey has been conducted biannually since 1997, we follow Blau & Kahn (2017) and impute experience information for the skipped years using retroactive questions. Because the same individuals are interviewed over the course of the survey, job histories can be constructed by tracking the trajectory of reported occupations each year an individual is in the survey. Jobs are initially encoded using census codes that vary over the course of the survey. We standardize these occupational categories, using a crosswalk to transform jobs into one of 330 `occ1990dd` occupational categories (Autor &

[9]The PSID survey reports an individual's sex, but it does not report their gender. In this study, we follow the convention of the existing literature and use the term "gender wage gap" despite relying on sex-based data from the PSID. The PSID variable for an individual's sex does not change over time, so this variable is fixed for our analyses that follow individuals over time.

Dorn, 2013). We add seven categories for when an individual's occupation is unavailable but their employment status is available. These categories are: employed, temporarily laid off, unemployed, disabled, retired, homemaker, and student. In all of our analyses, these categories are treated as special kinds of occupations.

The PSID only began asking females for their employment status in 1979, so we do not include any occupational information for either males or females from before 1979. Since our goal is to understand how occupational trajectories contribute to the gender wage gap, it is important that our sample contain the most detailed trajectories possible. Thus, we restrict our sample to the surveys conducted between 1990-2019 (still keeping occupational information from 1979-1989 when it is available). We further restrict our sample to non-farm and non-military wage and salary workers between 25 and 64 years old who worked for at least 26 weeks in non-farm jobs, following Blau & Kahn (2017). In total, we are left with 91,391 observations over the 19 surveys conducted between 1990 and 2019. The surveys ask about occupational experience for each previous year, so we label results as being for the year before each survey is conducted.

Individuals are not always added to the PSID at the beginning of their careers. Additionally, since the PSID has only conducted interviews biannually since 1997, the majority of occupational trajectories in the PSID do not have comprehensive, year-by-year observations for the full duration of a worker's career. The maximum number of observations that could be available is 30, one for each year of the survey between 1979 and 2019. For both males and females, the median number of occupational observations available is 13.

The PSID provides longitudinal sample weights for each family. These weights are designed to adjust for differences in the probability of selection into the sample. We incorporate these weights into all our analysis.

As discussed in Section 2, each term of the decomposition can only be estimated in a population where there is overlap, i.e. $\nu(G = f|h, x) < 1$. In order to ensure overlap, we trim the population so that only samples are included where there is sufficient overlap (Crump et al., 2006). Specifically, we use CAREER's propensity model to evaluate male and female propensities, discarding examples where male or female propensities are above 99%. All results reported in this paper are for the

trimmed population. This changes the population for whom we are decomposing wage gaps: we are now decomposing the change in gender wage gaps for the individuals whose *whose work histories and covariates are in the middle 98% of the gender distribution*.

After trimming, the total number of observations decreases from 91,391 to 90,074. Table 15 in Appendix E performs our main analysis without trimming, and the top-level results are similar.

Passively-collected job sequences. Our method for incorporating career trajectories into wage gap estimation leverages large-scale data to help learn representations of work trajectories. Since the PSID survey follows a relatively small number of individuals, we augment the PSID data with a massive dataset of work trajectories. We construct this dataset by transforming a large dataset of resumes and worker profiles collected by Zippia Inc., a career planning company, into sequences of work trajectories. Specifically, Zippia provided us with worker profiles, containing sequences of occupations of American workers. Each occupation in a worker profile had been encoded as an O*NET SOC occupational code, imputed from textual descriptions. We transformed these codes into `occ1990dd` codes to match those in our PSID sample. In total, this dataset contained 23.7 million individuals and 245 million occupational observations.

While this dataset of passively-collected career histories contains rich information about work trajectories, it cannot be used directly for wage gap estimation for a few reasons. Most crucially, necessary information for estimating gender wage gaps, such as genders and wages, are not available from passively-collected worker profiles. Additionally, noise and errors in histories will arise from imputing occupations from short textual descriptions along with individuals inaccurately listing their work experiences on worker profiles (Wexler, 2006). Finally, the worker profiles are not collected to be representative of the general population. Still, we will show that leveraging these worker profiles improves the predictions of wages on survey data.

5 Models

We consider various wage models for $\mu_g(x, h)$. We follow Blau & Kahn (2017) in including the following covariates X: years of full-time and part-time experience (and their squares), years of schooling, indicators for bachelors and advanced degrees, race and ethnicity indicators, census region indicators, an indicator for collective bargaining coverage, 15 industry category indicators, and 21 occupation category indicators.[10] Since our sample includes observations from multiple years, we also include year indicators and year-covariate interactions for each covariate.

The covariates X do not encode in an individual's complete work trajectory; only the number of years of full- and part-time experience. The PSID collects detailed labor market experience beyond these coarse summary statistics, recording an individual's occupation each year they are in the survey. However, as discussed in Section 3, this occupational information is too high-dimensional for classical econometric approaches to incorporate for estimating the conditional wage function μ.

We consider two broad approaches for modeling wage as a function of observables: models that do not condition on full history, and models that do condition on history. For the models that do not condition on full history, we apply traditional econometric methods that incorporate linear assumptions. For the models that condition on history, we transform histories into lower-dimensional representations using the technique described in Section 3. Below, we detail these models in more detail.

Linear models, no history. For modeling wage without using detailed history, we employ the wage model used by Blau & Kahn (2017) for estimating wage gaps. Given the covariates X described above, the conditional wage mean is modeled as:

$$\mu_g(x, h) = \alpha_g + \beta_g \cdot x, \tag{49}$$

[10] Unlike Blau & Kahn (2017), we do not include metro area indicators since they are only included in the restricted PSID sample, which we do not have access to. As Blau & Kahn (2017) find that region variables explain very little of the gender wage gap (0.4% in 1980 and 0.1% in 2010), we do not expect this exclusion to affect our analysis.

where $\alpha_g \in \mathbb{R}$ and $\beta_g \in \mathbb{R}^P$ are gender-specific intercepts and regression coefficients, respectively. This model does not consider fine-grained occupations or detailed work histories. Instead, occupation is summarized using the 21 coarse-grained categories in X, and work experience is summarized by the number of part-time and full-time years worked (also in X). We refer to this model as using **coarse-grained occupations** since it does not use detailed occupational information — only the 21 high-level occupational categories in X.

In contrast to the 21 occupational categories present in X, the PSID encodes occupations into a fine-grained taxonomy containing more than 300 occupations. While Blau & Kahn (2017) use coarse-grained occupations to estimate the gender wage gap on single-year slices of the PSID, we use a larger sample consisting of observations over a 30-year span. This larger data size enables efficient estimation from the more detailed occupational titles. Thus, we also consider using the detailed occupational labels as part of X, otherwise following (49). We refer to this model as using **fine-grained occupations**.

We consider two approaches for fitting the regression coefficients α_g and β_g: ordinary least squares and LASSO. For LASSO, we use cross-validation to set the regularization parameter.

Nonlinear models, with history (CAREER). The main approach we consider models wage as a function of fine-grained occupations and detailed history data. We refer to this model as **CAREER**, and it is described in more detail in Section 3. We use an ensemble of 10 transformers with $D = 64$ dimensions for the representation, 4 encoder layers, 2 attention heads, and 256 hidden units for the feedforward neural networks, for a total of 2,519,060 parameters. This is considerably smaller than the transformers used to train large language models, which have billions of parameters (Vaswani et al., 2017; Devlin et al., 2019; Ouyang et al., 2022). We find that smaller transformer models are effective at modeling sequence of occupations, which are generally shorter and less complex than the sequences of words large language models are applied to.

This model differs from the linear models in a few ways. Most notably, it conditions on an individual's full career history, h_i, instead of relying solely on the summary statistics in x_i. Additionally, its representation $\lambda_\theta(h_i)$ summarizes not just career history but also an individual's

current occupation. That is, while the linear model in (49) encodes current job as an indicator variable, CAREER also encodes it in $\lambda_\theta(h_l)$. Thus if the predictions of the full-history model are superior to those of the linear model in (49), this improvement cannot be solely attributed to its inclusion of history. Rather, the improvement may be due to incorporating current occupation in an improved functional form.

Thus, differences between the linear models that do not use history and CAREER cannot be solely attributed to history. In order to understand how including history affects predictions, we train a version of the CAREER model in Section 3 that only includes an individual's current job in the representation $\lambda_\theta(h)$. More specifically, we use the following term to predict wage from history:

$$\mu_h^H(h) = \rho_g(\lambda_\theta([h_L, d_L])), \tag{50}$$

where h_L denotes an individual's current job, d_L denotes the year, and $\rho_g : \mathbb{R}^D \to \mathbb{R}$ is a two-layer feedforward neural network. In other words, history is truncated to only include the current job and year. All other model and estimation details are as described in Section 3. We refer to this model as **CAREER (current job only)**.

We train one more variation of CAREER that conditions on an individual's historic workforce participation rather than their full history. An individual's work history contains their occupation H_l for each year they are working. When they are not working, their occupation is encoded using one of six special labels: temporarily laid off, unemployed, disabled, retired, homemaker, and student. In order to assess the degree to which CAREER's representation is capturing workforce participation status rather than actual occupation trajectories, we create modified histories $\tilde{H} = ((\tilde{H}_1, D_1), \ldots, (\tilde{H}_L, D_L))$ where each previous occupation $\tilde{H}_l$ only encodes an individual's workforce participation status:

$$\tilde{H}_l = \begin{cases} \text{“employed”} & \text{if the individual is working at time } l \\ H_l & \text{otherwise,} \end{cases} \tag{51}$$

for $l < L$. For an individual's current occupation, we include the full occupation $\tilde{H}_l = H_l$. We

train a version of CAREER using the modified histories $\tilde{H}$ instead of the full histories. All other model and estimation details are as described in Section 3. We refer to this model as **CAREER (participation and current job only)**.

All models are trained on the full pooled sample, containing 91,391 individuals. For the decompositions, the models are then adjusted for the decomposition subpopulation following the debiasing procedure in Section 3.3. During training, each individual is weighted according to their family weight. For all analyses that pool across years, weights are normalized so that the summed weight for each year is constant. For example, when estimating the MSE between 1989 and 1994, weights are normalized so they sum to 1 for each year.

In our analyses, we estimate standard errors by bootstrapping. The bootstrap we employ resamples individuals rather than observations. If an individual is resampled, their observations for each year are included in the dataset, while if they are not resampled, all of their observations are dropped. This resampling strategy preserves the correlation structure between repeated observations from the same individual. We do not re-train models for each bootstrap sample. Instead, we keep models fixed, evaluating each model on the bootstrapped sample. Since the decompositions require adjusting models to be unbiased for decomposition subpopulations, we re-adjust the models for each bootstrap so that the standard error estimates incorporate the variance due to this adjustment. For all analyses, standard errors are estimated using 100 bootstraps.

6 Predictive Performance

We begin by evaluating each model's ability to predict wages on held-out data. We demonstrate that the model that uses history, CAREER, makes the best predictions, both in terms of mean-squared error and calibration. In order to understand the source of CAREER's predictive improvement, we analyze the learned representations $\lambda_\theta(H)$. We create clusters of histories that the model views as similar. We demonstrate that coarse-grained occupation categories omit elements of history that are important for wages; for example, these clusters group together individuals who are listed as general managers but whose previous jobs include engineering roles. Finally, we introduce a

heuristic to reclassify workers into fine-grained occupations based on their histories. We show that models trained on the reclassified dataset are more accurate for predicting wages.

6.1 Mean-square error

We compare the predictive performance of six wage models:

- **Coarse-grained regression:** The linear model in (49), which uses 21 coarse-grained occupational categories. All covariates are interacted with year. Parameters are fit with ordinary least squares.
- **Coarse-grained LASSO:** The linear model in (49), which uses 21 coarse-grained occupational categories. All covariates are interacted with year. Parameters are fit with LASSO, using cross-validation for the regularization parameter.
- **Fine-grained LASSO** The linear model in (49), although 330 fine-grained occupational categories are used in X in addition to the 21 coarse-grained occuptional categories for the above models. All covariates are also interacted with year. Parameters are fit with LASSO, using cross-validation for the regularization parameter.
- **CAREER (current job only):** The CAREER model with one exception: only an individual's most recent job is included in the representation (50). This model is helpful for distinguishing how much of CAREER's predictive improvement is due to a more effective functional form of an individual's current job as opposed to including history.
- **CAREER (participation and current job only):** Another variant of the CAREER model that captures the effect of historic workforce participation. An individual's most recent job is included in CAREER's representation. All previous positions are encoded using only the individual's employment status (e.g. unemployed, student) rather than the actual occupation. See (51) for more details.
- **CAREER (no resumes):** The full CAREER model except a large-scale dataset of resumes is not used to aid its representations. Instead, the model is only fit to PSID survey responses.

	MSE (full)	MSE (male)	MSE (female)	R^2 (full)
Coarse-grained regression	0.215 (0.004)	0.195 (0.006)	0.232 (0.004)	0.417 (0.010)
Coarse-grained LASSO	0.210 (0.004)	0.190 (0.006)	0.228 (0.004)	0.430 (0.010)
Fine-grained LASSO	0.201 (0.004)	0.179 (0.006)	0.220 (0.004)	0.456 (0.010)
CAREER (current job only)	0.200 (0.004)	0.178 (0.006)	0.219 (0.004)	0.458 (0.010)
CAREER (participation and current job only)	0.193 (0.003)	0.171 (0.005)	0.213 (0.003)	0.475 (0.009)
CAREER (no resumes)	0.187 (0.003)	0.164 (0.005)	0.208 (0.004)	0.491 (0.009)
CAREER (with resumes)	0.174 (0.003)	0.153 (0.005)	0.193 (0.003)	0.527 (0.009)

Table 1: Log-wage predictive performance on PSID, pooled across years (1989-2018). The total sample consists of 90,074 individuals. All models are fit by cross-fitting, dividing the data into five folds. All results are reported for held-out data. Estimated standard errors are in parentheses.

- **CAREER (with resumes):** The full CAREER model in Section 3, which learns a representation of full job history and leverages a dataset of 23.7 million resumes to aid its representation learning.

