The Role of Skills and Sorting in Explaining Wage Inequality

Paul Diegert[*]

Job Market Paper

November 8, 2022
Click here for the latest version

Abstract

A large literature argues that technological change since the 1980s altered the demand for workers' skills, increasing wage inequality and polarization. By estimating a model of occupational choice using panel data from the Survey of Income and Program Participation (SIPP), I find that changes in the *supply* of workers' skills were also major driving factors in increasing inequality and polarization. Specifically, I find that (1) as tasks in high-skill jobs have become increasingly complex, the distribution of workers' ability to perform those tasks has become more dispersed, (2) workers' ability to perform low-skill work tasks has become more homogenous, and (3) workers have increasingly sorted into occupations by skill level, even if this does not maximize their income. These results suggest that skill formation has been a key channel through which long run changes in the nature of work have affected wage inequality. Finally, to obtain my estimates I prove a new identification result in a multi-dimensional potential outcome model and show how to robustly estimate it semiparametrically adapting results from mixture models.

JEL classification: J24; J31; O33; C33; C14

Keywords: Wage inequality; labor market polarization; occupational choice; skills; job tasks

[*] Department of Economics, Duke University. Email: paul.diegert@duke.edu. Address: 213 Social Sciences, 419 Chapel Drive, Box 90097, Durham, NC 27708-0097 USA. I am grateful to Arnaud Maurel, Joe Hotz, Matt Masten, and Peter Arcidiacono for many helpful discussions and comments.

1 Introduction

Wage inequality has been increasing in the United States since the 1980s, but this trend has not be uniform. Notably, there was an episode of polarization around the turn of the century, in which wages and employment in middle-income occupations fell relative to both high- and low-income occupations. One of the most influential explanations for this has been the *Routine-Biased Technological Change* (RBTC) hypothesis that technology began to substitute for workers who performed tasks that could be automated. This theory was formalized most notably in a model proposed in Acemoglu and Autor (2011), which has inspired an empirical literature focused on testing the predictions of the model, particularly that the relative prices for occupations involving routine tasks fell during the period of polarization (e.g., Cortes 2016; Cavaglia and Etheridge 2020; Gottschalk, Green, and Sand 2015; Böhm 2020).

The model in Acemoglu and Autor (2011) and related work (e.g., Jung and Mercenier 2014; Costinot and Vogel 2010) provide a compelling explanation for how automation could cause a decline in employment and wages in middle-income jobs. However, the predictions of this stylized model focus on the short-run effects of a labor demand shock, maintaining two key assumptions on labor supply: (1) that workers' skills are fixed before and after the labor demand shock, and (2) that workers respond to a labor demand shock by switching into the occupation that maximizes their income. The first assumption is plausible in the short-run, but growing evidence suggests that there have been important changes in the skills of workers on a longer time frame (e.g., Altonji, Bharadwaj, and Lange 2012). The second assumption is called into question by a growing literature that has shown that non-pecuniary considerations play a significant role in workers' occupational choices (e.g. Arcidiacono et al. 2020; Wiswall and Zafar 2018).

In this paper, I develop an approach to identifying and estimating a model of occupational choice which relaxes these two assumptions in order to explore how changes in workers' skills and occupational sorting patterns have contributed to trends in wage inequality. Following the insight in Acemoglu and Autor (2011) that workers cannot immediately update their skills when skill prices change, I maintain the assumption that individual workers' skills are approximately fixed in a period of three years. However, I allow the distribution of workers' skills to change over a longer time frame as new cohorts enter the labor market who received different training and education, and as workers build different skills on the job. To explore the importance of changing occupational sorting patterns, my identification strategy does not impose any restrictions on how workers sort into occupations by skill level.

Estimating changes to the skill distribution over time without imposing assumptions on the way workers sort into occupations presents an empirical challenge. Many of the seminal works on wage polarization and RBTC in the United States use cross-sectional data such as the Census and the American Community Survey (ACS) (Autor et al. 2006; Autor, Katz, and Kearney 2008; Autor and Dorn 2013; Beaudry, Green, and Sand 2016). But it is well known that even a model with pure Roy sorting — in which workers choose occupations to maximize their earnings — is not point identified from cross-sectional labor market data without strong parametric assumptions (Heckman and Honore 1990; Mourifie, Henry, and Meango 2018). One strand of the empirical literature on wage polarization has focused on combining the assumption of pure Roy sorting with additional data such as skill measurements or panel data (Böhm 2020; Gottschalk, Green, and Sand 2015). On the other hand, common identification approaches to a generalized Roy model typically impose stringent data requirements, relying on exogenous measurements such as cognitive test scores or instruments for occupational choice with large support (see e.g., Carneiro, Hansen, and Heckman 2003).

In this application, however, the assumption that individuals' skills are fixed in the short term suggests an alternative panel data approach to identification using only commonly available labor market data. In particular, I provide a result showing that the joint distribution of skills and occupational choice probabilities can be point identified using a short panel of wage observations. The key assumptions in this approach are that skills are low dimensional compared to the number of wage measurements, and that time-varying productivity shocks are independent from skills. These are similar to the identifying assumptions in standard panel data models with multidimensional individual effects (Freyberger 2018). Under a suitable rank condition essentially requiring that wages in all occupations are sufficiently informative about skills, I show that identification can be established using wages in endogenously selected occupations as measurements of skills.

I estimate the model using data from the Survey for Income and Program Participation (SIPP). The SIPP is an ideal fit for this empirical strategy because it provides a sequence of nationally representative samples who are tracked for 3-4 years. While the panel dimension of this data is short compared to some alternative panel datasets, it is long enough to satisfy the identifying assumptions of the identification result for each panel. The sequence of nationally representative samples provides a basis for measuring change in workers' skills over time. Practically, the distinction between the short-term stickiness of skills and the long-term evolution of the skill distribution is operationalized by assuming that skills are fixed during each panel, but that the skill distribution can evolve between panels. Because of the focus on flexibly estimating changes in workers skills and occupational sorting patterns, I employ a semiparametric estimation approach which estimates the joint distribution of skills and choice probabilities nonparametrically. As this can be a high-dimensional object, I develop an accelerated Expectation Maximization (EM) method, which builds on advances in semiparametric estimation of mixture models to make computation feasible.

Estimation results reveal that changes in the skills and occupational sorting patterns of workers played an important role in the major trends in wage inequality. First, I find workers have increasingly sorted into occupations by skill level, which has increased wage inequality. An interesting aspect of this result is that this sorting pattern does not appear to be driven by income maximization. Under the estimated results, many workers in lower-income occupations pass up potential wage premia they could get from switching to high-income occupations, and this has increased over time. This result adds to the evidence in several recent papers, which have highlighted the importance of non-pecuniary factors in occupational choices (e.g. Arcidiacono et al. 2020; Wiswall and Zafar 2018). While prior work has focused on the occupational choices of college graduates, the findings in this paper suggests this is true more broadly across all workers.

The finding that this sorting pattern has increased wage inequality underscores the importance of flexibly estimating workers' occupational sorting patterns. Classic analysis of the Roy model (Heckman and Honore 1990) showed that under the relatively mild conditions, pure Roy sorting decreases overall dispersion in wages. However, as this result does not hold in the absence of pure Roy sorting, the relationship between occupational sorting and inequality is a question for empirical research. I find that sorting does appear to be part of the story of rising wage inequality and that the role of occupational sorting reversed during this period of study. Compared to randomly assigning workers to occupations, the estimated sorting pattern decreases wage inequality at the beginning of the period but increasing inequality at the end.[1]

The second major finding is that the aggregate pattern of wage polarization was driven by changes in the skills of workers within occupations. One measure of wage polarization is increasing inequality in the upper tail of the wage distribution and decreasing inequality in the lower tail. The RBTC hypothesis focuses on explaining this pattern through declining skill prices in middle-income occupations, which implies that polarization was driven by changes between occupational groups. I find evidence of a slight decline in the potential wages for production occupations, relative to the other two occupations, which is a central prediction of the theoretical analysis of wage polarization. But the magnitude of that decline does not account for most of the change in wage inequality. The decline in lower tail inequality was largely driven by declining dispersion in the skills of workers in service occupations. This effect of homogenization in the skills of service workers appears to account for nearly all the decline in lower tail inequality in the wage distribution. On the other hand, I find that the increase in upper tail inequality was driven primarily by increasing dispersion in the wages of professional workers.

These results highlight the importance of relaxing the assumptions of pure Roy sorting and long-run fixed skills in order to understand the most important trends in the market for workers' skills. Occupational sorting on nonpecuniary factors is important not only in understanding the short-term response of workers to labor demand shocks, but also contributes to the long-term increase in wage inequality. On the other hand, the finding that changes in the skill distribution were the major driver of the observed patterns of wage polarization suggests that any theory of technological change needs to focus on the direct and indirect effects of technological change on shaping workers' skills.

The rest of the paper is organized as follows. I conclude this section with a brief review of several literatures related to this paper. In section 2, I describe the SIPP data and explore the basic trends in wage inequality since the 1980s. I show that the major trends described in the literature on wage polarization are present in the SIPP and highlight several trends suggesting there have been major changes in workers' skills. In section 3, I describe a flexible model of occupational choice and wage determination, and show how it can be specialized to encompass the task-based model used frequently in the literature. In section 4, I show how this model can be nested by a general multidimensional potential outcome model, and develop a general identification result for models with endogenously selected measurements. In section 5, I describe the semiparametric specification I will use to estimate the model, and describe the Accelerated EM algorithm that I will use to feasibly estimate the model. In Section 6, I present the results from estimating the model. In section 7 I summarize the findings and discuss future directions to extend this research.

Related Literature

This paper contributes more broadly to an extensive literature which examines how the skill-biased technological change (SBTC) or routine-biased technological change (RBTC) affected wage inequality in this period. This paper returns to a classic concern in the SBTC literature on measuring and interpreting changes in residual wage inequality after account for demographic change (Autor, Katz, and Krueger 1998; Lemieux

[1] Interestingly, the finding that sorting decreased inequality at the beginning of the sample is consistent with work in Heckman and Sedlacek (1985) estimating a parametric version of the generalized Roy model from cross-sectional data.

2006; Autor, Katz, and Kearney 2008). While increasing in wage inequality within demographic groups was sometimes taken to be evidence of increases in the price for high-skill labor, it was not generally possible to measure these price changes directly. I provide an approach to identifying and estimating the contributions of changes in skill prices and the distribution of skills in explaining residual wage inequality.

After the seminal work of Autor, Levy, and Murnane (2003), and the observation that changes in wage inequality was not monotone after the 1980s, much of the focus in the literature shifted to *task* framework. The core insight of this framework is that the price paid for labor is not for a worker's skills (as it is e.g., in the "cannonical model" described in Acemoglu and Autor (2011)), but rather for the tasks they perform. The arguments laid out in it provide a compelling way to explain the decline in middle-wages jobs after the 1980s: automation made it possible to use machines to replace "routine" tasks, driving down the price for those tasks. I adopt the task framework in estimating prices for job tasks, which empirically are tied to groupings of occupations as is common in the literature. The main innovation in my approach is to provide an empirical strategy that allows me to relax the two assumptions on labor supply discussed above.

Most closely related to my approach is the small but growing literature of papers that adopt the task framework and estimate the evolution of (unobserved) prices for the tasks performed by workers. Aside from the papers mentioned above that use Roy sorting as part of their identification strategy, another set of papers uses a panel data approach more closely related to my approach (Cortes 2016; Cavaglia and Etheridge 2020). Using a slightly stronger time-homogeneity assumption,[2] these papers estimate changes in task prices by taking first differences among workers who stay in the same occupation. The main difference in my approach is that I estimate changes in the distribution of skills as well as skill prices. The short-panel approach taken here makes it possible to distinguish between short-term effects when skills are fixed, and long-term evolution in the distribution of skills.