Table 4 shows the mean-squared error (MSE) of the six wage models. All results are for held-out predictions. The full CAREER model, which incorporates history into wage predictions, exhibits the best performance. It shows the lowest MSE values (0.174 for the full dataset, 0.153 for males, and 0.193 for females) and the highest held-out R^2 value (0.527 for the full dataset). Its predictive performance is not stemming from including a better functional form for an individual's current job; an identical model that only uses an individual's current job performs on par with a linear model that uses fine-grained occupations and no history. However, using the large-scale dataset of resumes is important for its predictive advantage. A version of CAREER that does not use these resumes performs as close to the LASSO as it does to the full model that uses resumes.

In Table 10 in Appendix C, we break out the predictive performance of each wage model into various year buckets. The same trends as Table 4 hold within each pool bucket, with the CAREER model performing best for each year. Its comparative advantage is smallest for the 1989-1994 bucket, which has the shortest history lengths among all buckets.

Of the wage models that do not use CAREER, the best-performing model is consistently the model that includes fine-grained occupational categories and is fit with LASSO. This model uses an individual's fine-grained occupation in addition to the 21 coarse-grained categories to predict wage. The effectiveness of this model indicates the importance of detailed occupational categories,

even in the absence of job history. For the remainder of the analysis, we will focus on comparing the full CAREER model to this linear model that does not use full history.

While our analysis focuses on the pooled PSID sample, our wage model is also effective when trained on only single-year slices of the PSID. Table 11 in Appendix C shows the predictive performance of each model when it is trained on only a single year of the PSID survey. These datasets have much fewer observations than the pooled sample, so each model performs worse than on the pooled sample. Still, our model performs the best for each year. CAREER (with resumes) has a larger advantage over CAREER (no resumes) for the single-year surveys than it does for the pooled sample. When data is small, it is even more important to augment small survey datasets with large-scale auxiliary data to estimate the representation function.

Figure 6 in Appendix C compares the predictive performance of CAREER wage models trained with and without enforcing sufficient representations. We find that our sufficiency constraint does not hurt the wage predictions of the model; in fact, sufficiency-constrained optimization improves the predictions of wages.

6.2 Calibration

While mean-squared error (MSE) gauges a model's predictive precision, it doesn't reflect the reliability of predictions across the distribution of inputs. A model can have low MSE yet consistently under-estimate wage for some groups of characteristics and over-estimate wage for others. Calibration, on the other hand, specifically evaluates the reliability of model predictions across groups of characteristics. A well-calibrated model has predicted averages that mirror observed averages closely for various pre-defined groups.

Here, we assess and compare the calibrations of the best-performing econometric wage model, the fine-grained occupation LASSO model, and our proposed wage model, CAREER. We use the following procedure to assess calibration:

1. Define a bucketing function $\Pi : (\mathbb{R}^P, \mathscr{H}) \to [K]$ that maps observed characteristics to K discrete buckets. For each bucket k, define C_k as the index set of observations assigned to bucket k, $C_k = \{i : \Pi(x_i, h_i) = k\}$.

2. Compute the true average log-wage for each bucket $\bar{y}_k = \frac{1}{|C_k|} \sum_{i \in C_k} y_i$.

3. Compute the predicted average log-wage for each bucket, $\hat{y}_k = \frac{1}{|C_k|} \sum_{i \in C_k} \hat{\mu}_{g_i}(x_i, h_i)$, using held-out data.

If a wage model $\hat{\mu}$ is well-calibrated, the mean predicted log-wage for each category $\hat{y}_k$ should be close to the true average log-wage for the category $\bar{y}_k$. Systemic differences between $\hat{y}_k$ and $\bar{y}_k$ reveal model unreliability. We quantify a model's calibration by estimating its average absolute calibration error:

$$\sum_k \frac{|C_k|}{N} |\bar{y}_k - \hat{y}_k| , \tag{52}$$

where N is the total number of observations. The average absolute calibration error penalizes group predictions that stray from the true group average wage. The calibration error is evaluated using predictions on held-out data.

In this exercise, a model's calibration depends on the bucketing function Π. An ideal bucketing function partitions observables into regions that are homogeneous in wage, while ensuring that the true wage distribution differs across regions. If the buckets are too broad, the wage distribution within each bucket could be heterogeneous, which can mask model deficiencies. For example, if a bucketing function partitions observations uniformly at random, any model will be well-calibrated for the partition as long its marginal prediction is close to the true population average.

We consider bucketing functions based on the quantiles of wage model predictions. This bucketing strategy assesses how a model performs across the full distribution of wage predictions. If its predictions are well-calibrated, the true wages for the lowest-quantile bucket should be smallest. The true average wage should then increase for each quantile.

We can create quantiles based on the wage predictions of either the linear LASSO wage model that does not use history or the wage model that uses CAREER to represent full job histories. Define $\hat{\mu}_{g,L}$ to be the log-wage predictions from the LASSO model and define $\hat{\mu}_{g,C}$ to be the predictions from the CAREER model. Denote by Π_L the bucketing function that partitions data based on centiles of the LASSO model predictions. That is, an observation (g_i, x_i, h_i) is assigned to bucket $q \in \{1, \ldots, 100\}$ if $\hat{\mu}_{g_i,L}(x_i, h_i)$ falls in the range defined by the $(q-1)$'th and q'th percentile of all

$\hat{\mu}_{g,L}$ predictions. Denote by Π_C the analogous bucketing function using centiles from the CAREER model predictions $\hat{\mu}_{g_i,C}$.

Figure 1 depicts the calibration for clusters created from the LASSO predicted centiles, Π_L. Both the LASSO predictions and CAREER predictions are reasonably well-calibrated, both overall and for each gender individually. The average absolute calibration error is 0.019 for the LASSO model, compared to 0.016 for CAREER.

However, the results change when CAREER's predicted centiles are used to determine buckets instead of LASSO's. Figure 2 shows the calibration for clusters created from CAREER's predicted centiles. While CAREER is well-calibrated, the LASSO model's calibration deteriorates. The average absolute calibration error is 0.055 for the LASSO model, compared to 0.023 for CAREER. The LASSO model consistently under-estimates wage for observations that the CAREER model predicts to have low-wage, while it consistently over-estimates wage for observations the CAREER model predicts to have high-wage. On the other hand, CAREER is calibrated over its own predictions, for both males and females.

Why is there a shift in calibration performance? LASSO's predictions do not incorporate job history, so buckets created from these predictions may not group together job histories with similar values. Instead, histories with both low and high value may be regularly grouped together, thus averaging away effects of history. On the other hand, when the CAREER model's predicted centiles are used to define the buckets, each bucket contains observations with more similarly valued job histories. Since histories are more homogeneous within buckets, the LASSO's predictions will not average out to equal the true average wages.

6.3 Interpreting model improvement

Here, we investigate the representations of history captured by our model to understand why its predictions of wage are better than those of models that do not use history.

We first demonstrate that occupation labels omit important elements of history that may refine an individual's current occupation. For example, managers who were previously engineers make higher wages than those that were never engineers. We show that these factors are captured by

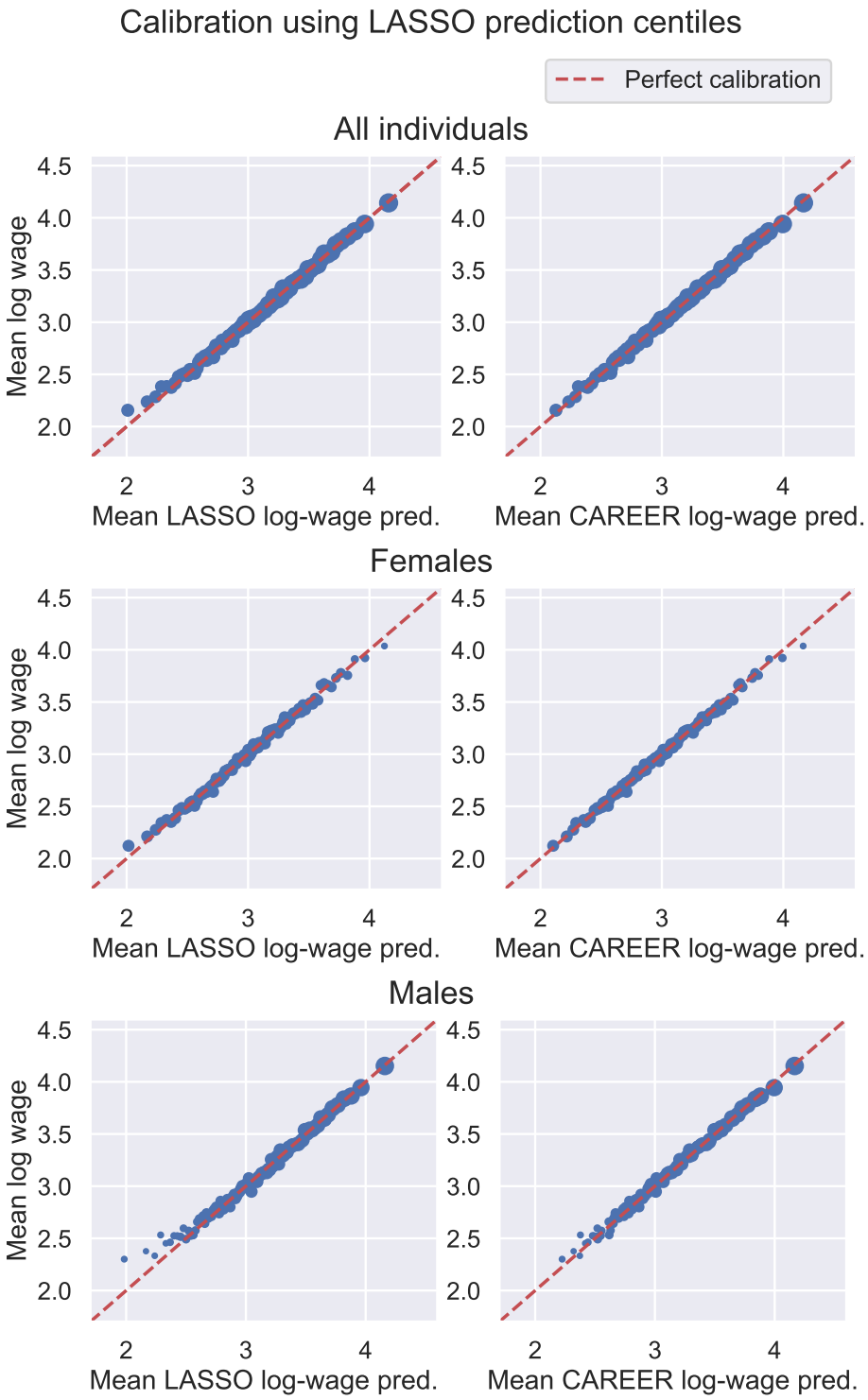

Figure 1: Calibration of log-wage predictions on PSID, using LASSO prediction centiles to create calibration buckets. Each dot represents a group of observations, determined by centiles of the LASSO wage model's held-out predictions. The size of each dot is proportional to the size of group (weighted using PSID's assigned weights). The X-axis shows each model's average held-out log-wage prediction for the group, while the Y-axis shows the group's true average log-wage. Calibration buckets are determined from the pooled population, and the same buckets are used for all the full population, females, and males.

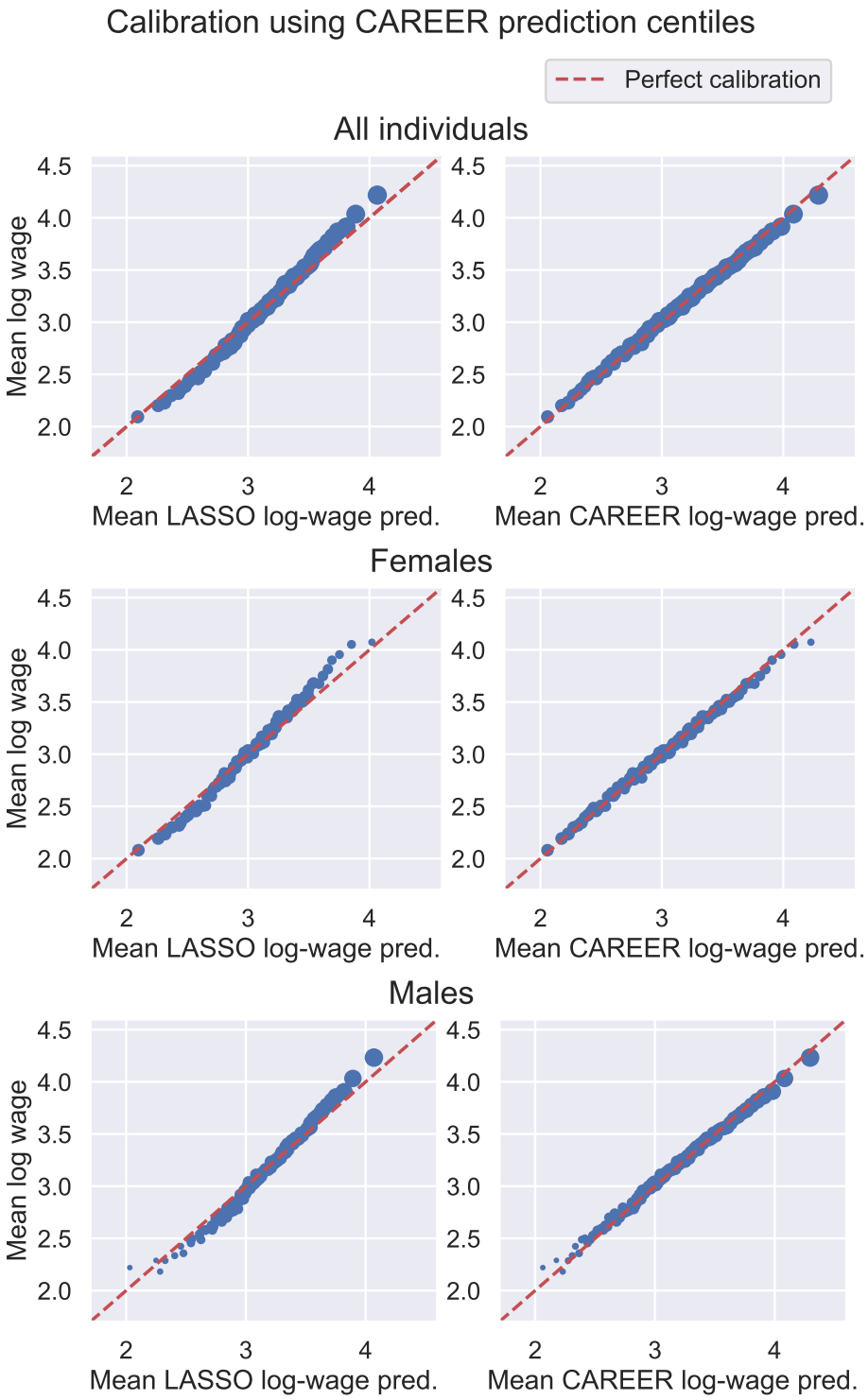

Figure 2: Calibration of log-wage predictions on PSID, using CAREER prediction centiles to create calibration buckets. Each dot represents a group of observations, determined by centiles of the CAREER wage model's held-out predictions. The size of each dot is proportional to the size of group (weighted using PSID's assigned weights). The X-axis shows each model's average held-out log-wage prediction for the group, while the Y-axis shows the group's true average log-wage. Calibration buckets are determined from the pooled population, and the same buckets are used for all the full population, females, and males.

our model, and that they can be incorporated into classical econometric models to improve their predictions.

Next, we introduce a heuristic for refining occupation labels based on CAREER's representation of histories. This heuristic involves fitting a model to predict an individual's current job from CAREER's representation of history. If the predicted job is different from the actual job, it suggests that there are aspects from their work history that are more informative for their current work than their listed job. We then "correct" listed jobs, changing them to the predicted job from CAREER's representation. We show that when baseline models that do not incorporate history are trained on the corrected jobs, they make better predictions than when they are trained on the listed jobs.

Representations of history refine occupations. The question we seek to answer here is: what are the aspects of history that improve CAREER's wage predictions? A challenge for this analysis is that CAREER's representations are difficult to interpret directly; each history has a different representation, and the representations are too high-dimensional to gain insights from directly. Instead of interpreting the representations directly, we partition them into discrete clusters. These clusters can be treated like any discrete variables. We feed them directly into classical wage models to improve their wage predictions. Then, by analyzing the clusters of history that are most helpful for improving wage, we can analyze characteristics of jobs for each cluster.

Specifically, we form clusters of histories by dividing an individual's current occupation into partitions based on the representations of history.[11] As described in Section 4, each occupation belongs to one of 21 coarse-grained categories. We aggregate all observations for each occupational category, and then partition the histories for these observations into 30 clusters using a machine learning clustering algorithm.[12]

Our goal is to find the clusters of history that best enhance the predictive power of the baseline LASSO model. Instead of refitting the LASSO model, we aim to predict the difference between

[11] We use the held-out representation function to represent each history in a fold. This is to ensure that the representations don't encode observe wages. We find that although representation functions vary slightly across cross-fit models due to training dynamics, the final representations are comparable across folds. This is because they are initialized using the same representation, so the updates will be the same in expectation across folds.

[12] We first use UMAP (McInnes et al., 2018) to project each representation into 2-dimensional space, on top of which we use K-Means to form 30 clusters for each coarse-grained occupation.

the original model's wage prediction and the actual wage, using these clusters. If a cluster can accurately predict this difference on new, unseen data, it means that it would improve the original model's predictions.

We use a regression tree to add history clusters to the predictions of the LASSO model. A regression tree incorporates one cluster at a time into its prediction of the output. At each step, it selects the cluster that would most decrease the training MSE if added. As a result, the sequence in which the regression tree selects clusters provides us with an understanding of their relative importance.

Figure 7 in Appendix C shows the held-out log-wage MSE when adding history clusters to the original predictions of the LASSO baseline. The first 10 or so clusters substantially improve the MSE, after which the curve begins to plateau. Even when all clusters are included in the regression tree, the performance does not reach that of the full CAREER model. There are a few reasons for this. CAREER, which learns continuous representations for each history, is more flexible than models that treat groups of similar histories discretely. All histories in a cluster are not identical. So by differentiating between histories in a cluster, CAREER can more flexibly model wages than a model that treats all histories in a cluster the same. Additionally, although each cluster is treated separately by the regression tree, CAREER's continuous representations allow the model to pool information across histories. For example, if software engineering managers and hardware engineering managers are clustered into different categories, the regression tree cannot use the wages of software engineering managers to help predict those of hardware engineering managers. Meanwhile, CAREER is able to pool this information since its representations of history are not discrete.

What are the characteristics of the most important clusters? Rather than exhaustively enumerate each occupation in a cluster, we use a heuristic to identify the most prevalent current jobs and historical jobs for each cluster. Specifically, for each cluster and job, we compute the proportion of histories in the cluster that contain the job. We also compute the proportion of histories in the broader occupational category that contain the job. We take the jobs that are most prevalent in the cluster relevant to the rest of the broad occupational category, only taking jobs that are present in

Occupational category	Most prevalent current job	Most prevalent previous jobs	Held-out MSE
manager	chief executive	software developer, electrical engineer, computer systems analyst	0.208
manager	manager/administrator	household appliance repairer, electrical equipment repairer, writer and author	0.207
manager	manager/administrator	cashier, homemaker, secretary/stenographer	0.207
manager	manager/administrator	cook, housekeeper, food preparataion worker	0.206
manager	manager/administrator	truck driver, machine operator, freight laborer	0.205
office/administrative support	bank teller	child care worker, bank teller, homemaker	0.205
sales	retail salesperson	child care worker, homemaker, cashier	0.204
construction/extraction/installation	machine operator	homemaker, cashier, retail salesperson	0.204
sales	sales supervisor	chief executive, sales supervisor, insurance sales	0.204
lawyer/physician	lawyer/judge	barber, bookkeeper, homemaker	0.203

Table 2: The history clusters that most improve the predictive performance of the LASSO wage model, shown in the order they were added to a regression tree. The regression tree is trained on 80% of the pooled PSID data, and evaluated on the remaining 20%. The held-out MSE for the original LASSO model without history indicators is 0.210. The most prevalent current and previous job columns show the most common occupations in comparison to the rest of the coarse-grained occupational category, as determined by the heuristic described in Section 6.

more than 2.5% of clusters in the history. We compute an analogous heuristic for current jobs.

Table 2 shows the top clusters identified by the regression tree. Half of these clusters are formed by partitioning manager jobs into finer-grained histories. These clusters reveal important aspects of history that refine an individual's current occupation. For example, the top cluster consists of managers who were previously software developers, electrical engineers, and computer systems analysts. Managers with these jobs in their history get paid more than managers without them. These refinements may reflect differences in occupations that are not captured by the initial encoding (there is no occupational category for "engineering manager"). While models that do not incorporate history omit these factors, they are captured by our model.

Reclassifying occupations. The analysis above showed that CAREER's representations partition encoded occupations into finer-grained categories that improve wage predictions. The finer-grained partitions identified were not present in the original occupation encoding scheme; for example, there is no occupation category for managers who have previously been engineers. Thus, these partitions form new kinds of occupations. So to incorporate them as part of traditional analyses, one must introduce additional job categories, similar to the regression tree exercise.

Here we seek to identify occupations that can be refined into other occupations that are *already present in the data*. For example, if an individual has worked for 15 years as a nurse and they are listed as a "manager" for the following year, a more accurate encoding may be "medical manager" (which is one of the categories present in the PSID occupational encoding scheme). Then, to incorporate these refined occupations into existing analyses, one can use the original taxonomy of occupations; one can simply change individual occupation labels.

Here, we describe a straightforward heuristic based on CAREER's representations that can "correct" an individual's current occupation. This heuristic centers on building a model to predict an individual's current occupation from CAREER's representation of their history and current occupation. For most representations, the likeliest current occupation will be an individual's true occupation. However, if there are representations where the likeliest current occupation is different from the true occupation, it reveals that CAREER groups them with other occupations that it

believes are more relevant for predicting its wage. We demonstrate that models trained on the corrected occupation, as opposed to the original occupation, make better held-out predictions.

Denote by H_L an individual's current occupation. Recall that CAREER encodes an individual's trajectory of jobs, $H = ((H_1, D_1), \ldots, (H_L, D_L))$, into a representation $\lambda_\theta(H) \in \mathbb{R}^D$. We first build a model, $p(H_L = h_L | \lambda_\theta(h))$, to predict an individual's current job given their history. We use K-nearest neighbors with $K = 100$ as our model, using cosine distance to quantiy distances between representations.[13] Then, for each individual, we estimate the top predicted job:

$$\hat{h}_{Li} = \arg\max_j p(H_L = j | \lambda_\theta(h_i)). \tag{53}$$

We define "corrected" occupations $\tilde{h}_i$ based on the predictions of the model. Specifically,

$$\tilde{h}_{Li} = \begin{cases} \hat{h}_{Li} & \text{if } \hat{h}_{Li} \neq h_{Li} \text{ and } p\left(H_{Li} = \hat{h}_{Li} | \lambda_\theta(h_i)\right) > 0.50. \\ h_{Li} & \text{otherwise.} \end{cases} \tag{54}$$

In other words, if the model predicts an occupation that differs from the one listed and assigns it a probability higher than 50%, we then adopt this predicted occupation as the "corrected" occupation. However, if the model doesn't meet this threshold, we retain the originally listed occupation.

Table 3 shows examples of corrected occupations. Most examples consist of long strings of the same occupation, from which the current occupation differs. For example, one example is an individual who has been recorded as a public service detective for their whole career, with the exception of the current year, when they are recorded as a construction inspector. The model "corrects" the occupation to public service detective.

To assess whether the corrected occupations are informative, we first evaluate the calibration of the wage models among the population of individuals whose occupations are corrected. We consider two bucketing functions $\Pi(x, h)$: one that buckets individuals based on their listed current

[13] We again use cross-fitting to make sure that the representations do not encode wage. For each individual i, we use the representation function from their held-out fold to represent their job history. To build the K-nearest neighbors classifier, we consider all neighboring histories, using the same representation function as individual i. Although the representation will be applied to individuals that it was trained on, individual i's predicted occupation can depend on the wages of other examples in the dataset as long as i's predicted occupation doesn't depend on i's observed wage.

Previous occupations	Listed occupation	Corrected occupation
employed, employed, welder, driller, miner, plant operator, plant operator, plant operator, truck driver, truck driver, truck driver, freight handler, freight handler, truck driver, freight handler, truck driver, truck driver, truck driver	miner	truck driver
employed, lawyer/judge, lawyer/judge, lawyer/judge, lawyer/judge, lawyer/judge, lawyer/judge	human resource clerk	lawyer/judge
lawyer/judge, lawyer/judge, lawyer/judge, lawyer/judge, lawyer/judge, financial manager, lawyer/judge, lawyer/judge, lawyer/judge, lawyer/judge, lawyer/judge, lawyer/judge, lawyer/judge	financial manager	lawyer/judge
public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective, public service detective	construction inspector	public service detective

Table 3: Occupation histories where the highest-likelihood predicted occupation using CAREER's representation is different from the labeled current occupation. The listed examples include a random sample of trajectories for which the top predicted occupation is assigned at least a 99% likelihood of being the listed occupation.

occupation, h_L; and another that buckets individuals based on their "corrected occupation", $\tilde{h}_L$.

Figure 3 shows the calibration of both the LASSO and CAREER models using both bucketing functions. When individuals are bucketed by their listed occupation, both models are reasonably well-calibrated. Both models condition on an individual's listed job, so it makes sense that even the baseline is calibrated here. However, when bucketing by corrected job, the LASSO model is poorly calibrated, while CAREER maintains its strong calibration. Neither the predictions nor the true wages are changing between plots; just the bucketing function. These differences suggest that the predicted occupation from history is an important omitted variable for the LASSO model.

If the corrected occupations provide more signal than the listed occupations, then models trained on the corrected occupations instead of the listed ones should make better held-out predictions of wage. We use our heuristic to replace an original's listed occupation with the one imputed from CAREER's history. We compare models trained on the listed and corrected occupations, evaluating their performance on a held-out sample of occupations whose labels were corrected.

Table 4 compares the predictive performance of LASSO and regression models on the subset

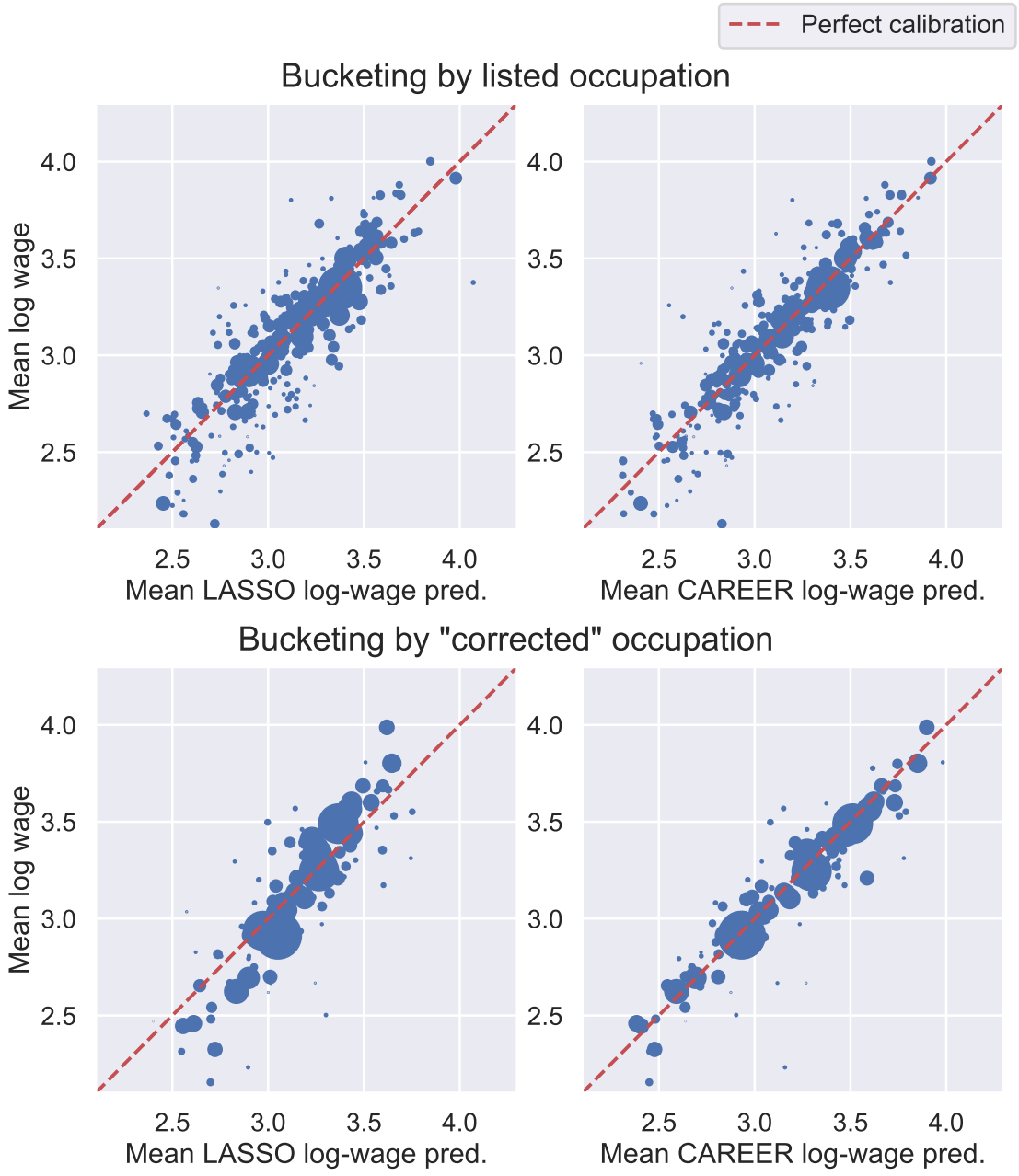

Figure 3: Calibration of log-wage predictions on the sample of observations where the top predicted occupation using CAREER's representation is different from the labeled current occupation. Each dot represents a group of observations, determined by the bucketing function. The first row buckets individuals based on their listed current occupation, while the second row buckets individuals based on the occupation predicted from CAREER's representations. The size of each dot is proportional to the size of group (weighted using PSID's assigned weights). The X-axis shows each model's average held-out log-wage prediction for the group, while the Y-axis shows the group's true average log-wage.