The finding in this paper that nonpecuniary considerations are important in workers' occupational choices is not new. In early work generalizing the Roy model to allow for nonpecuniary preferences over occupations, Heckman and Sedlacek (1985, 1990) reject the hypothesis of pure Roy sorting between manufacturing and nonmanufacturing occupations. Much of the subsequent work on the generalized Roy model has focused on the related question of the returns to schooling. A major theme in this literature is analyzing how a cluster of unobserved personality traits or "non-cognitive skills" jointly affect labor market outcomes and educational choices, potentially accounting for the "psychic costs" of schooling that lead workers to pass up monetary gains (Heckman, Stixrud, and Urzua 2006). More recent work (Arcidiacono et al. 2020; Wiswall and Zafar 2018) has confirmed that nonpecuniary factors play an important part in occupational choices of college graduates. In another related context, Sorkin (2018) finds that worker's preferences over firms is also not captured well by maximizing income.

Building on this body of literature, this paper confirms the finding that nonpecuniary considerations are important in workers' occupational choices and provides new quantitative estimates of its importance. My approach complements the work in Arcidiacono et al. (2020) and Wiswall and Zafar (2018) by considering choices over broad occupational categories without a restriction to college graduates. Connecting this

[2] In addition to assuming that skills are fixed, Cortes (2016), and Cavaglia and Etheridge (2020) also assume that the productivity of skills in each occupation is fixed over time. This follows the theory of RBTC discussed in Acemoglu and Autor (2011), but seems to rule on the possibility that technology changed the way people work.

population-level measure of sorting to the overall trends in inequality, I am able to quantify the effects of these sorting papers on the secular trends in wage equality. The methodology I employ builds on the econometric work on the generalized Roy model especially Carneiro, Hansen, and Heckman (2003). Notably, I follow their approach of modeling the joint distribution of potential wages as a factor model, but I focus on using endogenous wage measurements rather than exogenous measurements and instruments for occupational choice with large support.

2 Trends in Wage Inequality

In this section I describe several major trends in wage inequality in the United States since the 1980s using data from the Survey of Income and Program Participation (SIPP). The SIPP is an ideal dataset for studying changes in the skills and occupational sorting patterns of workers because it combines a short panel dimension with repeated representative samples of the US population. The panel dimension of the data helps reveal information about the underlying abilities and preferences of each sample of workers, while the sequence of representative samples makes it possible to analyze change over time at a population level. Data from the SIPP has been used for a number of related applications such as analysis of wage volatility (Moffitt et al. 2022; Carr, Moffitt, and Wiemers 2020), but research on wage polarization has typically used either standard cross-sectional datasets (e.g., Census and ACS) or long panels (e.g., the Panel Study of Income Dynamics for the United States (PSID)). I confirm here that the major empirical trends outlined in Autor, Katz, and Kearney (2008), Acemoglu and Autor (2011) characterizing wage polarization hold in the SIPP. I also highlight change in the distribution of wages within occupations, which does not fit neatly with the theory of RBTC, and calls for a more flexible investigation of how the distribution of workers' skills and their occupational sorting patterns changed over this period.

Data

The SIPP is a nationally representative survey of the U.S. population, which began in 1984 and continues through the present day. In contrast to other popular longitudinal surveys such as the Panel Study of Income Dynamics for the United States (PSID) or the National Longitudinal Survey of Youth (NLSY), the SIPP collects data on each panel for only two to five years, but a new representative sample is formed for each panel. To date, there have been a total of 17 SIPP panels. While this short panel design can make some analysis of lifecycle dynamics difficult, it achieves a sample size similar to the CPS while providing some repeated measurements of individuals over time. In this section I focus on cross-sectional aspects of the data. In the remainder of the paper, I build an identification strategy that exploits the panel dimension of the data to identify the unobserved skills of works. An advantage of the SIPP in this application is that it provides a sequence of new representative samples from 1985 to the present day with sample sizes that are large enough to analyze how the distribution of skills has changed over time.

The SIPP collects data on wages and job characteristics for up to two concurrent employment spells. I construct primary employment spells by choosing the longer spell whenever there is an overlap. For each employment spell, data is collected on the total monthly earnings, usual hours worked per week, and hourly wage for workers who are paid by the hour. I use the hourly wage as the primary measure of wages, because it corresponds to the concept of the price per unit of labor output. In order to include workers who are not

Table 1: Sample Size with Sample Restrictions

SIPP Panel Starting Year	(A) Demographic Subsample	(B) Employed During Period	(C) Three Periods of Employment Data
1984	15,639	14,076	7,500
1985	10,802	9,523	4,861
1986	8,957	8,001	3,943
1987	8,802	7,926	4,880
1988	9,173	8,062	3,754
1990	19,616	17,416	9,728
1991	12,728	11,208	6,214
1992	17,502	15,530	7,786
1993	17,548	15,459	7,840
1996	32,279	28,766	15,958
2001	29,875	26,152	13,067
2004	37,093	32,538	18,081
2008	36,628	31,534	12,174

Table 2: Demographics of Sample under Sample Restrictions

Year	1984	1985	1986	1987	1988	1990	1991	1992	1993	1996	2001	2004	2008
Age Group													
18 to 24	0.20	0.18	0.18	0.17	0.18	0.18	0.16	0.17	0.16	0.15	0.15	0.15	0.14
	(0.18)	(0.16)	(0.17)	(0.15)	(0.16)	(0.15)	(0.13)	(0.14)	(0.14)	(0.13)	(0.13)	(0.13)	(0.08)
25 to 39	0.41	0.43	0.44	0.45	0.44	0.44	0.46	0.44	0.43	0.42	0.39	0.36	0.35
	(0.45)	(0.47)	(0.47)	(0.48)	(0.46)	(0.47)	(0.48)	(0.47)	(0.46)	(0.44)	(0.41)	(0.38)	(0.37)
40 to 64	0.38	0.37	0.37	0.37	0.37	0.36	0.37	0.39	0.40	0.43	0.45	0.48	0.50
	(0.36)	(0.36)	(0.36)	(0.37)	(0.37)	(0.37)	(0.38)	(0.38)	(0.40)	(0.43)	(0.46)	(0.49)	(0.55)
Education Level													
Less than High School	0.16	0.16	0.15	0.15	0.14	0.14	0.12	0.12	0.12	0.13	0.14	0.12	0.11
	(0.15)	(0.15)	(0.13)	(0.13)	(0.13)	(0.13)	(0.11)	(0.11)	(0.11)	(0.12)	(0.12)	(0.11)	(0.09)
High School Degree	0.59	0.59	0.60	0.59	0.59	0.60	0.60	0.61	0.60	0.63	0.60	0.62	0.61
	(0.60)	(0.59)	(0.61)	(0.60)	(0.59)	(0.59)	(0.59)	(0.61)	(0.60)	(0.63)	(0.61)	(0.62)	(0.60)
College Degree	0.24	0.25	0.25	0.26	0.27	0.26	0.28	0.27	0.28	0.24	0.26	0.26	0.28
	(0.25)	(0.26)	(0.27)	(0.26)	(0.28)	(0.28)	(0.30)	(0.28)	(0.29)	(0.26)	(0.27)	(0.27)	(0.31)

paid on an hourly basis, I follow the usual practice of constructing an hourly wage variable by calculating the approximate number of hours worked each month and dividing the monthly earnings by monthly hours worked. Wages are deflated using the CPI and pegged to 1985 prices.

The occupational classification for each job is recorded using the the most recent U.S. census classification of occupations. Throughout the analysis, I follow the approach taken in Acemoglu and Autor (2011) to group occupations into broad groupings largely defined by the census occupational coding. These groups are:

1. Professional, technical, managerial
2. Production, operations, sales, clerical, administrative support
3. Service

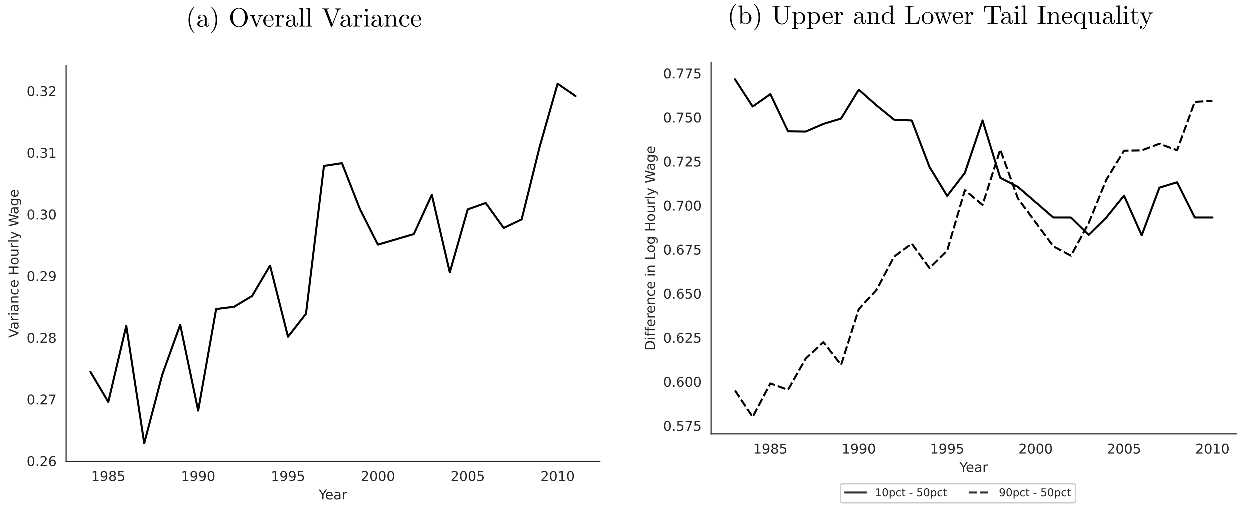
For brevity, I refer to these simply as professional, production, and service occupations throughout the text.

Acemoglu and Autor (2011) argue that this occupational classification broadly captures the primary tasks performed in different jobs, with the first corresponding to jobs with "non-routine cognitive" tasks, the second corresponding to jobs with "routine" tasks, and the third corresponding to jobs with "non-routine manual" tasks. I adopt this framework primarily to be consistent with this theoretical analysis and the subsequent empirical work on job polarization which also use this classification of occupations.

In the analysis that follows, I consider a restricted sample of male workers aged 25-60.[3] Sample sizes for each SIPP panel under this restriction are shown in Column (A) of Table 1. The sample size under the restriction that an individual is employed sometime during the panel is given in column (B). Finally, in the empirical strategy laid out in the rest of the paper, I will further restrict the sample to workers who have complete employment data for three periods. This restriction ensures that there is enough information to identify the underlying skills of the sample of workers and reduces the sample size roughly in half.

Table 2 reports demographic characteristics of the SIPP samples. In line with other standard data source, it shows the workforce became older and more educated during this period. Demographics with the sample restriction to workers with three periods of data are reported in parentheses. The overall levels and trends of these demographics closely follow the employed sample, but there is over-representation of older and more educated workers, which is around 1-2 percentage points in most periods.

Figure 1: Trends in Wage Dispersion



(a) Overall Variance

(b) Upper and Lower Tail Inequality

Stylized Facts

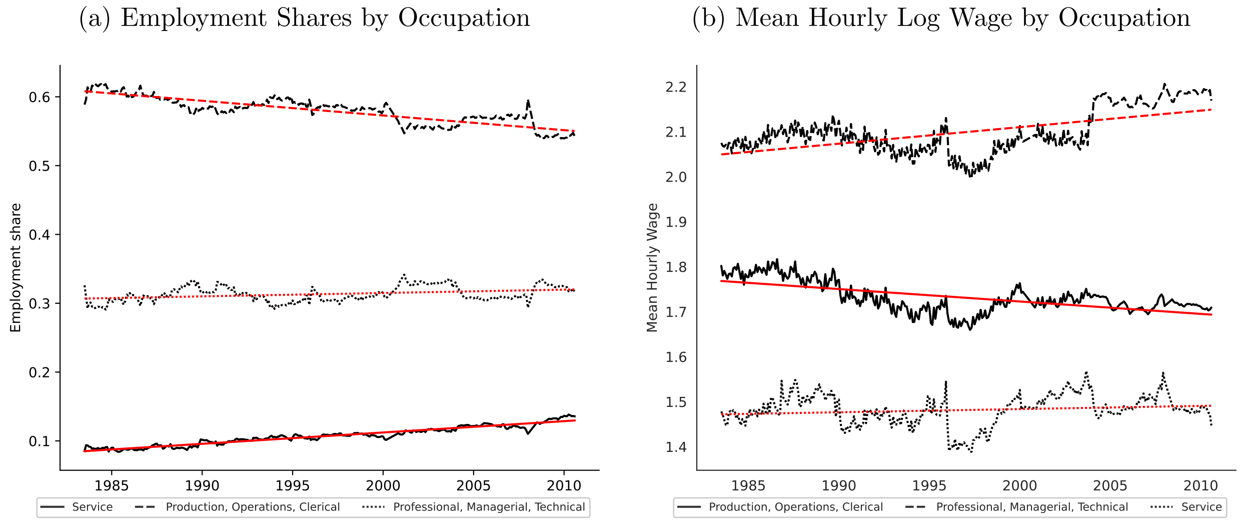
The simplest measure of wage inequality is the variance of observed wages. Panel (a) of Figure 1 shows that overall variance in wages increased during this period. In addition to overall increasing variance, a central

[3] The restriction to male workers is made for several reasons. First, the trends in wage inequality between men and women follow very different patterns during this period and have to be treated separately. Second, endogenous non-participation is an important complication in the empirical strategy I develop in the rest of the paper, so I focus here on the sample with the highest labor force participation. Extending the approach in this paper to account for endogenous non-participation will be an important avenue for further research.

finding in the literature on labor market polarization is that dispersion has increased at the top of the wage distribution, but declined at the bottom. The most common way of illustrating this is to look at changes over time in the difference between the 90th and 50th percentile of the wage distribution, and the difference between the 50th and 10th percentile. These trends are shown in Panel (b) of Figure 1. Similar to the results reported in Autor, Katz, and Kearney (2008), this shows that inequality in the upper tail of the distribution increased over the period, while lower tail inequality declined.

One of the most prominent explanations that has been proposed for the divergent trends in upper and lower tail wage inequality has been that labor demand for middle-wage occupations declined relative to high- and low-wage occupations. Descriptive evidence for this hypothesis is presented in Figure 2. Panel (a) shows the monthly employment shares in each of the three occupations, while panel (b) shows the monthly mean hourly wage in each occupation. Regression lines are shown in red on each panel. Employment in the production occupations declined by 5%, while employment in professional occupations increased by 2% and employment in the service category increased by 3%. These changes in employment combined with the trend of increasing mean wages in professional occupations and decreasing mean wages in professional occupations could have contributed to the changes in inequality observed.

Figure 2: Between Occupation Contributions to Wage Polarization

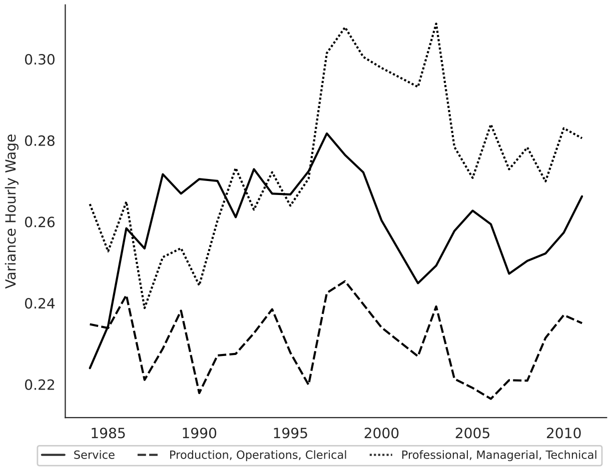

(a) Employment Shares by Occupation

(b) Mean Hourly Log Wage by Occupation

While the theory of RBTC emphasizes the importance of changing inequality between occupations, changes among workers in each occupation played an important role in the growth in inequality and polarization. Figure 3 shows the variance of log wages within each occupational category. Variance increased within the wages of workers in production occupations, while they remained relatively flat in the other two categories. However, considering the changes in lower and upper tail inequality separately in Figure 4 shows that there were significant distributional changes among workers within occupations. Within both the production and service occupations, upper- and lower-tail inequality diverge mirroring the overall pattern in wages.

These patterns of wage polarization within occupations could be explained by changes in skill prices, the distribution of skills, occupational sorting, or some combination of these. To gain some intuition for how these different factors could affect within-occupation wage distributions, consider the following simplified

Figure 3: Within Occupation Wage Dispersion: Variance



version of the model I develop in the next section. Assume that workers, indexed by i, have a scalar skill, $x_i \in \mathbb{R}$, and they can choose to work in one of the three occupations, $\mathcal{D} = \{\text{service}, \text{production}, \text{professional}\}$. In each occupation, $d \in \mathcal{D}$, worker i can produce $\exp(\lambda(d)x)$ units of output per hour, and is paid a rate of $P(d)$ per unit of output. Under these assumptions, the log of the potential hourly wage a worker with skills x can earn by working in occupation d is

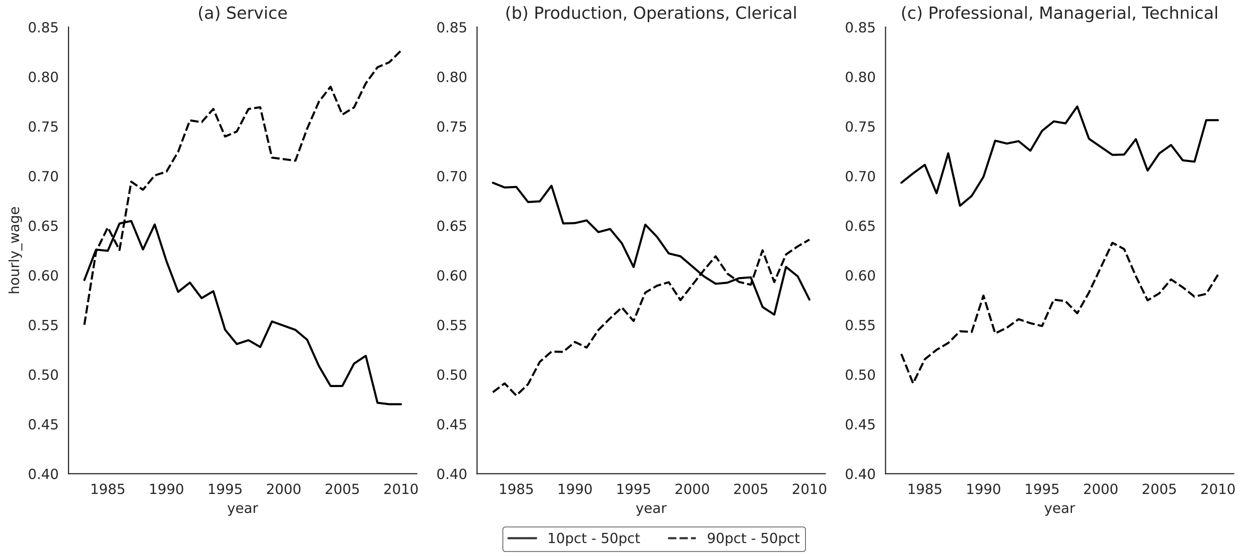
$$y(d, x) = \log(P(d)) + \lambda(d)x$$

In this setting, $P := \{P(d) : d \in \mathcal{D}\}$ are equilibrium market prices, while $\lambda := \{\lambda(d) : d \in \mathcal{D}\}$ are technological parameters representing the productivity of skills. The first thing to notice about this model is that $\log(P(d))$ is additively separable from $\lambda(d)x$, so holding the assignment of occupations fixed, changing the prices P cannot change upper or lower tail inequality in log wages. Similarly, increasing $\lambda(d)$ increases both lower and upper tail inequality. This is because $\frac{\partial(y(d,x')-y(d,x))}{\partial\lambda(d)} = x' - x \geq 0$ for any $x' \geq x$, which holds with x' and x selected to be any two quantiles of the distribution of workers' skills in the population.

In the context of this simplified model, therefore, diverging trends in upper and lower tail inequality within an occupation must be explained by changes in the distribution of skills of workers who select into that occupation. In the analysis of Acemoglu and Autor (2011) and Jung and Mercenier (2014), the only mechanism for skills to change in response to a change in P or λ is for workers to reoptimize their occupational choices to maximize income. However, the trends displayed in Figure 4 are difficult to fully explain with this adjustment mechanism. In the case where the price for the production occupation declines, holding the other two prices fixed, these models predict that the highest and lowest skilled workers in production occupations would switch into service or professional occupations respectively. However, this would decrease both upper and lower tail inequality in the wages of production workers, whereas upper tail inequality increases in the data.

The preceding analysis suggests that in the context of this simplified model, changes in within-occupation distribution of workers' skills are an important part of the story of polarization and are not fully explained by Roy sorting. In the following section, I turn to developing a model which is flexible enough to explain these changes in within-occupation wage inequality. An important motivation in the following sections will be to develop an approach to identifying the roles that changes in skill prices, the distribution of skills, and occupational sorting play in these trends.

Figure 4: Within Occupation Wage Dispersion: Upper and Lower Tail Inequality



3 Model

In this section, I develop a model of wage determination and occupational choice, which I will use to empirically evaluate how the distribution of workers' skills and their sorting patterns have changed over time. The model is sufficiently general to evaluate the role of three factors in the evolution of wage inequality: (1) the distribution of workers' skills, (2) the patterns of worker sorting into occupations, and (3) the prices of skills in each occupation. In order to flexibly evaluate the importance of each of these factors, the model allows for an arbitrary relationship between workers' skills and occupational choices.

Assumptions

The follow assumption describes a general model of wage determination and occupational choice.

Assumption M For the set of workers, I, the following hold,

1. Worker $i \in I$ can produce $g_t(d, x_{it})$ units of output in occupation d in period t, where x_{it} are a set of factors that affect productivity.
2. Worker i derives a utility $u_t(x_{it}, z_{it})$ from working in occupation d in period t where z_{it} are additional factors that affect their decision, and chooses a sequence of occupations $(d_{i1}, \dots, d_{i\bar{t}})$ to maximize their utility in each period.
3. A worker who works in occupation d in period t and produces y units of output is paid a wage of $w_t(d, y)$, where w_t is strictly increasing in y for each choice $d \in D$

Notice that these assumptions have not yet imposed any restrictions on the joint distribution of (x_{it}, z_{it}). Hence this model is consistent with any arbitrary pattern of productivity and occupational choices. For example, assumption M.2 is consistent pure Roy sorting. Assumption M.2 appears to impose that agents choose between all occupations each period, but z_{it} can contain a variable specifying which choices are feasible.