Model	Metric	Training on listed occupation	Training on corrected occupation
Coarse-grained regression	MSE	0.216 (0.006)	0.206 (0.006)
	R^2	0.358 (0.015)	0.385 (0.014)
Fine-grained LASSO	MSE	0.213 (0.007)	0.197 (0.006)
	R^2	0.368 (0.015)	0.407 (0.015)

Table 4: The log-wage MSE and R^2 of the regression and LASSO wage models on a held-out sample of trajectories for which the current occupation is "corrected" using CAREER's representation of history. 8,740 of the 90,074 total histories are corrected. The models in the "Training on listed occupation" column are trained using the listed occupations, while the models in the "Training on corrected occupation" column are trained using the corrected occupations. Estimated standard errors are in parentheses.

of corrected occupations. In both cases, training on the corrected occupations improves predictive performance. The improvement is larger for the LASSO model, which uses fine-grained occupations. The corrected occupations may have the same coarse-grained occupations as the uncorrected ones, making them identical to the original features for the coarse-grained regression. The corrected features differ more from their original values for LASSO than they do for the regression, so LASSO's relative improvement is sensible.

7 Decomposing Wage Gaps

Having demonstrated the predictive capabilities of our model in Section 6, we now use it to decompose gender wage gaps. We begin by decomposing cross-sectional wage gaps, where we show that the model with history explains about 25% of the wage gap that is unexplained when history is not included. We then move on to decomposing changes in gender wage gaps over careers. We apply our wage model to decompose changes in the gender wage gap into differences in starting characteristics and transitions. We study two populations over 12-year intervals: one that begins early in the workers' careers (ages 25 to 35), and the other later in careers (ages 40 to 50). We find that for the younger population, the gender wage gap increases over time, and the change is driven by differences in career transitions. In contrast, for the older population, the gender wage gap narrows over time; however, the change is not being driven by female transitions being more valuable, but rather by differences in starting characteristics.

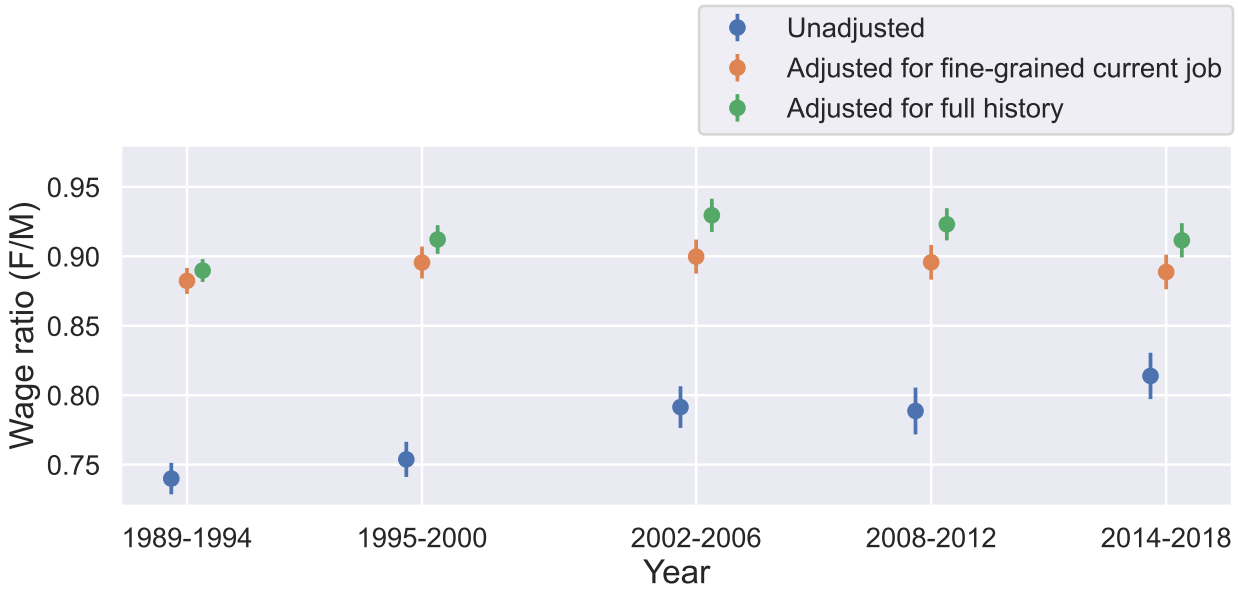

Figure 4: Estimates of the unexplained wage ratio (the exponentiated unexplained wage gap) on PSID. The LASSO model conditions on an individual's current occupation and summary statistics about experience, while the CAREER model conditions on full history. Meanwhile, the unadjusted model shows the raw wage gaps. The sample includes wage and salary workers between 25 and 64 years old who worked for at least 26 weeks in non-farm jobs, and whose work histories are in the middle 98% of the gender distribution (to assure overlap). Standard errors are estimated by bootstrapping.

7.1 Decomposing cross-sectional wage gaps

We begin by analyzing the gender wage gap at a fixed point in time. We decompose the gender wage gap into unexplained (6) and explained (7) components using a Blinder-Oaxaca decomposition (Blinder, 1973; Oaxaca, 1973). We further break up the explained wage gap into per-variable explanations, taking advantage of the linear separability of our model.

We partition observations from the PSID into multi-year buckets and perform the Blinder-Oaxaca decomposition for each bucket. We compare the LASSO model with fine-grained occupations to the wage model that uses CAREER to represent complete histories. Figure 4 presents the unexplained wage ratio for each year bucket (the unexplained wage ratio is the exponentiated unexplained wage gap). Adjusting for history explains a material portion of the wage gap for each year bucket after 1994. It is difficult to draw conclusions for why history does not explain any additional portion of the wage gap between 1989-1994. While it is possible that history indeed

	LASSO		CAREER	
	log points	Percent of gap explained	log points	Percent of gap explained
Total (F-M) wage gap	-0.241 (0.013)	—	-0.241 (0.013)	—
Explained by experience	-0.044 (0.004)	18.4% (1.5%)	-0.016 (0.001)	6.8% (0.6%)
Explained by education	0.011 (0.006)	-4.5% (2.4%)	0.007 (0.003)	-2.8% (1.3%)
Explained by region	-0.002 (0.001)	1.0% (0.6%)	-0.002 (0.001)	1.0% (0.5%)
Explained by demographic	-0.006 (0.001)	2.4% (0.5%)	-0.003 (0.001)	1.1% (0.2%)
Explained by year	-0.000 (0.000)	0.2% (0.2%)	-0.000 (0.000)	0.0% (0.1%)
Explained by union	-0.003 (0.002)	1.4% (1.0%)	-0.003 (0.002)	1.1% (0.9%)
Explained by industry	-0.031 (0.003)	12.6% (1.0%)	-0.024 (0.002)	10.1% (0.7%)
Explained by occupation	-0.054 (0.005)	22.3% (1.9%)	-0.007 (0.002)	3.1% (0.7%)
Explained by all non-history variables	-0.130 (0.010)	54.0% (4.1%)	-0.050 (0.005)	20.5% (1.9%)
Explained by history	—	—	-0.107 (0.008)	44.5% (3.2%)
Unexplained	-0.109 (0.009)	45.2% (3.9%)	-0.086 (0.008)	35.7% (3.5%)
Cross-fitting error	-0.002 (0.003)	0.8% (1.1%)	0.002 (0.002)	-0.7% (0.9%)

Table 5: Decomposing the cross-sectional 1995-2018 gender wage gap (61,032 total observations) using an Oaxaca-Blinder decomposition. The LASSO model summarizes history with summary statistics about experience, while the CAREER model adjusts for full history. The sample includes wage and salary workers between 25 and 64 years old who worked for at least 26 weeks in non-farm jobs, and whose work histories and covariates are in the middle 98% of the gender distribution (to assure overlap). The amount explained by each non-history covariate also includes an interaction with year. The cross-fitting error term arises from the fact that the decomposition is performed on held-out data, so the residual isn't exactly zero. Estimated standard errors are in parentheses.

explains less of the wage gap for these years, it is also possible that the comparatively smaller explained gap is due to estimation error. Detailed histories were only collected beginning in 1979, so the histories are shorter for this group of years (see Table 10). CAREER makes better predictions when longer histories are available, and indeed its predictive performance is closer to those of the baselines for 1989-1994 than it is for the other year buckets.

Table 5 depicts a Blinder-Oaxaca decomposition for the pooled samples from 1995-2018. The raw gender wage gap is -0.241 log points in this sample. Without adjusting for history, -0.130 log points are explained. Adjusting for history increases the amount explained to -0.157 log points. Thus, adding history explains about 24% of the remaining wage gap.

When history is not included, the covariates that explain the most of the gap are occupation (-0.054 log points), followed by experience (-0.044 log points) and industry (-0.031 log points). Meanwhile, when adjusting for history, the representation of history explains almost half of the wage

gap. This representation includes an individual's current occupation along with partial measures of experience, since it includes observations for each year an individual is in the survey. The amounts explained by all other covariates are similar between models.

Additional decompositions are included in Appendix D. Table 12 includes the decomposition for the first year bucket, 1989-1994, where the model with history does not explain any more of the wage gap than the model without it. Table 13 decomposes the wage gap for the last bucket of the sample, 2014-2018, where the model with history explains an additional 22% of the wage gap.

7.2 Decomposing changes in gender wage gaps

We now decompose the change in wage gaps over time using the decomposition described in Section 2. We focus on studying changes in the gender wage gap for a consistent set of individuals who are full-time wage workers at two endpoints of 12-year intervals. Most of the gender wage gap literature focuses on cross-sections who are working at a single point in time; by focusing on individuals who are present at both endpoints, our sample consists of individuals who are more committed to the labor force.

We perform two sets of decompositions for two separate populations. The first population consists of individuals who are between 25-35 years of age at the beginning of a 12-year interval, taking place between 1989-2018. The second population follows an older population, consisting of individuals between 40-50 years old at the beginning of a 12-year interval in the same time period.

We consider three wage models for the decomposition. The main model we consider is the full CAREER model described in Section 3. In order to put the results for CAREER into perspective, we also consider two baseline models: the LASSO model with fine-grained occupation that does not incorporate an individual's full job history, and the CAREER model that incorporates an individual's current job and workforce participation trajectory but not their previous occupations.