The most restrictive assumption is M.3, which imposes that wages are strictly increasing in a worker's productivity. This is critical to the identification argument I make in the following section, because it allows us to treat wages as a measurement of underlying productivity. This assumption holds most naturally if the labor market is competitive, in which cases the wage schedule is $w_t(d, y) = w_t(d)y$.

The following additional assumptions provide sufficient conditions to identify the model,

Assumption I

1. The factors that affect productivity can be separated into a vector of skills x_i which is constant for $t = 1, \ldots, \bar{t}$, and time-varying productivity shocks, $\{\epsilon_{it}(d) : d \in \mathcal{D}\}$.
2. The productivity shocks, $\{\epsilon_{it}\}$ are serially independent, and independent of skills and preferences $\{x_i, z_{it}\}$.
3. Utilities do not depend on the time-varying productivity shocks, i.e., $u_t(x_{it}, z_{it}) = u_t(x_i, z_{it})$.
4. The productivity function g is multiplicatively separable in ϵ_{it} and x_i, i.e., $g_t(d, x_i, \epsilon_{it}) = h_t(d, x_i)\exp(\epsilon_{it})$

Assumptions I.1 and I.2 are motivated by the permanent income hypothesis. This is one of the starting points for a large literature which attempts to decompose income dynamics into permanent and transitory components (see e.g., Hu, Moffitt, and Sasaki 2019; Shin and Solon 2011; Moffitt and Gottschalk 2012; Gu and Koenker 2017). The canonical variance components model used in that literature decomposes earnings y_{it} as,

$$y_{it} = p_t\mu_i + \epsilon_{it}$$

Assumptions M and I generalize this canonical model in three ways: First, it allows workers to endogenously choose between multiple potential occupations which have different wage schedules; second, it allows the permanent component to be multidimensional; and third, it allows the distribution of the transitory shocks to be different for different occupations.

Recent work decomposing income dynamics into transitory and permanent components have focused on relaxing the canonical model in a different direction. Hu, Moffitt, and Sasaki (2019), for example, allow the "permanent" component to change over time following a random walk and for the ϵ_{it} to follow an ARMA process. These models are non-nested, since assumption I maintains analogues to the canonical assumptions that x_i is fixed over time and $(\epsilon_{it}(0), \ldots, \epsilon_{it}(\bar{d}))$ are independent over time.

Maintaining these assumptions can be justified in this setting for several reasons. First, in contrast to that literature which typically uses a long panel covering much of the life cycle, I focus on short time frame (3-4 years) during which skills can be considered to be approximated fixed. Second, while this literature is agnostic about what the shocks to permanent income are, the model laid out in M provides a mechanism to explain shocks to permanent income through changes in labor prices and productivity across occupations. This model can be understood then as a way of explaining the "black box" shocks to permanent income under the assumption that skills are fixed.

On the other hand, these assumptions do not require that the distribution of skills in the population is fixed over time. Therefore, estimating the model across successive panels can show how the distribution of skills have changed over time through the changing composition of the population.

Assumptions I.2 and I.3 require that transitory productivity shocks do not affect workers' occupational choices. This can be interpreted most naturally as an assumption that these shocks are realized after occupational choices are made. This assumption is common in many areas of labor economics such as analysis of firm and work heterogeneity (e.g. Abowd, Kramarz, and Margolis 1999; Bonhomme, Lamadon, and Manresa 2019; Card, Heining, and Kline 2013) and migration decisions (e.g. Kennan and Walker 2011). This the only restriction on the choice model. The assumption is consistent, for example with a dynamic choice model in which workers maximize future earnings, as well as labor market frictions such as switching costs, and nonpecuniary preferences which can be arbitrarily correlated with skills.

Finally, Assumption I.4 is a functional form assumption, which means that the log productivity is additively separable in the time-varying component. This assumption can be relaxed, but as discussed in section 4, this requires making a higher level rank assumption, which is generally hard to interpret or verify.

Relationship to task-based occupational choice

The model described by assumptions M and I is closely related to the task-based models of occupational choice. Autor and Handel (2013), for example, develop a cross-sectional model of the labor market which is nested by assumptions M and I, by setting

$$h(d, x_i) = \exp(\lambda(d)' x_i)$$
$$p(d, y) = p(d) y$$

This leads to the log-linear potential wage schedule in each occupation as,

$$w(d, x) = \mu(d) + \lambda(d)' x + \epsilon$$

where $\mu(d) = \log(p(d))$. In their model, this productivity function has the following interpretation: workers perform K tasks in one occupation. Each occupation produces a different output which firms combine to make a composite good. The productivity of performing one unit of each task varies across occupations, which is described by $\lambda(d) \in \mathbb{R}^K$. A worker can perform $x_i \in \mathbb{R}^K$ units of each tasks. Hence, a worker with skills x can produce $\lambda(d)' x$ log units of output in occupation d.

An important aspect of this model is that it provides a justification for skills to be priced different across different occupations (see e.g., Heckman and Scheinkman 1987; Autor and Handel 2013). Rosen (1978), Teulings (1995), Costinot and Vogel (2010), and Jung and Mercenier (2014) develop models of labor market sorting which use this as a basis to derive models of sorting across occupations, and explain trends in wage inequality (see e.g., Acemoglu and Autor 2011).

A focus in empirical work has been extending this to a panel-data model and estimating $\mu(d)$ over time, as this can be interpreted as a change in the labor price for each occupation. Under additional assumptions the changes of these equilibrium prices can be interpreted as evidence of changes in labor demand which have driven changes.

Cortes (2016), Gottschalk, Green, and Sand 2015, Böhm (2020), and Cavaglia and Etheridge (2020) provide various approaches to estimating $\mu_t(d)$. The approach in each of these papers, requires additional identifying assumptions. Cortes (2016) assumes that productivity is fixed over time, i.e., $g_t(d, x_i) \equiv g(d, x_i)$, implying that changes in tasks prices can be recovered from taking first differences among workers who stay in the same occupation. Cavaglia and Etheridge (2020) extends this approach by allowing productivity to also depend on time-varying observed variables. Gottschalk, Green, and Sand (2015) exploit an assumption of Roy-style selection in which workers choose the occupation which gives them the highest wage. Böhm (2020) requires non-selected measurements of x_i.

In contrast to these approaches, I do not require that the productivity of skills, g is fixed over time or that workers choose an occupation to maximize income, or that external measurements of skills are available. The identification strategy and estimation method I develop in this paper also diverges from this work by focusing on identification of the full joint distribution of potential wages and choices rather than focusing on task prices.

4 Identification

The model laid out in the previous section has a natural factor structure that can be used to establish point identification. In particular, the low dimensional vector of skills can be viewed as an unobserved latent variable, while observed wages can be viewed as measurements of these unobserved variables, where the time varying shocks are assumed to be independent from the factor. If there were no endogenous selection into occupations, a viable approach to establishing identification would be to verify the conditions of high-level theorems from the non-classical measurement error literature, e.g., Theorem 2.4.2 in Hu (2017), or specializing to the linear factor model of Theorem 2 in Freyberger (2018). However, a complication in using these existing results is that the particular set of wage measurements observed is endogenously selected. In this section, I develop a new identification result which extends the work of these papers to a setting with endogenously selected measurements. I then establish that the assumptions of this theorem are satisfied under the model of occupational choice laid out in the previous section.

Notation

Throughout this section I will use the notation $[n] = \{1, \ldots, n\}$ for $n \in \mathbb{N}$. I define $a^n = (a_1, \ldots, a_n)$ for $n \in \mathbb{N}$

Setup

Consider a vector of potential outcomes of length $\bar{t} \in \mathbb{N}$: $Y(d) = (Y_1(d_1), \ldots, Y_{\bar{t}}(d_{\bar{t}}))$, where $d_t \in [\bar{d}_t]$, $\bar{d}_t \in \mathbb{N}$ for each t. The discrete random variable D with support $[\bar{d}_1] \times \ldots, \times [\bar{d}_{\bar{t}}]$ can be thought of as a vector of endogenous choices or treatments. The realized outcomes and choices are (Y, D), where $Y =$

$(Y_0(D_0), \ldots, Y_{\bar{t}}(D_{\bar{t}}))$ and $D = (D_0, \ldots, D_{\bar{t}})$. In the application, the components of Y are outcomes in different time periods, but t could also index other outcomes such as multiple test scores.

The following is the main identification assumption.

Assumption PO-I

1. Potential outcomes can be written as:

$$Y_t(d_t) = \mu_t(d_t, X) + U_t(d_t)$$

   Where, $U_t = (U_t(0), \ldots, U_t(\bar{d}_t))$ is a random variable on $\mathbb{R}^{\bar{d}_t}$, X is random variable on $\mathcal{X}$, and $\mu_t : [\bar{d}_t] \times \mathcal{X} \mapsto \mathbb{R}$.
2. U_t is independent of (D_t, X) for all $t \in [\bar{t}]$.
3. $\{U_t : t \in [\bar{t}]\}$ are mutually independent.

These assumptions are closely related assumptions M and I in the previous section, which are substantive assumptions about the labor market. Assumption PO-I.1 is implied by assumptions M.1, M.3, and I.1, and I.4, setting potential wages in each outcome $Y_t(d)$. Assumption PO-I.2 is implied by assumptions M.2, I.2, and I.3. Finally, assumption PO-I.3 is implied by assumption I.2.

By themselves, these assumptions are not restrictive, because we have not restricted the domain of $\mathcal{X}$ or the μ_t functions. If $\mathcal{X} = \mathbb{R}^{\prod_t \bar{d}_t}$, and $\mu_t(d, x) = x_t(d)$ and $U_t(d) = 0$ then potential outcomes can be arbitrarily correlated across time and outcomes. In order to get some identifying power, it is necessary to restrict the dimension of $\mathcal{X}$.

The following assumption provides a set of sufficient conditions to point identify the model. For sequences t^n and d^n, I will denote $\mu^{t^n}(d^n, x) = (\mu_{t_1}(d_1, x), \ldots \mu_{t_n}(d_n, x))$:

Assumption PO-R

1. $\mathcal{X} \subseteq \mathbb{R}^{\bar{k}}$ for some $\bar{k} > 2\bar{t}$.
2. For each sequence of periods and choices of length $\bar{k}$, denoted $d^{\bar{k}}$ and $t^{\bar{k}}$ respectively, $\mu^{t^{\bar{k}}}(d^{\bar{k}}, \cdot)$ is continuous and $\mu^{t^{\bar{k}}}(d^{\bar{k}}, \mathcal{X}) = R^{\bar{k}}$.
3. X is continuous and has a density which is strictly positive for almost every $x \in \mathcal{X}$
4. For each (t, d_t), $U_t(d_t)$ is continuous and it's characteristic function is infinitely differentiable and strictly positive almost everywhere.
5. The function $\mu^{t^{\bar{k}}}(d^{\bar{k}}, \cdot)$ is injective for every pair of $\bar{k}$-length sequences $t^{\bar{k}}, d^{\bar{k}}$.

Assumptions PO-R.2 captures the idea that any set of realized outcomes of the same dimension of the latent variable X can be viewed as a measurement of X. Assumption PO-R.1 allows us to then treat outcomes as repeated measurements, by ensuring that there are enough outcomes to form more than two full measurements of $\mathcal{X}$. Assumption PO-R.4 is a regularity condition on the idiosyncratic errors, which is common in the deconvolution literature. It holds for the noise-free case, where $U_t(d) = 0$ and for all of the typical families of distributions such as Gaussian.

Under assumptions PO-I and PO-R, the parameters of the model are the functions $\mu = (\mu_0, \ldots, \mu_{\bar{t}})$, the joint distribution of X and D, F and the distributions of U_t for each t, G_t, with $G = (G_0, \ldots, G_{\bar{t}})$. The identified set, Θ_0, as those (μ, F, G) for which assumptions M and I hold, and for which the distribution of $(D, Y(D, X))$ under (μ, F, G) is equal to the true distribution, $F^0_{D,Y}$

Since the functions μ and the distributions of both X and U_t are unknown, there are two fundamental indeterminacies in the model. The first is the labeling of X. Suppose π is a bijective transformation on $\mathcal{X}$, we can define $\bar{\mu}_t(d, x) = \mu_t(d, \pi^{-1}(x))$ so that $\mu_t(d, x) = \bar{\mu}_t(d, \pi(x))$. The second comes from additivity of the idiosyncratic errors. For any model, there is an equivalent model with $U_t(d) + \alpha$ and $\mu_t(d, x) - \alpha$.

The following theorem says the parameters of this model are identified up to these two indeterminacies.

Theorem 1. For any $(\mu, F, G), (\bar{\mu}, \bar{F}, \bar{G}) \in \Theta_0$, there is a bijective function $\pi : \mathcal{X} \mapsto \mathcal{X}$ and a set of constants, $\alpha_t(d)$ such that:

- $\bar{G}_t(u) = G_t(u - \alpha_t)$
- $\bar{F}(d, x) = F(d, \pi(x))$
- $\bar{\mu}_t(d, x) = \alpha_t(d) + \mu_t(d, \pi^{-1}(x))$

The full proof is given in Appendix A, which I sketch here. The first step in the proof is to show that for each sequence of choices $d = (d_0, \ldots, d_{\bar{t}})$, the conditional distribution of X, $x \mapsto F_{X|D}(x, d)$, and the mean function $x \mapsto (\mu(d_0, x), \ldots \mu(d_{\bar{t}}, x))$ are identified up to the labels of x. This is established by adapting results from the literature on nonclassical measurement error with three measurements (Hu and Schennach 2008). These results provide sufficient conditions to identify the joint distribution of a random vector (Y_1, Y_2, Y_3, X) under the conditions that (Y_1, Y_2, Y_3) is observed $Y_1 \perp Y_2 \perp Y_3 | X$. A key condition in these results is that for $j = 1, 2, 3$, variable X induces a sufficient amount variation in Y_j. More precisely, this requires that the operator, L_j defined by,

$$[L_j g](y) = \int f_{Y_j|X}(y|x) g(x) dx$$

is invertible. This high-level condition, sometimes called *bounded completeness* can be difficult to verify, but D'Haultfoeuille (2011) provides sufficient conditions which exactly match the assumptions of Theorem 1. In particular, the assumptions that $\mu(d, X)$ is additively separable from $U_t(d)$ and that $x \mapsto \mu^t(d^t, x)$ is invertible are sufficient to establish that the invertibility of L_j for each j.

Finally, since identification is established conditionally for each choice sequence, it remains to show that it is possible to choose a consistent labeling of the support of X for the entire joint distribution of (X, D). This established following an argument similar to the proof of Theorem 2 in Freyberger (2018): because for any two choice sequences, we can consider a third which includes elements of both. Each choice sequence must have consistent labeling, so these overlapping choice sequences provide a link between conditional distributions.

Theorem 1 is similar to Theorem 2 in Freyberger (2018) and Theorem 1 in Bonhomme, Lamadon, and Manresa (2019). While Freyberger (2018) uses the results of D'Haultfoeuille (2011) to establish the injectivity condition, he makes the stronger assumption that μ is linear, and does not allow the factor loadings to depend on an endogenous variable. Bonhomme, Lamadon, and Manresa (2019), on the other hand, places fewer restrictions on the functional form of μ and do not require additive separability of $\mu(d,x)$ and $U(d)$, but rely on higher level identification conditions. In particular, they assume directly that the bounded completeness condition holds[4]. To the best of my knowledge, the result in D'Haultfoeuille (2011) is the most flexible set of primitive conditions known to guarantee the injectivity condition in this setting. A valuable extension of this result would be to explore when deviations from additive separability still satisfy this condition.

Sufficient Conditions for Injectivity

Theorem 1 provides non-parametric identification of the function μ with the condition that μ is injective. It is also possible to restrict the function space that μ lies in and derive specific conditions to establish injectivity. One simple choice is to assume that μ is linear:

Assumption PO-R.5′ $\mu_t(d_t,x)=\alpha_t(d_t)+\lambda_t(d_t)'x$, and for any sequence $(t^{\bar{k}},d^{\bar{k}})$ of length $\bar{k}$, the $\bar{k}\times\bar{k}$-dimension matrix $\Lambda(t^{\bar{k}},d^{\bar{k}}):=[\lambda_{t_1}(d_1)\dots\lambda_{t_{\bar{k}}}(d_{\bar{k}})]$ has full rank.

By modeling μ as a linear function, we can establish the required injectivity of μ with this rank condition on λ.

Under assumption PO-R.5', we can also derive a more concrete representation of the mappings, π, between parameters in the identified set. This is given in the following theorem. The theorem will use the notation $\Sigma_{Z|D}(d)=Var(Z|D=d)$ for a random vector $Z=(Z_1,\dots,Z_T)$ and $D=(D_1,\dots,D_T)$,

Theorem 2. Suppose assumptions PO-I, PO-R.1-4 and PO-R.5', then for any $(\Lambda,\Sigma_{X|D}(x|d))$, $(\bar{\Lambda},\bar{\Sigma}_{X|D}(x|d))$ such that,

$$\Sigma_{Y|D}(d)-\Sigma_{U|D}(d)=\Lambda(d)^T\Sigma_{X|D}(d)\Lambda(d)=\bar{\Lambda}(d)^T\bar{\Sigma}_{X|D}(d)\bar{\Lambda}(d)$$

There is a nonsingular matrix M such that for each choice sequence d:

$$\bar{\Lambda}(d)=\Lambda(d)M^{-1} \tag{1}$$

$$\bar{\Sigma}_{X|D}(d)=M\Sigma_{X|D}(d) \tag{2}$$

The rank condition under linearity does impose some constraints. Notably, when $\bar{k}>1$, it rules out the possibility that $\mu_t(d_t,\cdot)=\bar{\mu}(d_t,\cdot)$ for all t. If that were the case, then for the choice sequence $d^{\bar{k}}=(1,\dots,1)$, the matrix $\Lambda([\bar{k}],d^{\bar{k}})$ would have rank 1.

Another limitation of restricting μ to linear functions is that it imposes that the marginal effect of a component of x potential outcomes is constant for every potential outcome. This can be relaxed, by expanding this set to include monotonic transformations of each component of x. That is: $\mu_t(d_t,x)=\alpha_t(d_t,x)=\alpha_t(d_t)+\lambda_t(d_t)'h_t(d_t,x)$, where $h_t(d_t,x)=(h_{t1}(d_t,x_{t1}),\dots,h_{t\bar{k}}(d_t,x_{t\bar{k}}))$, where each component function is a strictly increasing function. These functions are also injective since each h_{tj} function is injective and the rank condition on β holds.

[4] In the statment of the theorem and the proof they also require that unobserved heterogeneity is discrete. In this case, the operator L_j is simply a matrix and bounded completeness is equivalent to invertibility of the matrix. They argue their results can be extended to the continuous case. If this is correct, it would require assuming the invertibility condition directly.

Linearity also rules out the possibility of complementarities between components of the latent variable. There are many common classes of functions, however, that would allow for more complementarity between inputs. Perhaps the most straightforward is a constant elasticity of substitution production function:

$$\mu_t(d_t, x) = \alpha_t(d_t) + \frac{\nu_t(d_t)}{\gamma_t(d_t)} \log\left(\sum_{k=1}^{\bar{k}} \lambda_{tk}(d_t)\exp(\gamma_t(d_t)x_k)\right)$$

This is a composition of strictly increasing functions, so the rank condition on λ is still sufficient to establish the injectivity of μ.

5 Estimation

In this section, I describe the specification of the model I will estimate and the estimation procedure employed. Because of the goal of understanding the changes in workers' skills and selection into occupations, I focus on estimating the joint distribution of skills and choices nonparametrically. In order to make this feasible, I impose several simplifying assumptions on the other aspects of the model. First, I assume that skills are scalar-valued. While this assumption is restrictive, it is commonly made in the theory of wage polarization (e.g., Jung and Mercenier 2014; Acemoglu and Autor 2011), and I show that a scalar-value model fits the data well. Second, while the distribution of $U_t(d)$ is identified nonparametrically, I assume that $U_t(d)$ have a Gaussian distribution. I also maintain the assumption that potential outcomes are linear as is common in the task framework.

Under these assumptions, it is possible to extend a semiparametric estimation methods used to estimate mixture models to this application. In a simple mixture model, likelihood contributions can be written as:

$$\ell(y) = \int g(y, x)dF(x),$$

where the distribution F is estimated nonparametrically. Several recent papers (e.g., Koenker and Mizera 2014; Fox, Kim, and Yang 2016; Chernozhukov et al. 2013) have proposed estimating F by searching over discrete distributions with a fixed (but potentially large) grid of support points. This fixed-grid estimator can be viewed as a particular sieve estimator (Chen 2007), and is motivated by a result that with a sample size of n, there exists a finite distribution with at most n points of support that maximizes the likelihood function (Lindsay 1995). In practice Koenker and Mizera (2014) show that this estimator can be computed efficiently and has good performance in simulations. Fox, Kim, and Yang (2016) and Chernozhukov et al. (2013) show that the estimator is consistent in two contexts.

In this setting, unobserved skills X can depend arbitrarily on the endogenous choice variables $D = (d_1, \ldots, d_T)$. The component mixture distribution g is also parameterized by a finite-dimensional vector of parameters, θ, which has a non-trivial dimension in several of the model specifications I estimate.

I therefore propose several refinements to the implementation of the fixed-grid estimator to allow for efficient computation in this context. First, I propose an accelerated expectation maximization (EM) algorithm which exploits the computational properties of the fixed-grid estimator and the linearity of the outcome equation. The algorithm reduces the optimization problem to alternately solving two simpler problems. Next, I use the linearity of the outcome equation to derive a consistent first-stage estimator of the finite parameters, and show that the full MLE estimator converges to the global minimum when using this first-stage estimator for a starting value of the finite parameters.

Likelihood function.

Under the assumptions PO-I, and PO-R.1-4, PO-R.5$'$, the potential wage in occupation $d_t \in [\bar{d}_t]$ in period $t \in [\bar{t}]$ is,

$$Y_t(d_t) = \alpha_t(d_t) + \lambda_t(d_t)X + U_t(d_t) \tag{3}$$

I also assume that $U_t(d_t)$ is a mean zero Gaussian random variable with variance $\sigma_t^2(d_t)$. The finite-dimensional parameters of the model to be estimated are:

$$\theta = \{\alpha_t(d_t), \lambda_t(d_t), \sigma_t^2(d_t) : t \in [\bar{t}], d_t \in [\bar{d}_t]\}$$

Together with the joint distribution of (X, D), $(\theta, F_{X,D})$ fully specifies the model.