Table 6 shows the decomposition for the younger population, who begin the interval aged between 25-35. The gender wage gap at the beginning of the interval is -0.196 log points, and it widens to -0.244 log points after 12 years. The baseline wage models attribute most of this change to the effect of changing unexplained gap. Meanwhile, the full CAREER model attributes

	LASSO	CAREER (Participation status only for previous jobs)	CAREER (Full information for previous jobs)
Number of individuals	6858	6858	6858
Total change in (F-M) gap	-0.049 (0.020)	-0.049 (0.020)	-0.049 (0.020)
Effect of changing history	—	-0.010 (0.008)	-0.037 (0.011)
Effect of differential starting history	—	0.047 (0.011)	0.057 (0.013)
Effect of differential history transitions	—	-0.057 (0.013)	-0.094 (0.015)
Effect of changing all non-history covariates	-0.000 (0.013)	0.001 (0.009)	0.004 (0.006)
Effect of differential starting non-history covariates	0.081 (0.016)	0.043 (0.011)	0.034 (0.008)
Effect of differential non-history covariate transitions	-0.081 (0.019)	-0.042 (0.014)	-0.030 (0.010)
Effect of changing experience	-0.003 (0.005)	-0.006 (0.003)	-0.001 (0.002)
Effect of differential starting experience	0.002 (0.005)	-0.003 (0.004)	0.000 (0.002)
Effect of differential experience transitions	-0.005 (0.008)	-0.003 (0.005)	-0.002 (0.003)
Effect of changing education	0.023 (0.008)	0.020 (0.007)	0.014 (0.004)
Effect of differential starting education	0.037 (0.009)	0.030 (0.008)	0.020 (0.005)
Effect of differential education transitions	-0.014 (0.011)	-0.010 (0.009)	-0.006 (0.006)
Effect of changing region	-0.001 (0.001)	-0.000 (0.001)	-0.000 (0.001)
Effect of differential starting region	0.004 (0.002)	0.004 (0.002)	0.004 (0.001)
Effect of differential region transitions	-0.005 (0.002)	-0.004 (0.002)	-0.004 (0.002)
Effect of changing demographic	-0.003 (0.001)	-0.002 (0.001)	-0.001 (0.000)
Effect of differential starting demographic	-0.004 (0.001)	-0.003 (0.001)	-0.002 (0.001)
Effect of differential demographic transitions	0.001 (0.001)	0.001 (0.001)	0.001 (0.001)
Effect of changing union	0.006 (0.004)	0.005 (0.003)	0.005 (0.003)
Effect of differential starting union	0.006 (0.003)	0.005 (0.003)	0.004 (0.003)
Effect of differential union transitions	0.000 (0.004)	0.000 (0.003)	0.000 (0.003)
Effect of changing industry	-0.005 (0.005)	-0.004 (0.004)	-0.005 (0.003)
Effect of differential starting industry	0.002 (0.005)	0.003 (0.004)	0.000 (0.003)
Effect of differential industry transitions	-0.007 (0.006)	-0.007 (0.005)	-0.006 (0.004)
Effect of changing occupation	-0.015 (0.008)	-0.009 (0.005)	-0.006 (0.003)
Effect of differential starting occupation	0.038 (0.010)	0.009 (0.005)	0.005 (0.003)
Effect of differential occupation transitions	-0.053 (0.012)	-0.017 (0.006)	-0.012 (0.004)
Effect of changing year	-0.003 (0.003)	-0.002 (0.002)	-0.001 (0.001)
Effect of changing unexplained gaps	-0.032 (0.019)	-0.034 (0.020)	-0.011 (0.021)
Effect of changing model	-0.015 (0.009)	-0.006 (0.008)	-0.002 (0.007)
Cross-fitting error	-0.002 (0.010)	-0.000 (0.010)	-0.001 (0.009)
Initial (F-M) gap	-0.196 (0.023)	-0.196 (0.023)	-0.196 (0.023)
Initial unexplained gap	-0.112 (0.018)	-0.090 (0.018)	-0.103 (0.019)
Final (F-M) gap	-0.244 (0.028)	-0.244 (0.028)	-0.244 (0.028)
Final unexplained gap	-0.144 (0.018)	-0.124 (0.018)	-0.114 (0.020)

Table 6: Decomposing the 12-year change in the gender wage gap for individuals aged 25-35 into differential starting characteristics and transitions using the decomposition in Section 2. The population consists of full-time workers who worked at least 26 weeks in non-farm jobs and whose characteristics are in the middle 98% of the gender distribution at the beginning and end of 12-year intervals taking place between 1989-2018. The amount explained by each non-history covariate also includes an interaction with year (this is why the change explained by non-evolving covariates, such as demographic, is not always 0).

only −0.011 log points of this change to changing unexplained gaps. Instead, it attributes most of the change in the wage gap to changing histories, explaining −0.037 of the −0.049 log point differential.

The effect of changing history is decomposed into effects of differential starting histories and differential transitions using the decomposition from Section 2. Table 6 reveals that for the full CAREER model, the change explained by history is being driven by differences in male and female transitions. The model expects the wage gap explained by history to shrink in magnitude by 0.057 log points for males and females making the same transitions with different starting histories. However, females on average transition to substantially lower-value histories than males; the wage gap explained by history is expected to widen by 0.095 log points for males and females with the same starting histories making different transitions.

Comparing the full CAREER model to the version of CAREER that only conditions on previous workforce participation reveals that the change in explained gaps is not just being driven by differences in workforce participation. The representation of workforce participation and current job explains only −0.010 log points of the −0.049 log point wage gap differential, compared to −0.037 for the full representation of history. Both models attribute a similar effect to differences in male and female initial histories. However, females transition to lower-value histories than males even after controlling for workforce participation; the model with workforce participation attributes −0.057 log points of the change explained by history to differences in history transitions, while CAREER attributes −0.094 log points.

Table 7 further breaks down the decomposition based on initial occupations. Individuals are divided into four quartiles based on the average wage of their initial-period occupation, and we decompose the change in gender wage gaps for each quartile using the full CAREER wage model. The wage gap increases by the largest magnitude for individuals with initial occupations in the highest-wage quartile, from −0.133 log points to −0.228 log points. Meanwhile, the wage gap decreases in magnitude for the lowest-wage quartile, from −0.222 log points to −0.185 log points. However, female histories become less valuable relative to male histories after 16 years for all quartiles. For the lower quartiles, the decomposition suggests that female histories stand to become

	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Number of individuals	1838	1596	2038	1386
Share female	0.516 (0.022)	0.446 (0.023)	0.402 (0.022)	0.356 (0.023)
Initial average wage	2.692 (0.021)	2.831 (0.021)	3.068 (0.013)	3.312 (0.021)
Initial (F-M) gap	-0.222 (0.047)	-0.157 (0.034)	-0.090 (0.028)	-0.133 (0.044)
Total change in (F-M) gap	0.037 (0.038)	-0.028 (0.037)	-0.058 (0.032)	-0.095 (0.049)
Effect of changing history	-0.040 (0.019)	-0.001 (0.020)	-0.041 (0.017)	-0.017 (0.027)
Effect of differential starting history	0.071 (0.041)	0.150 (0.049)	-0.001 (0.019)	0.038 (0.034)
Effect of differential history transitions	-0.111 (0.037)	-0.152 (0.047)	-0.040 (0.021)	-0.055 (0.036)
Effect of changing all non-history covariates	0.009 (0.014)	-0.010 (0.013)	0.016 (0.009)	-0.018 (0.016)
Effect of differential starting non-history covariates	0.069 (0.023)	0.020 (0.035)	0.034 (0.016)	0.014 (0.027)
Effect of differential non-history covariate transitions	-0.060 (0.026)	-0.030 (0.037)	-0.018 (0.017)	-0.032 (0.028)
Effect of changing unexplained gaps	0.073 (0.045)	-0.035 (0.037)	-0.052 (0.031)	-0.067 (0.056)
Effect of changing model	-0.017 (0.026)	-0.004 (0.018)	-0.009 (0.013)	-0.000 (0.023)
Cross-fitting error	0.013 (0.029)	0.022 (0.027)	0.028 (0.020)	0.006 (0.039)

Table 7: Decomposing the 12-year change in the gender wage gap for individuals aged 25-35 grouped by quartiles of initial occupation average wage. For each quartile, the decomposition from Section 2 is performed using the CAREER model to decompose the change in the gender wage gap into differential starting characteristics and transitions. The population consists of individuals who worked at least 26 weeks in non-farm jobs and whose characteristics are in the middle 98% of the gender distribution at the beginning and end of 12-year intervals taking place between 1989-2018.

more valuable relative to those of males for the same transitions. However, this effect is offset by female transitions becoming less valuable than those of males'. Meanwhile, for the higher quartiles, the female transitions are less valuable than those of males' but by a smaller magnitude than for the lower quartiles. Instead, they have relatively less to gain from making the same transitions as males.

In Table 8, we perform the decomposition for the older population, consisting of individuals between 40-50 at the beginning of 12-year intervals. Here, the raw wage gap shrinks in magnitude, from -0.362 log points to -0.310 log points. The decreased wage gap is partially due to female histories becoming 0.015 log points more valuable relative to male histories after 12 years. However, this effect isn't being driven by females transitioning to higher-value histories than males during the interval; in fact, females make modestly less valuable transitions (-0.023 log points, although with a large standard error). The increased value of female histories is instead being driven by differences in starting period histories (0.038 log points). Although female transitions are not more valuable than those of males, the relative value of female histories increases because females are

	LASSO	CAREER (Participation status only for previous jobs)	CAREER (Full information for previous jobs)
Number of individuals	5857	5857	5857
Total change in (F-M) gap	0.052 (0.021)	0.052 (0.021)	0.052 (0.021)
Effect of changing history	—	0.021 (0.009)	0.015 (0.009)
Effect of differential starting history	—	0.081 (0.019)	0.038 (0.024)
Effect of differential history transitions	—	-0.060 (0.020)	-0.023 (0.025)
Effect of changing all non-history covariates	0.059 (0.015)	0.034 (0.010)	0.025 (0.007)
Effect of differential starting non-history covariates	0.103 (0.025)	0.061 (0.019)	0.048 (0.015)
Effect of differential non-history covariate transitions	-0.044 (0.028)	-0.028 (0.020)	-0.023 (0.015)
Effect of changing experience	0.045 (0.005)	0.023 (0.004)	0.019 (0.003)
Effect of differential starting experience	0.040 (0.011)	0.023 (0.009)	0.020 (0.007)
Effect of differential experience transitions	0.006 (0.013)	0.001 (0.010)	-0.002 (0.007)
Effect of changing education	0.011 (0.007)	0.010 (0.006)	0.006 (0.004)
Effect of differential starting education	0.010 (0.017)	0.008 (0.014)	0.005 (0.010)
Effect of differential education transitions	0.001 (0.018)	0.002 (0.016)	0.001 (0.010)
Effect of changing region	0.001 (0.001)	0.001 (0.001)	0.001 (0.001)
Effect of differential starting region	0.001 (0.002)	0.002 (0.002)	0.002 (0.002)
Effect of differential region transitions	-0.000 (0.002)	-0.001 (0.002)	-0.001 (0.002)
Effect of changing demographic	-0.002 (0.001)	-0.001 (0.001)	-0.000 (0.000)
Effect of differential starting demographic	-0.001 (0.001)	-0.000 (0.001)	0.000 (0.001)
Effect of differential demographic transitions	-0.001 (0.001)	-0.001 (0.001)	-0.001 (0.001)
Effect of changing union	-0.000 (0.004)	-0.001 (0.003)	-0.000 (0.003)
Effect of differential starting union	0.011 (0.006)	0.009 (0.006)	0.008 (0.005)
Effect of differential union transitions	-0.011 (0.007)	-0.010 (0.006)	-0.008 (0.006)
Effect of changing industry	0.001 (0.004)	0.003 (0.004)	0.001 (0.003)
Effect of differential starting industry	-0.008 (0.008)	-0.002 (0.007)	-0.004 (0.005)
Effect of differential industry transitions	0.009 (0.009)	0.005 (0.008)	0.004 (0.006)
Effect of changing occupation	0.002 (0.009)	-0.000 (0.005)	-0.000 (0.003)
Effect of differential starting occupation	0.038 (0.015)	0.009 (0.009)	0.007 (0.005)
Effect of differential occupation transitions	-0.036 (0.017)	-0.009 (0.010)	-0.008 (0.005)
Effect of changing year	-0.000 (0.004)	-0.001 (0.002)	-0.000 (0.001)
Effect of changing unexplained gaps	-0.002 (0.026)	-0.009 (0.026)	0.024 (0.024)
Effect of changing model	-0.018 (0.012)	-0.010 (0.013)	-0.023 (0.012)
Cross-fitting error	0.013 (0.011)	0.015 (0.010)	0.011 (0.010)
Initial (F-M) gap	-0.362 (0.025)	-0.362 (0.025)	-0.362 (0.025)
Initial unexplained gap	-0.110 (0.021)	-0.086 (0.020)	-0.092 (0.021)
Final (F-M) gap	-0.310 (0.030)	-0.310 (0.030)	-0.310 (0.030)
Final unexplained gap	-0.111 (0.025)	-0.095 (0.024)	-0.067 (0.025)

Table 8: Decomposing the 12-year change in the gender wage gap for individuals aged 40-50 into differential starting characteristics and transitions using the decomposition in Section 2. The population consists of full-time workers who worked at least 26 weeks in non-farm jobs and whose characteristics are in the middle 98% of the gender distribution at the beginning and end of 12-year intervals taking place between 1989-2018. The amount explained by each non-history covariate also includes an interaction with year (this is why the change explained by non-evolving covariates, such as demographic, is not always 0).

12-year lags				
	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Number of individuals	1546	1437	1651	1223
Share female	0.707 (0.023)	0.419 (0.026)	0.465 (0.022)	0.319 (0.025)
Initial average wage	2.751 (0.023)	3.078 (0.020)	3.360 (0.019)	3.607 (0.021)
Initial (F-M) gap	-0.185 (0.047)	-0.148 (0.042)	-0.286 (0.032)	-0.274 (0.045)
Total change in (F-M) gap	0.096 (0.047)	0.085 (0.035)	0.042 (0.028)	-0.028 (0.036)
Effect of changing history	0.033 (0.019)	0.028 (0.017)	-0.011 (0.014)	0.056 (0.020)
Effect of differential starting history	-0.011 (0.070)	0.108 (0.053)	0.027 (0.032)	0.107 (0.052)
Effect of differential history transitions	0.043 (0.077)	-0.080 (0.053)	-0.038 (0.033)	-0.051 (0.051)
Effect of changing all non-history covariates	0.026 (0.014)	0.004 (0.013)	0.035 (0.010)	0.008 (0.012)
Effect of differential starting non-history covariates	0.018 (0.056)	0.056 (0.033)	0.022 (0.031)	0.020 (0.042)
Effect of differential non-history covariate transitions	0.008 (0.056)	-0.053 (0.034)	0.013 (0.032)	-0.012 (0.042)
Effect of changing unexplained gaps	0.023 (0.048)	0.062 (0.042)	0.033 (0.035)	-0.079 (0.071)
Effect of changing model	0.005 (0.042)	-0.008 (0.022)	-0.024 (0.023)	-0.046 (0.029)
Cross-fitting error	0.009 (0.049)	-0.001 (0.031)	0.008 (0.023)	0.032 (0.049)

Table 9: Decomposing the 12-year change in the gender wage gap for individuals aged 40-50 grouped by quartiles of initial occupation average wage. For each quartile, the decomposition from Section 2 is performed using the CAREER model to decompose the change in the gender wage gap into differential starting characteristics and transitions. The population consists of individuals who worked at least 26 weeks in non-farm jobs and whose characteristics are in the middle 98% of the gender distribution at the beginning and end of 12-year intervals taking place between 1989-2018.

beginning with less valuable histories on average.

Table 9 breaks down the decomposition based on quartiles of initial-period average occupational wage for the older population. The effects are closer to zero, but the results follow generally the same patterns as before. For the second, third, and fourth quartiles, differences in initial histories shrink the explained wage gap, while differences in transitions expand it. The results differ for the first quartile, where the magnitude of the wage gap explained by history decreases after 12 years. Here, although the model does not expect the wage gap to shrink for males and females making the same transitions with different starting histories, females transition to higher-value histories than males do. However, the estimated effects have large standard errors, making it difficult to draw firm conclusions.

8 Conclusion

In this paper, we investigated how changes in the gender wage gap over careers are related to male and female characteristics evolving in different ways. We proposed a decomposition of the change in the explained gender wage gap into two portions: differential transitions (keeping initial characteristics fixed) and differential initial characteristics (keeping transitions fixed). We demonstrated that to accurately estimate these terms from data, it is crucial to incorporate detailed descriptions of occupational trajectories. We proposed new methods based on machine learning to more fully account for full worker history when predicting wages.

We applied our methods to survey data from the Panel Study of Income Dynamics, where our model outperformed classical econometric approaches for predicting wages. The representations of history learned by this method explained more of the cross-sectional wage gap than approaches that omit full history. We studied two populations over 12-year intervals, one younger and one older. For the younger population, we found the explained wage gap to increase, driven by males transitioning to higher-earning characteristics than females. In contrast, for the older population, we found the explained wage gap to narrow, although it was not driven by females making more valuable transitions. Rather, the change was attributed to the female workforce beginning with lesser-earning characteristics, with more to gain from making the same transitions as males.

References

Abowd, J. M., Kramarz, F., and Margolis, D. N. High wage workers and high wage firms. *Econometrica*, 67(2):251–333, 1999.

Altonji, J. G. and Blank, R. M. Race and gender in the labor market. *Handbook of Labor Economics*, 3:3143–3259, 1999.

Autor, D. and Dorn, D. The growth of low-skill service jobs and the polarization of the U.S. labor market. *American Economic Review*, 103(5):1553–97, 2013.

Barth, E., Kerr, S. P., and Olivetti, C. The dynamics of gender earnings differentials: Evidence from establishment data. Technical report, National Bureau of Economic Research, 2017.

Benson, A. Rethinking the two-body problem: The segregation of women into geographically dispersed occupations. *Demography*, 51(5):1619–1639, 2014.

Bertrand, M., Goldin, C., and Katz, L. F. Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American Economic Journal: Applied Economics*, 2(3): 228–55, 2010.

Black, D. A., Haviland, A. M., Sanders, S. G., and Taylor, L. J. Gender wage disparities among the highly educated. *Journal of Human Resources*, 43(3):630–659, 2008.

Blau, F. D. and Kahn, L. M. The feasibility and importance of adding measures of actual experience to cross-sectional data collection. *Journal of Labor Economics*, 31(S1):S17–S58, 2013.

Blau, F. D. and Kahn, L. M. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865, 2017.

Blinder, A. S. Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4):436–455, 1973.

Breunig, R. and Rospabe, S. The male-female wage gap in France: An analysis using non-parametric methods. *Journal of School of Economics and Research School of Social Sciences Australian National University*, 2004.

Calamai, P. H. and Moré, J. J. Projected gradient methods for linearly constrained problems. *Mathematical Programming*, 39(1):93–116, 1987.

Card, D., Cardoso, A. R., and Kline, P. Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly Journal of Economics*, 131(2): 633–686, 2016.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), 2018.

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., and Syrgkanis, V. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022a.

Chernozhukov, V., Newey, W., Quintas-Martınez, V. M., and Syrgkanis, V. RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, 2022b.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report, National Bureau of Economic Research, 2006.

Devicienti, F. Shapley-value decompositions of changes in wage distributions: A note. *The Journal of Economic Inequality*, 8:35–45, 2010.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

Dietterich, T. G. Ensemble methods in machine learning. In *International Workshop on Mmultiple Classifier Systems*, pp. 1–15. Springer, 2000.

Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., and Smith, N. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Fortin, N., Lemieux, T., and Firpo, S. Decomposition methods in economics. In *Handbook of Labor Economics*, volume 4, pp. 1–102. Elsevier, 2011.

Goldin, C. The quiet revolution that transformed women's employment, education, and family. *American Economic Review*, 96(2):1–21, 2006.

Goldin, C. A Grand Gender Convergence: Its Last Chapter. *American Economic Review*, 104(4): 1091–1119, 2014.

Goldin, C., Kerr, S. P., Olivetti, C., and Barth, E. The expanding gender earnings gap: Evidence from the LEHD-2000 census. *American Economic Review*, 107(5):110–114, 2017.

Goraus, K., Tyrowicz, J., and Van der Velde, L. Which gender wage gap estimates to trust? A comparative analysis. *Review of Income and Wealth*, 63(1):118–146, 2017.

Guvenen, F., Karahan, F., Ozkan, S., and Song, J. What do data on millions of US workers reveal about lifecycle earnings dynamics? *Econometrica*, 89(5):2303–2339, 2021.

Heckman, J. J. and MaCurdy, T. E. A life cycle model of female labour supply. *The Review of Economic Studies*, 47(1):47–74, 1980.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., and Eck, D. Music transformer: Generating music with long-term structure. In *International Conference on Learning Representations*, 2019.

Huber, M. and Solovyeva, A. On the sensitivity of wage gap decompositions. *Journal of Labor Research*, 41:1–33, 2020.

Johnes, G. and Tanaka, Y. Changes in gender wage discrimination in the 1990s: A tale of three very different economies. *Japan and the World Economy*, 20(1):97–113, 2008.

Juhn, C., Murphy, K. M., and Pierce, B. Wage inequality and the rise in returns to skill. *Journal of Political Economy*, 101(3):410–442, 1993.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv:2001.08361*, 2020.

Killingsworth, M. R. and Heckman, J. J. Female labor supply: A survey. *Handbook of Labor Economics*, 1:103–204, 1986.

Kim, C. Decomposing the change in the wage gap between white and black men over time, 1980-2005: An extension of the Blinder-Oaxaca decomposition method. *Sociological Methods & Research*, 38(4):619–651, 2010.

Kimhi, A. and Hanuka-Taflia, N. What drives the convergence in male and female wage distributions in Israel? a Shapley decomposition approach. *The Journal of Economic Inequality*, 17:379–399, 2019.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Neural Information Processing Systems*, 2017.

Li, L., Jing, H., Tong, H., Yang, J., He, Q., and Chen, B.-C. NEMO: Next career move prediction with contextual embedding. In *World Wide Web Conference*, 2017.

Light, A. and Ureta, M. Early-career work experience and gender wage differentials. *Journal of Labor Economics*, 13(1):121–154, 1995.

Loprest, P. J. Gender differences in wage growth and job mobility. *The American Economic Review*, 82(2):526–532, 1992.

Manning, A. and Swaffield, J. The gender gap in early-career wage growth. *The Economic Journal*, 118(530):983–1024, 2008.

McInnes, L., Healy, J., and Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Menzel, A. and Woodruff, C. Gender wage gaps and worker mobility: Evidence from the garment

sector in Bangladesh. *Labour Economics*, 71:102000, 2021.

Miller, A. R. The effects of motherhood timing on career path. *Journal of Population Economics*, 24:1071–1100, 2011.

Monti, H., Stinson, M., and Zehr, L. How long do early career decisions follow women? The impact of employer history on the gender wage gap. *Journal of Labor Research*, 41(3):189–232, 2020.

Murphy, K. M. and Welch, F. Empirical age-earnings profiles. *Journal of Labor Economics*, 8(2): 202–229, 1990.

Oaxaca, R. Male-female wage differentials in urban labor markets. *International economic review*, 14(3):693–709, 1973.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Neural Information Processing Systems*, 2022.

Panel Study of Income Dynamics, public use dataset. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, 2023.

Rajkumar, K. *Causal and computational methods for the study of labor market frictions*. PhD thesis, Stanford University, 2021.

Regan, T. L. and Oaxaca, R. L. Work experience as a source of specification error in earnings models: Implications for gender wage decompositions. *Journal of Population Economics*, 22: 463–499, 2009.

Rubinstein, Y. and Weiss, Y. Post schooling wage growth: Investment, search and learning. *Handbook of the Economics of Education*, 1:1–67, 2006.

Schönberg, U. and Ludsteck, J. Maternity leave legislation, female labor supply, and the family wage gap. 2007.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019.

Shepherd, D. et al. Post-apartheid trends in gender discrimination in South Africa: Analysis through decomposition techniques. *Stellenbosch University, Department of Economics*, 5:2011, 2008.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. In *Neural Information Processing Systems*, 2019.

Sinning, M., Hahn, M., and Bauer, T. K. The Blinder–Oaxaca decomposition for nonlinear regression models. *The Stata Journal*, 8(4):480–492, 2008.

Smith, J. P. and Welch, F. R. Black economic progress after Myrdal. *Journal of Economic Literature*, 27(2):519–564, 1989.

Suh, J. Decomposition of the change in the gender wage gap. *Research in Business and Economics Journal*, 1:1, 2010.

Ulrick, S. W. A nonparametric analysis of the black/white wage gap. *Applied Economics Letters*, 12(13):811–815, 2005.

Vafa, K., Palikot, E., Du, T., Kanodia, A., Athey, S., and Blei, D. M. CAREER: Transfer learning for economic prediction of labor sequence data. *arXiv preprint arXiv:2202.08370*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Neural Information Processing Systems*, 2017.

Waldfogel, J. Understanding the 'family gap' in pay for women with children. *Journal of Economic Perspectives*, 12(1):137–156, 1998.

Wang, Z., Xie, Q., Ding, Z., Feng, Y., and Xia, R. Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*, 2023.

Wexler, M. N. Successful resume fraud: Conjectures on the origins of amorality in the workplace. *Journal of Human Values*, 12(2):137–152, 2006.

Wood, R. G., Corcoran, M. E., and Courant, P. N. Pay differences among the highly paid: The male-female earnings gap in lawyers' salaries. *Journal of Labor Economics*, 11(3):417–441, 1993.

A Training details

Training sufficient representations. The full procedure for training sufficient representations is depicted in Algorithm 1. This algorithm is identical to the procedure outlined in Section 3.1, with one addition. There is a risk that projecting to the space of sufficient representations degrades the quality of wage predictions. Thus, after the final round of projections, we perform an additional round of wage error minimization. Throughout our semi-synthetic experiments, we find that including this additional round of wage error minimization results in more accurate wage predictions and more accurate estimates of the unnexplained wage gap. Although this last step of the procedure does not then involve projecting the representation onto the space of sufficient representations, we find that wage gap estimates are still significantly more accurate than those from estimating the wage gap without projections. This is because this final round of wage error minimization is fine-tuned from a sufficient representation; the final representation is encouraged to be not only predictive of wage but also sufficient for gender.

Ensembling In practice, the conditional wage model can be sensitive to the initialization of the model parameters on resume data. To reduce variance, we take inspiration from ensemble methods in machine learning (Dietterich, 2000; Lakshminarayanan et al., 2017). Specifically, we use multiple models, or *ensembles*, to learn initial representations on the resume data, each model identical except initialized with a different random seed. We then perform the projection process for each model, and estimate wage gaps by averaging predictions for all models. Specifically, for K ensembles, each with estimated conditional means $\hat{\mu}_{g,k}$, we model

$$\hat{\mu}_g(x,h) = \frac{1}{K}\sum_{k=1}^{K}\hat{\mu}_{g,k}(x,h) = \frac{1}{K}\sum_{k=1}^{K}\hat{\mu}^X_{g,k}(x) + \frac{1}{K}\sum_{k=1}^{K}\hat{\mu}^H_{g,k}(\lambda_{\theta_k}(h)), \tag{55}$$

where $\mu^X_{g,k}(h)$ and $\mu^H_{g,k}(h)$ refer to the per-ensemble conditional wage covariate and history terms, respectively, and λ_{θ_k} refers to the ensemble-k representation of history,

This ensembling procedure reduces the variance of the conditional wage function with respect to randomness induced by the pretraining process. By the law of large numbers, as the number of ensembles goes to infinity, the variance induced by pretraining randomness will diminish. Additionally, we find that averaging the predictions of ensembles improves wage predictions compared to models that do not ensemble. We use $K = 10$ ensembles for our main experiments.

Using resumes to learn representations Our approach uses resume data to predict which jobs are in sequences together. Specifically, we begin by initializing the representation function λ_θ

Algorithm 1: Sufficiency-constrained optimization

Input: N samples $\{x_i, y_i, g_i, h_i\}$, initial representation $\hat{\theta}$

Output: Parameters $\hat{\theta}$ that induce approximate solution to (29) and (30), estimated wage function $\hat{\mu}_g(x, \lambda_\theta(h))$ and propensity model parameters $\hat{\gamma}, \hat{\eta}$.

Initialize: Feedforward parameters $\hat{\rho}_g$, regression coefficients $\hat{\alpha}_g, \hat{\beta}_g$, and logistic regression coefficients $\hat{\gamma}, \hat{\eta}$ randomly.

while wage validation error and gender propensity validation error are decreasing **do**

1. **Minimization Step:**
   - Initialize $\theta = \hat{\theta}$, $\rho_g = \hat{\rho}_g$, $\alpha_g = \hat{\alpha}_g$, $\beta = \hat{\beta}_g$.
   - Perform gradient descent to minimize $\frac{1}{N}\sum_i (y_i - \hat{\mu}_{g_i}(x_i, \lambda_\theta(h_i))^2$ with respect to θ, ρ_g, and α_g, β_g.
   - Define $\hat{\theta}$, $\hat{\rho}_g$, $\hat{\alpha}_g$, and $\hat{\beta}_g$ to be the parameters that induce the best validation loss.
2. **Projection Step:**
   - Initialize $\theta = \hat{\theta}, \gamma = \hat{\gamma}, \eta = \hat{\eta}$.
   - Perform gradient descent to maximize $\frac{1}{N}\sum_i 1(g_i = f)\log(p_i) + 1(g_i = m)\log(1 - p_i)$ for $p_i = \sigma\left(\frac{1}{1+\exp(-\gamma \cdot \lambda_\theta(h_i) + \eta \cdot x_i)}\right)$ with respect to representation parameters θ and logistic regression parameters γ, η.
   - Define $\hat{\theta}$, $\hat{\gamma}$, and $\hat{\eta}$, to be the parameters that induce the best validation score.

end

Perform: final minimization step until validation loss convergence.

return $\hat{\theta}$, $\hat{\rho}_g$, $\hat{\alpha}_g$, $\hat{\beta}_g$, $\hat{\gamma}_g$.

randomly. Then, we model:

$$p(H_t = j|h_S) = p_\theta(H_t = j|\lambda_{\theta,t}(h_S)) \propto \exp\left\{\xi_j^\top \lambda_{\theta,t}(h_S)\right\}, \tag{56}$$

where S is an index set of random timesteps, t indexes a timestep that is not present in S, and $\xi_j \in \mathbb{R}^D$ is a vector of coefficients. In other words, we mask jobs in a career at random and form likelihoods of the masked jobs using CAREER's representation of the unmasked jobs. We train by maximizing the likelihood

$$\mathbb{E}_H\left[\mathbb{E}_S\left[\sum_{t\notin S} \log p_\theta(H_t = j|\lambda_{\theta,t}(h_S))\right]\right], \tag{57}$$

where the outer expectation is over the empirical distributions of career trajectories H in resume data and the inner expectation is with respect to a masking distribution. We maximize the likelihood in (57) with respect to CAREER's parameters θ and the coefficients ξ_j by maximum likelihood estimation on sequences from resume data, re-sampling index sets S at each optimization step. Then, to model the history-adjusted wage gap with the projection algorithm (Algorithm 1), we initialize λ_θ with the fitted values.