Under these assumptions, the likelihood contribution for an observation with outcomes $Y = (y_1, \ldots, y_{\bar{t}})$ and choices $D = (d_1, \ldots, d_{\bar{t}})$ is,

$$\ell(y, d; \theta, F_{X,D}) = f_D(d) \int \prod_t \frac{1}{\sigma_t(d_t)} \phi\left(\frac{y_t - \alpha_t(d_t) - \lambda_t(d_t)^T x}{\sigma_t(d_t)}\right) dF_{X|D}(x|d)$$

where $f_D(d) = P(D = d)$, $F_{X|D}$ is the distribution of X conditional on D, and ϕ is the distribution of a standard normal random variable. Note that by writing the likelihood function with the conditional mixing distribution $F_{X|D}$, the nonparametric conditional choice probabilities $P(D = d|X = x)$ do not need to be estimated directly.

Profile Likelihood.

Maximizing the likelihood function, $L_n(\theta, F_{X|D}) = \sum_i \log \ell(y_i, d_i; \theta, F_{X,D})$, can be broken down into two steps. In the first step, the likelihood is maximized over $F_{X,D}$ for a fixed value of θ. A computationally feasible approach to estimating $F_{X,D}$ is to use the fixed-grid sieve estimator. Using this sieve space, this inner maximization problem can be written as a convex optimization problem which can be solved efficiently using modern convex optimization solvers.

The fixed-grid sieve space is defined as follows: for each n, let $(x_{n1}, \ldots, x_{nR(n)}) := \mathcal{X}_n \subset \mathcal{X}$ be a finite subset of the support of X with $R(n)$ points, where $R(n) \to \infty$. The sieve space, then is defined as the set of conditional distributions, which each have support:

$$\mathcal{F}_{X|D}^{(n)} = \left\{ \sum_{\bar{d} \in \mathcal{D}} \sum_{r \in [R(n)]} a_{rd} \chi_{(-\infty, x_r] \times \{\bar{d}\}} : \{a_{rd}, r \in [R(n)], d \in \mathcal{D}\} \in \Delta(|\mathcal{X}_n|) \right\}$$

Where $\Delta(A)$ is the set of probability distributions over the finite set of support points A. Given $(\theta, \{\mathcal{X}_n : n \in \mathbb{N}\})$, we can define,

$$g_{idr}^{(n)}(\theta) = \mathbf{1}(d = d_i) \left(\prod_t \frac{1}{\sigma_t(d_t)} \phi\left(\frac{y_{it} - \alpha_t(d_{it}) - \lambda_t(d_{it})^T x_{nr}}{\sigma_t(d_t)} \right) \right)$$

Fixing θ, this maximization problem over $\{a_{rd}, r \in [R(n)], d \in \mathcal{D}\} \in \Delta(|\mathcal{X}_n|)\}$ can be written as follows,

$$\begin{aligned} \text{maximize}_a \quad & \sum_i \log(\ell_i) \\ \text{subject to} \quad & \sum_d \sum_r g_{idr}^{(n)}(\theta) a_{rd} = \ell_i \text{ for each } i \in [n] \\ & \sum_r a_{rd} = 1 \text{ for each } d \in \mathcal{D} \end{aligned}$$

This is a well-understood convex optimization problem which can be written in terms of linear and exponential cone constraints, and can be solved very efficiently with an interior point solver. In practice, I use Mosek's interior point solver. In all the specifications I estimate in this paper, this subproblem can be solved in less than one second.

Accelerated EM algorithm

Writing the solution of the maximization problem above as a function of θ,

$$\theta \mapsto \operatorname*{argmax}_{F_{X,D}} L_n(\theta, F_{X,D}),$$

defines the profile likelihood function. One approach to solving the full MLE problem would be to simply search over the parameter space for θ to maximize the profile likelihood function. In this application, I propose an accelerated expectation maximization (EM) algorithm which takes advantage of the linearity of the outcome equation. This allows for efficient computation and bypasses the need to calculate the derivative of the profile likelihood function with respect to θ.

Let the following sequence (f^m, θ^m) be defined recursively as follows,

$$\begin{aligned} \theta^{m+1} &= \operatorname*{argmax}_{\theta \in \Theta} \sum_i \sum_t \sum_r q_{ir}^m \log g_t(y_i, d_i | x_r; \theta^m) \\ f^m &= \operatorname*{argmax}_{f \in \mathcal{F}_{X|D}^{(n)}} \sum_i \sum_r f(x_r, d_i) \prod_t g_t(y_{it}, d_{it} | x_r; \theta^m) \end{aligned}$$

where,

$$q_{ir}^m = \frac{f^m(x_r, d_i) g(y_i, d_i | x_r; \theta^m)}{\sum_r f^m(x_r, d_i) g(y_i, d_i | x_r; \theta^m)}$$

Further let (θ_n^m, f_n^m), which follows the recursive definition above for each n. The following theorem establishes that this sequence converges to the global maximum of the likelihood function.

Theorem 3. For any such sequence (θ^m, f^m) converge to a local maximizer or $L_n(\theta, f)$. Furthermore, if θ_n^0 is a consistent estimator for θ_0, then

$$\lim_{m \to \infty} (\theta_n^m, f_n^m) = \operatorname*{argmax}_{\theta, f} L_n(\theta, f)$$

with probability approaching 1.

The proof of Theorem 3 is given in Appendix A. This algorithm is closely related to the standard EM algorithm. In the EM algorithm, f^m maximizes $E_n(\log[g(y_i, d_i|x; \theta)f(x)]|y_i, d_i; \phi^{m-1})$. Instead, f^{m+1} maximizes the likelihood function itself given θ^{m+1}. Dempster, Laird, and Rubin (1977) show that the former is sufficient to increase the likelihood at each step. Since this choice of f^m is at least as good as the EM step, this is sufficient to show this sequence will converge to a local minimum. Since it improves the likelihood more at each step and is the global maximum given θ, we should expect significantly faster convergence.

The model specifications I estimate have the following functional form,

$$g_t(y_{it}, d_{it}|x_r; \theta) = \frac{1}{2\sigma_t(d_{it})}\left(\frac{y_{it} - \lambda_t(d_{it})^T x_r}{\sigma_t(d_{it})}\right)^2 \tag{4}$$

With this functional form, the θ update step has a particularly convenient form. To state the following result, I will use the following notation,

$$\begin{aligned}
\bar{q}^m_{td_tr} &= \sum_i \mathbf{1}(d_{it} = d_t) q^m_{ir} \\
\bar{y}^m_{td_tr} &= \sum_i \frac{q^m_{ir}}{\bar{q}^m_{td_tr}} y_{it} \\
\sigma^2_{td_tr} &= \sum_i \frac{q^m_{ir}}{\bar{q}^m_{td_tr}} (y_{it} - x_r'\lambda_t(d_t))^2
\end{aligned}$$

Corollary. When g_t is defined as in (4),

$$\begin{aligned}
\lambda^{m+1}_t(d_t) &= \left(\sum_r \bar{q}^m_{td_tr} x_r x_r^T\right)^{-1} \sum_r \bar{q}^m_{td_tr} \bar{y}^m_{td_tr} x_r \\
\sigma_t(d_t) &= \sum_r \bar{q}^m_{td_tr} \sigma^2_{td_tr}(d_t)
\end{aligned}$$

First step estimator

An important aspect of Theorem 3 is that it guarantees that the estimator will converge to the global maximum if we have a consistent first stage estimator for θ. To get this estimator, can use an approach suggested in Heckman and Scheinkman (1987) and stated formally in Freyberger (2018) to derive a consistent first stage estimator of θ.

Assume that $X \in \mathbb{R}^k$. Partitioning $[T]$ into (t^1, t^2, t^3), where t^1 and t^2 are of length k and t_3 is of length $T - 2k$, then for a sequence d, partitioned in the same way.

$$\begin{aligned}
Y^{t_1} &= \alpha_1(d^1) + \Lambda_1(d_1)X + U_1 \\
&= \big(\alpha_1(d^1) - \Lambda_1(d^1)\Lambda_2^{-1}(d^2)\alpha_2\big) + \Lambda_1(d^1)\Lambda_2^{-1}(d^2)Y^{t_2} + \big(U_1 - \Lambda_1(d^1)\Lambda_2^{-1}(d^2)U_2\big),
\end{aligned}$$

where $\Lambda_1(d^1) = (\lambda_{t^1_1}(d_1), \ldots, \lambda_{t^1_k}(d_k))^T$

Using Y_3 as an instrument for Y_2, we get,

$$\mathrm{Cov}(Y^{t_1}, Y^{t_3}|D = d) = \Lambda_1\Lambda_2^{-1}\mathrm{Cov}(Y^{t_2}, Y^{t_3}|D = d) \tag{5}$$

If every $k \times k$ submatrix of Λ is invertible, then there will be one of these moment conditions for every combination of the partitions of $[T]$ and choice sequences d. The following theorem provides tractable way of combining these moment conditions which maintains the linearity of the estimator:

Proposition 1. Let

$$M_n = \begin{bmatrix} m_n^{(1)}\phi_1^{(1)} - \phi_0^{(1)} \\ \vdots \\ m_n^{(Q)}\phi_1^{(Q)} - \phi_0^{(Q)} \end{bmatrix}$$

where $Q = |\mathcal{D}|\binom{T}{k}\binom{T-k}{k}$, m_n^j is a 2SLS estimand based on the moment condition in (5), $\phi_1^{(j)}, \phi_2^{(j)}$ are the $(k \times |\mathcal{D}_t|T)$ and $(1 \times |\mathcal{D}_t|T)$ matrices defined in appendix B.

For any positive definite weighting matrix W, the GMM estimator $\hat{\lambda}_n := \min_{\mathit{\Lambda}} \mathrm{vec}(\mathit{\Lambda})^T (I_k \otimes M_n)^T W (I_k \otimes M_n)\mathrm{vec}(\mathit{\Lambda}) \to^p \mathrm{vec}(\mathit{\Lambda}_0)$

Proof. See Appendix A □

The GMM estimator proposed in this proposition is less efficient than the full GMM estimator of λ, but it still uses all the moment conditions of the form (5). Because of the large number of potential choice sequences and different ways to partition the periods, this results in a large number of moment conditions. The simplification of stacking 2SLS estimands maintains the linearity of the minimization problem, so the solution is simply the solution to a linear least squares problem.

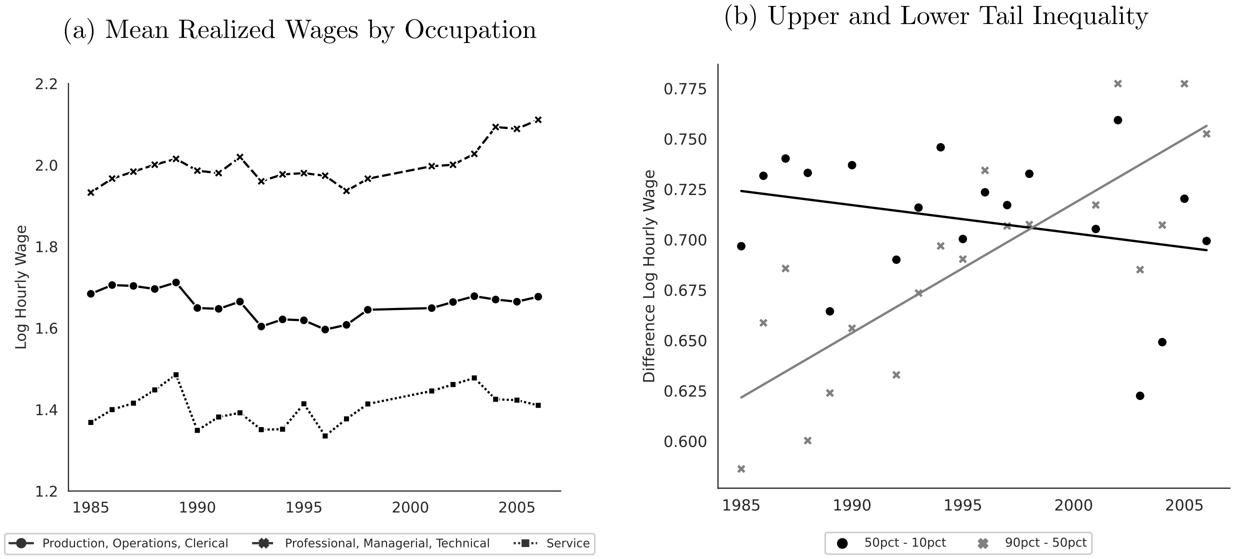
6 Results

In this section, I present the results from estimating the model. I begin by showing that the model fits the cross-sectional trends reviewed in section 2 as well as the joint distribution of individuals' wages over time. I next turn to exploring the role of endogenous sorting into occupations, and show that non-wage factors appear to play a large role in workers' occupational choices. I then explore how the patterns of occupational sorting have contributed to changes in wage inequality since the 1980s. I find that sorting across occupations by skill level has led to increasing wage inequality compared to the situation where workers are randomly allocated to occupations. Finally, I analyze the contribution of temporal changes in the distribution of workers' skills to the trends in wage inequality, and show that changes in the skill distribution among workers who select into service occupations account for the largest share of the changes to wage inequality.

Model Fit

A parsimonious specification with univariate skills and independent, Gaussian productivity shocks captures the main the cross-sectional trends in wage inequality outlined in section 2 and key patterns in the joint distribution of individual workers' wages over several periods. Beginning with the cross-sectional trends, Figure 5 shows the estimates of my model's parameters fit the trends reviewed in section 2. Panel (a) of Figure 5 shows that the mean log hourly wages by occupation follow the pattern shown in Panel (a) of Figure 2. As in the raw data, there is a slight pattern of polarization in the mean wages, with the high skill and low skill occupations increasing slightly, while the middle-skill occupation declined. Similarly, Panel (b) of Figure 5 shows that the trends observed in upper and lower tail inequality are captured in the fitted model. Under the estimation results, lower tail inequality decreases somewhat while upper tail inequality increases.

Figure 5: Estimated Trends in Inequality


(a) Mean Realized Wages by Occupation

(b) Upper and Lower Tail Inequality

Next, I explore how well the univariate skill model fits the observed patterns of individual workers' sequences of wages and occupational choices. Under assumptions PO-M and PO-I, each observation of an individual's wage can be viewed as a measurement of their skills. For a worker i who chooses occupation D_{it} in period t, we can define a measurement of individual i's skills as follows,

$$\bar{Y}_{it} := \frac{Y_{it} - \alpha_t(D_{it})}{\lambda_t(D_{it})} = X_i + \frac{U_{it}(D_{it})}{\lambda_t(D_{it})}$$

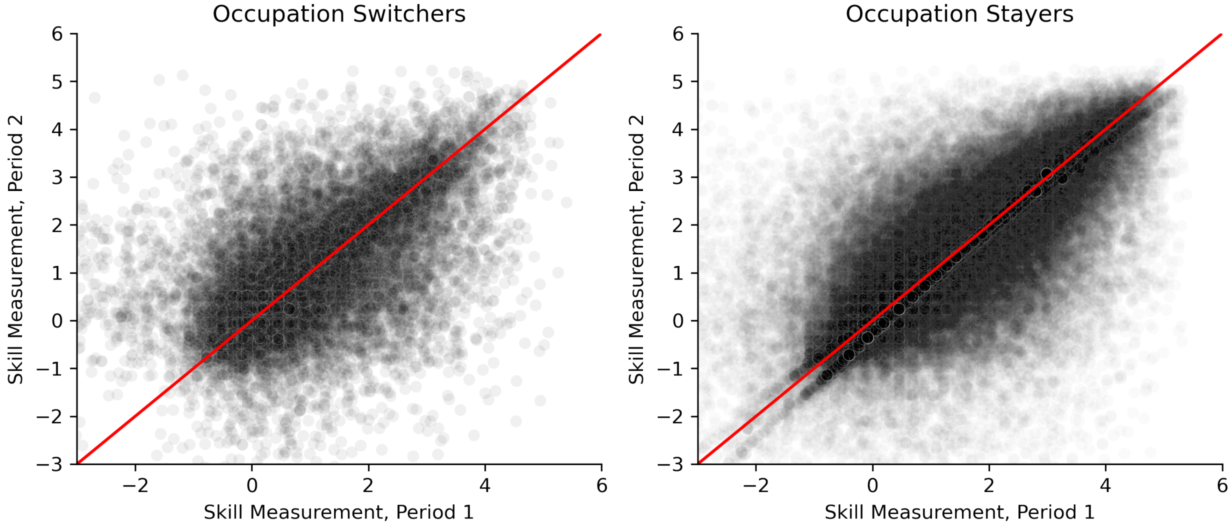
Now consider the distribution of a pair of skill measurements for worker i conditional on choosing the occupational sequence $D_i := (D_{i1}, D_{i2}) = (d_1, d_2)$. This can be written as,

$$\bar{Y}_i := \begin{bmatrix} \bar{Y}_{i1} \\ \bar{Y}_{i2} \end{bmatrix} = \begin{bmatrix} X_i \\ X_i \end{bmatrix} + \begin{bmatrix} U_{i1}(d_1)/\lambda_1(d_1) \\ U_{i2}(d_2)/\lambda_2(d_2) \end{bmatrix}$$

Notice that under assumption PO-I.2, these $\bar{Y}_i$ is the sum of two terms that are independent conditional on D_i. The first is perfectly correlated between period 1 and period 2 and the second is spherical noise centered at the origin. If the variance of U_i is small relative to the variance of X_i, then the bivariate distribution of $\bar{Y}_i$ should be clustered along the 45 degree line with symmetrical deviations around that line.

Figure 6, shows the bivariate relationship for all pairs of rescaled outcomes, $(\bar{Y}_t, \bar{Y}_s)$ for all individuals included in the sample for all survey years estimated. In order to evaluate whether the single factor model accurately captures the transferability of skills across occupations, the relationship is plotted separately for pairs of measurements that involve a change of occupation and those that do not. The figure shows that measurement pairs are clustered around the 45 degree line, consistent the model predictions.

Figure 6: Bivariate Relationship Between Pairs of Skill Measurements



These predictions do not hold when X_i is not univariate. In particular, when X_i is multidimensional, there is generally not a linear association between the pair of skill measurements. The association between pairs of skill measurements also varies depending on the choice sequences when X_i is multidimensional.

To see why, consider an alternative case where X_i is bivariate. A measurement of the form above for any arbitrary choices of $\ell := (\ell_1, \ell_2)$, conditional on the choice sequence $(D_{i1}, D_{i2}) = (d_1, d_2)$ can be written as,

$$\bar{Y}_i(\ell) := \begin{bmatrix} \bar{Y}_{i1}(\ell_1) \\ \bar{Y}_{i2}(\ell_2) \end{bmatrix} = \begin{bmatrix} \lambda_{11}(d_1)/\ell_1 \\ \lambda_{21}(d_1)/\ell_2 \end{bmatrix} X_{1i} + \begin{bmatrix} \lambda_{12}(d_2)/\ell_1 \\ \lambda_{22}(d_2)/\ell_2 \end{bmatrix} X_{2i} + \begin{bmatrix} U_{i1}(d_1)/\ell_1 \\ U_{i2}(d_2)/\ell_2 \end{bmatrix}$$

From this expression we can see that if X_{i1} and X_{i2} are not independent from each other or perfectly correlated, there would generally be a nonlinear association between $\bar{Y}_{i1}(\ell_1)$ and $\bar{Y}_{i2}(\ell_2)$. Consider, for example, setting $\ell_1 = \lambda_{11}(d_1)$, $\ell_2 = \lambda_{21}(d_2)$. As in the scalar case, the term associated with X_{1i} would be perfectly correlated between period 1 and 2, and the term associated with U_i would be spherical noise centered at the original. However, unless X_{2i} is independent from X_{1i}, the mean of the term associated with X_{2i} could vary with X_{1i}, inducing a nonlinear deviation from the 45 degree line in the association between $\bar{Y}_{1i}$ and $\bar{Y}_{2i}$. Notice also that the association between X_{1i} and X_{2i} would not generally be independent of the choice sequence. Therefore, the association between pairs of skill measurements could vary with the choice sequence.

While the evidence presented here is not a formal specification test, the linearity of the association between $\bar{Y}_t$ and $\bar{Y}_2$, and the similarity in the results for occupation switchers and occupation stayers gives some confidence that this model captures the main features of the data.

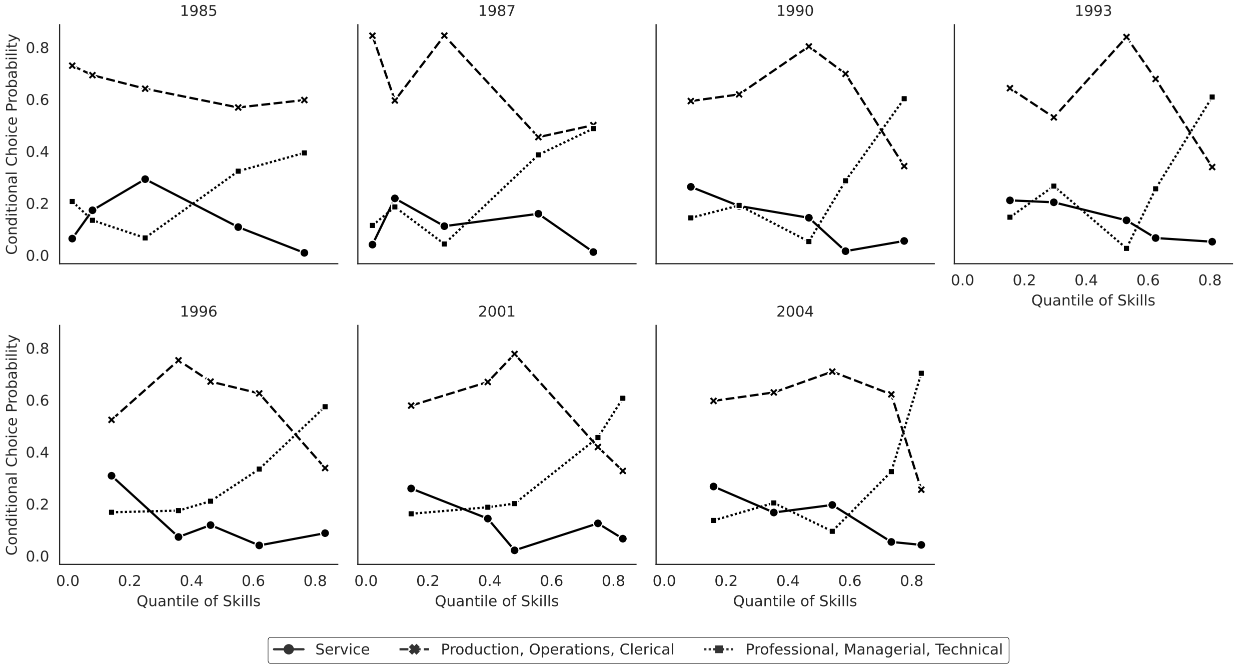
Endogenous sorting patterns

Turning to the main results, I first consider how workers sort into occupations. Since skills are univariate in the estimated model, choice probabilities can be viewed on a continuum from low to high skill workers.

Figure 7 shows the probabilities that workers at a given quantile of the skill distribution will choose each of the occupations. Each panel shows the choice probabilities for one SIPP panel on the year that panel began. The approach of comparing successive panels provides a way to explore the evolution of population-level sorting patterns over time.