Our end-to-end estimation procedure is given in Algorithm 2. This procedure includes resume pretraining with ensembles, projecting each representation with Algorithm 1, and predicting wages with cross-fitting.

B Semi-synthetic experiments

Ideally, we could assess the accuracy of wage gap estimates on real world data. However, ground truth wage gaps are not available in the real world. Instead, we compare approaches on semi-synthetic data. Semi-synthetic experiments are a common method to assess causal estimation strategies in the presence of high-dimensional confounders. In order for these experiments to reflect real-world data settings, we use real job histories H.

The main idea of the experiments are as follows: we generate data with a known, ground-truth adjusted wage gap. We then compare how close different estimation methods come to finding it.

These experiments begin by forming a confounder, a function of history that is correlated with both gender and wage. Models that do not account for this confounder will not be able to estimate

Algorithm 2: Predicting held-out wages with CAREER.

Input: Corpus of resumes $\mathscr{D}_R$ containing job sequences $\{h_i\}$; survey dataset $\mathscr{D}_S$ containing wages $\{y_i\}$, gender $\{g_i\}$, covariates $\{x_i\}$, and histories $\{h_i\}$; number of ensembles K; number of splits S for cross-fitting.

Output: Held-out estimates of male and female wages, $\hat{y}_i^{(m)}$ and $\hat{y}_i^{(f)}$.

Divide survey dataset $\mathscr{D}_S$ into S splits randomly.

for split s in $1 \dots S$ **do**

  Define $\mathscr{D}_{S,\text{test}}^{(s)} = \{(Y_i, G_i, X_i, H_i) \in \mathscr{D}_S \text{ s.t. } i = s\}$.

  Randomly split remainder of $\mathscr{D}_S^{(s)}$ into train set $\mathscr{D}_{S,\text{train}}^{(s)}$ and validation set $\mathscr{D}_{S,\text{valid}}^{(s)}$.

end

for ensemble k in $1 \dots K$ **do**

  Initialize parameters θ_k randomly.

  Pretrain on resume corpus $\mathscr{D}_R$ to optimize (56) with respect to λ_{θ_k}.

  for split s in $1 \dots S$ **do**

    Set $\hat{\mu}_{g,k}(x, h)$ by performing sufficiency-constrained optimization (Algorithm 1) with $\mathscr{D}_{S,\text{train}}^{(s)}, \mathscr{D}_{S,\text{valid}}^{(s)}, \theta_{k,s}$.

    for index i in $\mathscr{D}_{S,\text{test}}^{(s)}$ **do**

      Set $\hat{y}_{i,k}^{(m)} = \hat{\mu}_m(x_i, \lambda_{\theta_{k,s}}(h_i))$.

      Set $\hat{y}_{i,k}^{(f)} = \hat{\mu}_f(x_i, \lambda_{\theta_{k,s}}(h_i))$.

    end

  end

end

Set $\hat{y}_i^{(M)} = \frac{1}{K} \sum_k \hat{y}_{i,k}^{(M)}$.

Set $\hat{y}_i^{(F)} = \frac{1}{K} \sum_k \hat{y}_{i,k}^{(F)}$.

return $\hat{y}_i^{(M)}, \hat{y}_i^{(F)}$.

the true adjusted wage gap. Denoting the confounder as $\lambda_\phi(H)$, wages are then sampled as

$$Y_i = \tau * 1(G_i = f) + \gamma * (\hat{\pi}(\lambda_\phi(H_i)) - 0.5) + \epsilon_i,$$
$$\epsilon_i \sim \mathcal{N}\left(0, (0.1)^2\right),$$

where $\hat{\pi} : \mathbb{R}^K \to [0, 1]$ corresponds to a logistic regression fit from the confounder $\lambda_\phi(H)$ to predict gender, analogous to a propensity score. The parameter $\tau \in \mathbb{R}$ is the true parameter of interest, and $\gamma \in \mathbb{R}$ controls the confounding strength.

We use a transformer to model the confounder λ_ϕ. Specifically, we fit λ_ϕ to real data to predict wage. We use two settings: one where the confounder λ_ϕ is in the model class of the transformer used for estimation, and one where it is not. To generate a confounder λ_ϕ that is outside the model class used for estimation, we use a transformer architecture that is 20 times larger than than the one used for estimation. We overfit this model to the wage data to ensure that it is outside the model class of the smaller transformer. Then, $\hat{\pi}(\lambda_\phi(H))$ is generated by averaging the fitted propensity scores for this model and the one within the model class.

We compare five approaches for estimating the unexplained gender wage ratio (given by $\exp(\tau)$). The ununexplained gender wage ratio computes the difference in male and female average wages, while the non-history unexplained ratio models wage as a function of covariates X that includes summary statistics about history (Blau & Kahn, 2017). CAREER (no projections) uses CAREER to model wage without modeling gender. CAREER (joint optimization) follows Shi et al. (2019) and Chernozhukov et al. (2022b): it uses CAREER to jointly minimize the predictive error of wage and gender, controlled by a hyperparameter. Finally, CAREER (w/ projections) follows the sufficiency-constrained optimization procedure outlined in Algorithm 1.

Figure 5 depicts the results. Across settings, CAREER (with projections) estimates an unexplained wage ratio that is closest to the true underlying unexplained wage ratio. Although the unadjusted and non-history adjusted models perform fine when there is little confounding (as expected), they struggle as the confounding strength increases. We also find that the performance of the joint optimization technique is sensitive to the hyperparameter dictating the tradeoff between gender and wage predictive components.

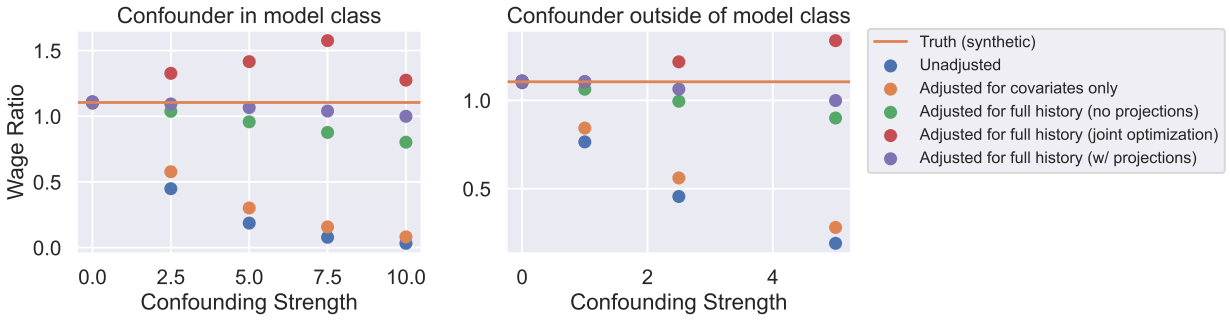

Figure 5: Semi-synthetic experiments comparing estimates of the unexplained wage ratio to the true wage ratio when the synthetic confounder is in and outside the model class. Five methods are compared: an unadjusted estimator; a classical estimator of the wage gap that conditions on only covariates and summary statistics about history ("unadjusted"); a version that uses CAREER as a wage model but does not try to enforce sufficiency ("no projections"); a version that jointly optimizes to predict wage and gender following Shi et al. (2019) ("joint optimization"); and a version that follows the sufficiency-constrained optimization approach we develop ("w/ projections"). The results use semi-synthetic data from a single year of PSID.

C Supplementary prediction results

	1989-1994	1995-2000	2002-2006	2008-2012	2014-2018
Number of observations	29042	17652	14206	14012	15162
Average history length	12.4	15.1	15.5	15.3	14.5
Total variance	0.334	0.350	0.383	0.388	0.410
Coarse-grained regression MSE	0.188 (0.004)	0.212 (0.005)	0.246 (0.007)	0.221 (0.006)	0.234 (0.006)
Coarse-grained LASSO MSE	0.184 (0.004)	0.207 (0.005)	0.240 (0.007)	0.216 (0.006)	0.229 (0.006)
Fine-grained LASSO MSE	0.176 (0.004)	0.199 (0.005)	0.229 (0.007)	0.207 (0.006)	0.218 (0.006)
CAREER (current job only) MSE	0.175 (0.004)	0.199 (0.005)	0.229 (0.007)	0.204 (0.006)	0.214 (0.006)
CAREER (partic. and current job only) MSE	0.167 (0.004)	0.190 (0.005)	0.224 (0.007)	0.201 (0.006)	0.213 (0.006)
CAREER (no resumes) MSE	0.166 (0.004)	0.183 (0.004)	0.217 (0.007)	0.193 (0.006)	0.202 (0.006)
CAREER (with resumes) MSE	0.152 (0.004)	0.170 (0.004)	0.205 (0.007)	0.178 (0.005)	0.190 (0.006)

Table 10: Log-wage predictive performance on PSID, separated by years. All models are fit by cross-fitting on the full pooled sample, dividing the data into five folds. All results are reported for held-out data. Estimated standard errors are in parentheses.

	2006	2008	2010	2012	2014	2016	2018
Coarse-grained regression	0.235	0.220	0.217	0.221	0.230	0.229	0.241
Fine-grained LASSO	0.227	0.217	0.216	0.218	0.223	0.227	0.236
CAREER (current job only)	0.227	0.216	0.213	0.215	0.222	0.222	0.236
CAREER (no resumes)	0.234	0.221	0.214	0.223	0.227	0.226	0.242
CAREER (with resumes)	0.197	0.194	0.192	0.194	0.203	0.205	0.217

Table 11: Log-wage MSE on PSID, separated by individual years. All models are fit by cross-fitting using only data from the year in question, dividing the data into five folds. All results are reported for held-out data.

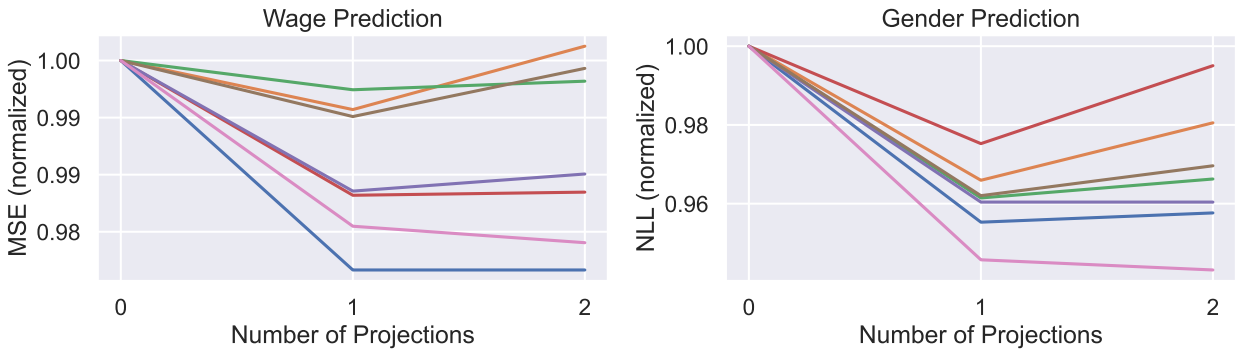

Figure 6: Log-wage mean-square error and gender negative-log likelihood as a function of projection round when training CAREER with sufficiency-constrained optimization Algorithm 2 . Each line is a different year of PSID, and each model is trained on only a single year. Prediction errors are normalized by dividing by the round-0 prediction error.

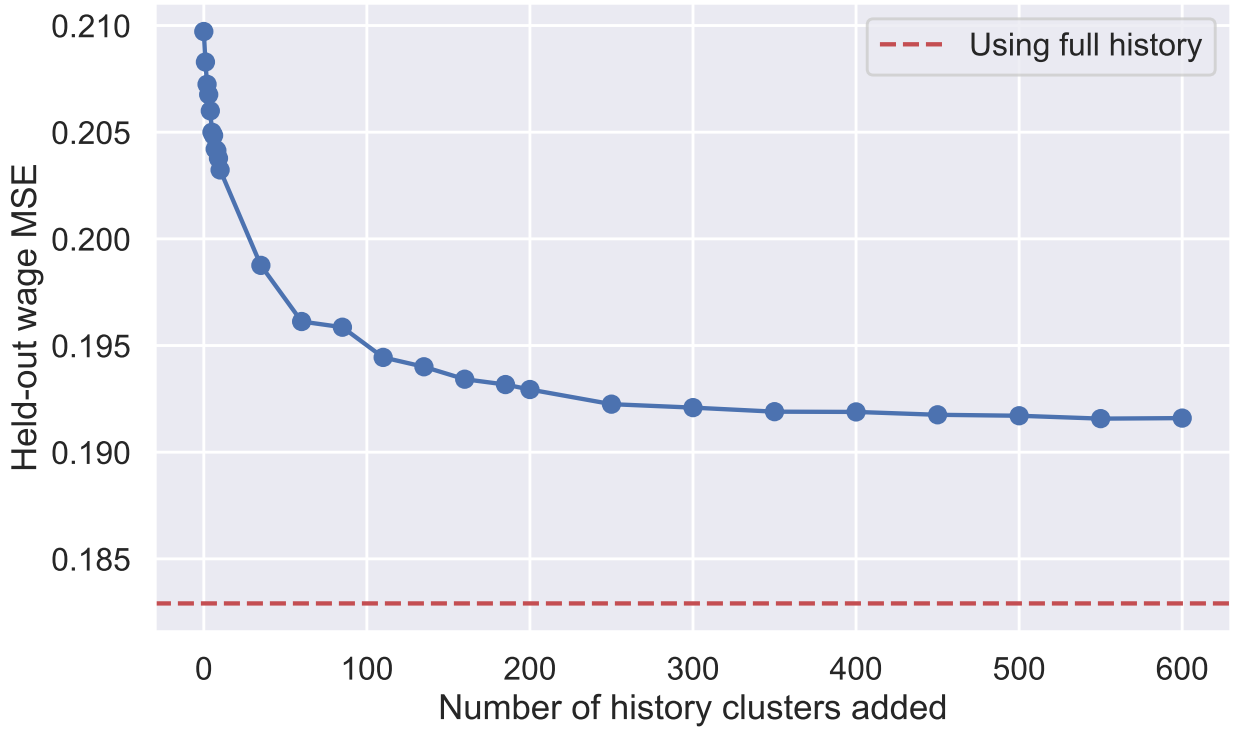

Figure 7: The held-out log-wage MSE of the LASSO wage model as more history cluster indicators are included in the regression tree. History clusters are determined from CAREER's representations of history. 80% of the pooled PSID data is used to train the regression tree, with the remaining 20% used for evaluation.

D Supplementary decompositions of cross-sectional wage gaps

	LASSO		CAREER	
	log points	Percent of gap explained	log points	Percent of gap explained
Total (F-M) wage gap	-0.301 (0.015)	—	-0.301 (0.015)	—
Explained by experience	-0.058 (0.004)	19.1% (1.4%)	-0.019 (0.002)	6.2% (0.6%)
Explained by education	-0.014 (0.006)	4.7% (2.0%)	-0.008 (0.004)	2.5% (1.2%)
Explained by region	-0.001 (0.002)	0.5% (0.6%)	-0.001 (0.002)	0.3% (0.6%)
Explained by demographic	-0.003 (0.001)	1.1% (0.3%)	-0.001 (0.000)	0.5% (0.2%)
Explained by year	0.000 (0.000)	-0.1% (0.1%)	0.000 (0.000)	-0.0% (0.1%)
Explained by union	-0.013 (0.004)	4.2% (1.2%)	-0.011 (0.003)	3.8% (1.1%)
Explained by industry	-0.027 (0.003)	8.9% (1.0%)	-0.020 (0.002)	6.6% (0.8%)
Explained by occupation	-0.065 (0.005)	21.7% (1.7%)	-0.010 (0.002)	3.4% (0.7%)
Explained by all non-history variables	-0.181 (0.013)	60.2% (4.2%)	-0.070 (0.007)	23.3% (2.4%)
Explained by history	—	—	-0.113 (0.008)	37.5% (2.8%)
Unexplained	-0.125 (0.011)	41.6% (3.5%)	-0.117 (0.009)	38.8% (3.1%)
Cross-fitting error	0.005 (0.004)	-1.8% (1.4%)	-0.001 (0.003)	0.4% (1.1%)

Table 12: Decomposing the cross-sectional 1989-1994 gender wage gap (29,042 total observations) using an Oaxaca-Blinder decomposition. The LASSO model summarizes history with summary statistics about experience, while the CAREER model adjusts for full history. The sample includes wage and salary workers between 25 and 64 years old who worked for at least 26 weeks in non-farm jobs, and whose work histories and covariates are in the middle 98% of the gender distribution (to assure overlap). The amount explained by each non-history covariate also includes an interaction with year. The cross-fitting error term arises from the fact that the decomposition is performed on held-out data, so the residual isn't exactly zero. Estimated standard errors are in parentheses.

	LASSO		CAREER	
	log points	Percent of gap explained	log points	Percent of gap explained
Total (F-M) wage gap	-0.206 (0.020)	—	-0.206 (0.020)	—
Explained by experience	-0.042 (0.006)	20.3% (2.9%)	-0.018 (0.003)	8.7% (1.2%)
Explained by education	0.041 (0.008)	-19.9% (4.0%)	0.022 (0.005)	-10.7% (2.3%)
Explained by region	-0.000 (0.002)	0.2% (1.0%)	-0.000 (0.002)	0.2% (0.8%)
Explained by demographic	-0.007 (0.002)	3.3% (0.8%)	-0.003 (0.001)	1.3% (0.3%)
Explained by year	0.000 (0.000)	-0.0% (0.0%)	0.000 (0.000)	-0.0% (0.0%)
Explained by union	0.000 (0.003)	-0.2% (1.3%)	0.001 (0.003)	-0.3% (1.4%)
Explained by industry	-0.031 (0.004)	14.9% (2.2%)	-0.024 (0.003)	11.9% (1.3%)
Explained by occupation	-0.049 (0.008)	23.6% (3.7%)	-0.009 (0.002)	4.3% (1.2%)
Explained by all non-history variables	-0.087 (0.014)	42.2% (6.9%)	-0.032 (0.007)	15.5% (3.5%)
Explained by history	—	—	-0.085 (0.012)	41.3% (5.6%)
Unexplained	-0.118 (0.014)	57.3% (6.8%)	-0.093 (0.014)	45.0% (6.6%)
Cross-fitting error	-0.001 (0.005)	0.5% (2.5%)	0.004 (0.004)	-1.8% (2.1%)

Table 13: Decomposing the cross-sectional 2014-2018 gender wage gap (15,162 total observations) using an Oaxaca-Blinder decomposition. The LASSO model summarizes history with summary statistics about experience, while the CAREER model adjusts for full history. The sample includes wage and salary workers between 25 and 64 years old who worked for at least 26 weeks in non-farm jobs, and whose work histories and covariates are in the middle 98% of the gender distribution (to assure overlap). The amount explained by each non-history covariate also includes an interaction with year. The cross-fitting error term arises from the fact that the decomposition is performed on held-out data, so the residual isn't exactly zero. Estimated standard errors are in parentheses.

E Supplementary decompositions of wage gap changes

	LASSO	CAREER (Participation status only for previous jobs)	CAREER (Full information for previous jobs)
Number of individuals	6858	6858	6858
Total change in (F-M) gap	-0.049 (0.020)	-0.049 (0.020)	-0.049 (0.020)
Effect of changing history	—	-0.011 (0.008)	-0.040 (0.011)
Effect of differential starting history	—	0.050 (0.011)	0.062 (0.013)
Effect of differential history transitions	—	-0.062 (0.014)	-0.101 (0.015)
Effect of changing all non-history covariates	0.003 (0.012)	0.003 (0.009)	0.004 (0.007)
Effect of differential starting non-history covariates	0.080 (0.012)	0.047 (0.009)	0.033 (0.007)
Effect of differential non-history covariate transitions	-0.076 (0.016)	-0.044 (0.013)	-0.029 (0.009)
Effect of changing experience	-0.000 (0.004)	-0.006 (0.003)	-0.001 (0.002)
Effect of differential starting experience	0.004 (0.003)	-0.002 (0.003)	0.001 (0.002)
Effect of differential experience transitions	-0.004 (0.004)	-0.003 (0.004)	-0.002 (0.003)
Effect of changing education	0.023 (0.008)	0.021 (0.007)	0.015 (0.004)
Effect of differential starting education	0.034 (0.007)	0.030 (0.006)	0.021 (0.004)
Effect of differential education transitions	-0.011 (0.009)	-0.009 (0.008)	-0.006 (0.006)
Effect of changing region	-0.000 (0.001)	-0.000 (0.001)	0.000 (0.001)
Effect of differential starting region	0.003 (0.001)	0.003 (0.001)	0.003 (0.001)
Effect of differential region transitions	-0.003 (0.002)	-0.003 (0.002)	-0.003 (0.001)
Effect of changing demographic	-0.003 (0.001)	-0.002 (0.001)	-0.001 (0.000)
Effect of differential starting demographic	-0.005 (0.001)	-0.004 (0.001)	-0.002 (0.001)
Effect of differential demographic transitions	0.001 (0.001)	0.001 (0.001)	0.001 (0.001)
Effect of changing union	0.005 (0.003)	0.005 (0.003)	0.004 (0.003)
Effect of differential starting union	0.005 (0.003)	0.004 (0.003)	0.004 (0.003)
Effect of differential union transitions	0.001 (0.003)	0.000 (0.003)	0.000 (0.003)
Effect of changing industry	-0.005 (0.005)	-0.005 (0.004)	-0.006 (0.003)
Effect of differential starting industry	0.003 (0.004)	0.004 (0.004)	0.001 (0.003)
Effect of differential industry transitions	-0.008 (0.005)	-0.009 (0.005)	-0.007 (0.004)
Effect of changing occupation	-0.013 (0.008)	-0.008 (0.005)	-0.006 (0.003)
Effect of differential starting occupation	0.038 (0.009)	0.012 (0.005)	0.006 (0.003)
Effect of differential occupation transitions	-0.052 (0.011)	-0.020 (0.007)	-0.013 (0.004)
Effect of changing year	-0.003 (0.003)	-0.002 (0.002)	-0.001 (0.001)
Effect of changing unexplained gaps	-0.035 (0.006)	-0.042 (0.005)	-0.013 (0.003)
Effect of changing model	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Cross-fitting error	-0.017 (0.020)	0.002 (0.019)	-0.000 (0.020)
Initial (F-M) gap	-0.196 (0.023)	-0.196 (0.023)	-0.196 (0.023)
Initial unexplained gap	-0.091 (0.005)	-0.055 (0.004)	-0.063 (0.003)
Final (F-M) gap	-0.244 (0.028)	-0.244 (0.028)	-0.244 (0.028)
Final unexplained gap	-0.126 (0.007)	-0.096 (0.005)	-0.076 (0.004)

Table 14: Decomposing the 12-year change in the gender wage gap for individuals aged 25-35 into differential starting characteristics and transitions using the decomposition in Section 2. For the results in this table, the **wage models are not adjusted to be unbiased for the decomposition population**. The population consists of full-time workers who worked at least 26 weeks in non-farm jobs and whose characteristics are in the middle 98% of the gender distribution at the beginning and end of 12-year intervals taking place between 1989-2018.

	LASSO	CAREER (Participation status only for previous jobs)	CAREER (Full information for previous jobs)
Number of individuals	6911	6911	6911
Total change in (F-M) gap	-0.048 (0.021)	-0.048 (0.021)	-0.048 (0.021)
Effect of changing history	—	-0.009 (0.010)	-0.038 (0.011)
Effect of differential starting history	—	0.048 (0.011)	0.060 (0.016)
Effect of differential history transitions	—	-0.057 (0.014)	-0.098 (0.017)
Effect of changing all non-history covariates	0.005 (0.014)	0.004 (0.010)	0.004 (0.007)
Effect of differential starting non-history covariates	0.087 (0.016)	0.048 (0.011)	0.036 (0.008)
Effect of differential non-history covariate transitions	-0.083 (0.019)	-0.045 (0.015)	-0.033 (0.009)
Effect of changing experience	-0.001 (0.004)	-0.005 (0.003)	-0.001 (0.002)
Effect of differential starting experience	0.001 (0.006)	-0.004 (0.005)	0.001 (0.003)
Effect of differential experience transitions	-0.002 (0.008)	-0.001 (0.006)	-0.002 (0.004)
Effect of changing education	0.025 (0.008)	0.021 (0.007)	0.014 (0.004)
Effect of differential starting education	0.036 (0.010)	0.028 (0.008)	0.022 (0.005)
Effect of differential education transitions	-0.011 (0.013)	-0.007 (0.010)	-0.007 (0.007)
Effect of changing region	-0.000 (0.001)	-0.000 (0.001)	-0.000 (0.001)
Effect of differential starting region	0.003 (0.002)	0.003 (0.002)	0.003 (0.001)
Effect of differential region transitions	-0.004 (0.002)	-0.004 (0.002)	-0.003 (0.001)
Effect of changing demographic	-0.004 (0.001)	-0.002 (0.001)	-0.001 (0.000)
Effect of differential starting demographic	-0.004 (0.001)	-0.003 (0.001)	-0.002 (0.001)
Effect of differential demographic transitions	0.001 (0.001)	0.001 (0.001)	0.001 (0.000)
Effect of changing union	0.005 (0.003)	0.004 (0.002)	0.004 (0.002)
Effect of differential starting union	0.005 (0.003)	0.004 (0.003)	0.004 (0.003)
Effect of differential union transitions	0.001 (0.003)	0.000 (0.003)	-0.000 (0.003)
Effect of changing industry	-0.006 (0.005)	-0.005 (0.004)	-0.006 (0.003)
Effect of differential starting industry	0.003 (0.006)	0.004 (0.005)	0.000 (0.003)
Effect of differential industry transitions	-0.009 (0.006)	-0.009 (0.005)	-0.006 (0.003)
Effect of changing occupation	-0.012 (0.010)	-0.007 (0.005)	-0.005 (0.003)
Effect of differential starting occupation	0.042 (0.010)	0.012 (0.005)	0.006 (0.003)
Effect of differential occupation transitions	-0.053 (0.011)	-0.019 (0.006)	-0.012 (0.003)
Effect of changing year	-0.003 (0.003)	-0.002 (0.002)	-0.001 (0.001)
Effect of changing unexplained gaps	-0.027 (0.021)	-0.027 (0.022)	-0.002 (0.023)
Effect of changing model	-0.026 (0.009)	-0.015 (0.008)	-0.006 (0.007)
Cross-fitting error	0.000 (0.010)	-0.001 (0.010)	-0.006 (0.009)
Initial (F-M) gap	-0.197 (0.022)	-0.197 (0.022)	-0.197 (0.022)
Initial unexplained gap	-0.121 (0.018)	-0.100 (0.018)	-0.119 (0.017)
Final (F-M) gap	-0.245 (0.030)	-0.245 (0.030)	-0.245 (0.030)
Final unexplained gap	-0.147 (0.020)	-0.127 (0.020)	-0.121 (0.018)

Table 15: Decomposing the 12-year change in the gender wage gap for individuals aged 25-35 into differential starting characteristics and transitions using the decomposition in Section 2. For the results in this table, **the population is not trimmed to ensure overlap**. The population consists of full-time workers who worked at least 26 weeks in non-farm jobs at the beginning and end of 12-year intervals taking place between 1989-2018.