Across panel years, there is a consistent pattern of selection into occupations by skill level. The probability of selecting service occupations declines at higher quantiles in the skill distribution, while the probability of choosing professional occupations increases. Production occupations have a higher overall probability of selection, but peak toward the middle of the skill distribution. There is a trend for workers to increasingly sort by skill level. This is most clear in the selection into professional occupationss: workers at the top of the skill distribution increasingly select professional occupations toward the end of the sample.

Figure 7: Occupational Choice Probabilities conditional on quantile of skills



This occupational sorting pattern could be consistent with selection based on comparative advantage, as implied by the pure Roy model. Jung and Mercenier (2014) provide a theoretical analysis under which the equilibrium wage schedule could accommodate this sorting pattern with pure Roy sorting even when skills are univariate. In particular, this sorting pattern can arise if there is a negative relationship between the intercepts (α) and the coefficients on skills (λ) in the occupation-specific log wage equations given in (3). The necessity of this negative association was noted in Autor and Handel (2013) and elaborated in Jung and Mercenier (2014) for the case where unobserved skills are scalar-valued.

The estimation results however, suggest that many workers do not choose an occupation to maximize their income as implied by a pure Roy model. In order to provide a quantitative assessment of the wage premia that workers forgo, I calculate the potential gains in wages that workers could get from switching occupations.

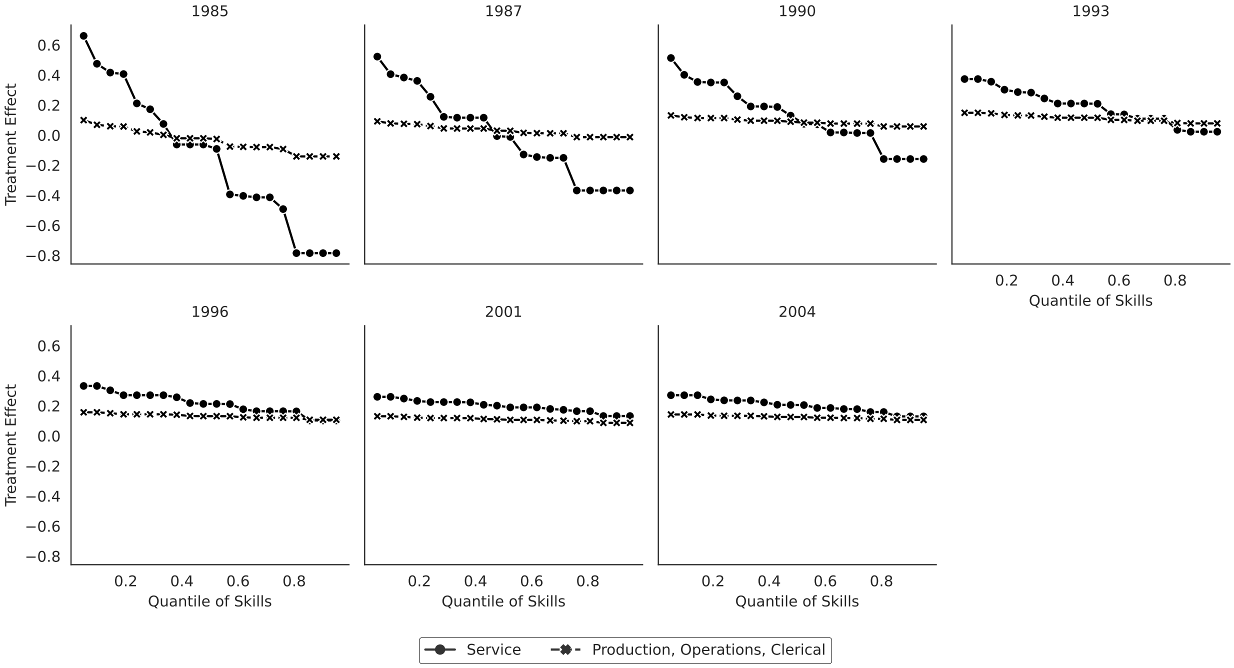
In particular, for each quantile of the distribution of skilles, X, I calculate the difference in potential log wages for each pair of occupations. To state this formally, for a random variable, Z, let the α-quantile be defined as, $Q_\alpha(Z) = \inf\{q : P(Z < q) \leq \alpha\}$. We can then define the following parameter,

$$\varDelta_{\alpha,\alpha'}(d, d') = E(Y_t(d) - Y_t(d')|Q_\alpha(X) < X \leq Q_{\alpha'}(X))$$

This is the average gain in log wages that a worker with skills between the α and α' quantiles of the skill distribution, in occupation d' would get by switching to occupation d. These parameters describe the comparative advantage of workers at different skill levels. A worker with skills X has a comparative advantage in occupation d if he has a positive potential wage gain by switching to occupation d from any other occupation. These parameters can be calculated from the estimation results, because of the flexible approach to recovering the skill distribution.

Figure 9 presents the estimated potential wage gain of switching to a professional occupation from either of the two alternatives at different quantiles in the skill distribution. One important feature of the results is that the treatment effects of switching from either service or production occupations to professional occupations is larger at lower skill levels. Recall that the results presented in Figure 7 showed that the probability of choosing a professional occupation increased with skill level under the model estimates. Given that the wage premium for choosing this occupation is actually smaller at higher skill levels, this indicates that occupational choice is not generated by a Pure Roy model of earnings maximization.

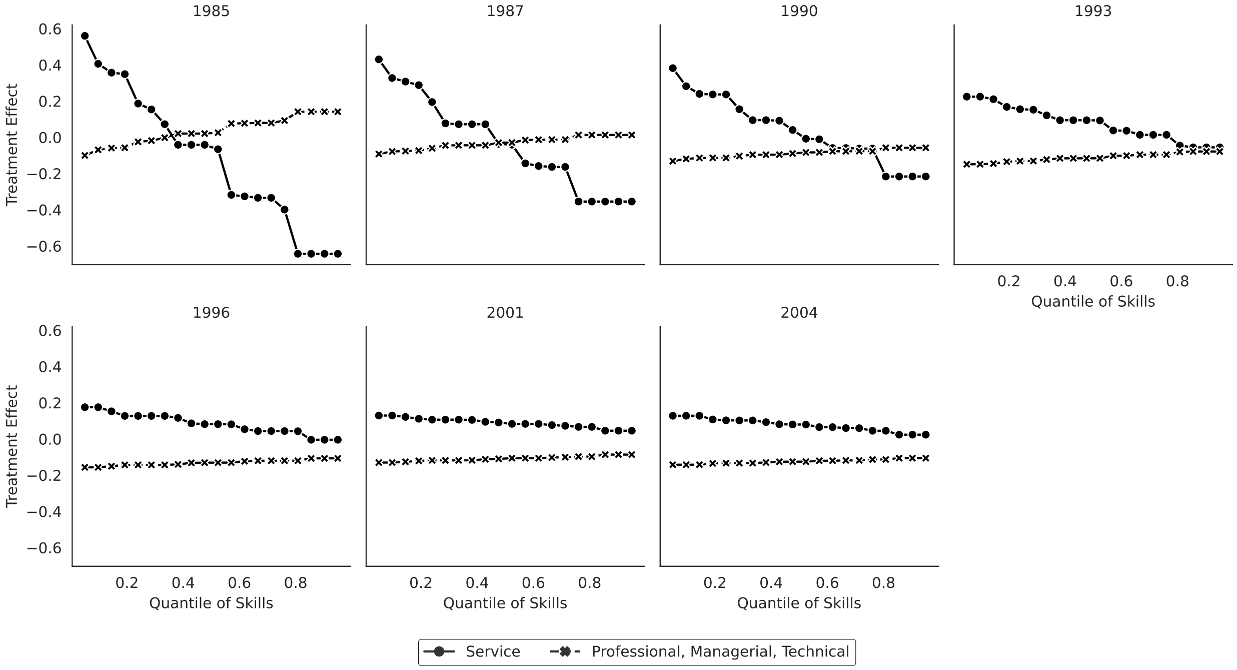
Figure 8: Potential Wage Gains of Switching to Professional Occupations from either Service or Production Occupations



The evolution of potential wage gains for workers who switch out of service occupations points to notable change in the structure of this sector. At the beginning of the period, there was much more heterogeneity in the potential wage gains for service workers, including some workers who earned more in service occupations than they would in professional occupations. Over time, however, these positive wage premia for service occupations disappeared, until at the end of the period, all service workers would have earned more in the professional occupation. This pattern could reflect a homogenizing effect of technology. In early periods high skill workers may have been relatively more productive at specialized service tasks. With the adoption of new technologies that simplified service tasks, there may be a smaller role for a worker's skills to increase their productivity in these tasks.

Wage premia for production occupations relative to the other two occupations are presented in Figure 9. These results also suggest that non-wage factors played a large role in the choice to work in production occupations. In the later periods, starting in 1996, treatment effects are positive across all skill levels for service occupations and negative for professional occupations, suggesting that workers at all skill levels in production and service occupations could have increased their wage by changing occupations. Similar to the analysis of professional occupations, the premium for professional occupations also seems to decline at higher skill levels even though high-skill workers are more likely to choose this occupation. Before 1996, two features are of note. First, at high skill levels, there is a very slightly positive wage premium for production occupations relative to professional occupations in some early periods. As in the analysis above, this disappears over time and is consistent with the hypothesis about the homogenizing effect of technology.

Figure 9: Treatment Effect conditional on quantile of skills: Switching to "Routine" Occupation



Occupational Sorting and Wage Inequality

I next turn to the question of how sorting into occupations affected wage inequality. The results from the previous section showed that workers in service and production occupations often forgo potential wage gains they could get from switching to the professional occupation. Since these workers also tend to have a lower skill level than workers who choose professional occupations, this sorting pattern could potentially exacerbate the wage inequality arising from the skill distribution alone.

In order to investigate this and quantify the contribution of sorting to inequality, I compare the sorting pattern estimated in the model against counterfactual distribution in which workers are randomly assigned to occupations. This comparison helps to assess the effect of the estimated sorting pattern of the wage distribution compared to a neutral sorting pattern in which skills are not related to occupational choice.

For each panel of the SIPP, I have estimated the joint distribution of (X, D), $\hat{F}_{X,D}$. From this, we can obtain the estimated marginal distribution of X,

$$\hat{F}_X(x) = \sum_d \hat{F}_{X,D}(x, d).$$

The counterfactual distribution with random assignment to occupations can be constructed as follows,

$$\hat{F}_X^{RA}(x, d) := \hat{F}_X(x)\hat{P}(D = d).$$

I then calculate the counterfactual variance for each of $\hat{F}_X, \hat{F}_X^{RA}$,

$$\int y(d, x)^2 dF(x, d) - \left(\int y(d, x) dF(x, d) \right)^2$$

for each of $\hat{F}_X, \hat{F}_X^{RA}$. The standard deviations under the two distributions are, respectively, $\hat{\sigma}_Y$ and $\hat{\sigma}_Y^{RA}$.

The results of this analysis are presented in Table 3. At the beginning of the period, in 1985, sorting into occupations resulted in lower variance in wages than there would be under random assignment. Interestingly, the result in the initial period, is similar to the estimate in Heckman and Sedlacek (1985) showing that sorting reduced the standard deviation of wages by 5.8%. However, this pattern reverses over the course of the sample period, with sorting increasing the standard deviation of wages 4.12%.

Table 3: Contribution of Sorting to Wage Inequality: Standard Deviation of Log Wages under Estimated and Counterfactual Distributions

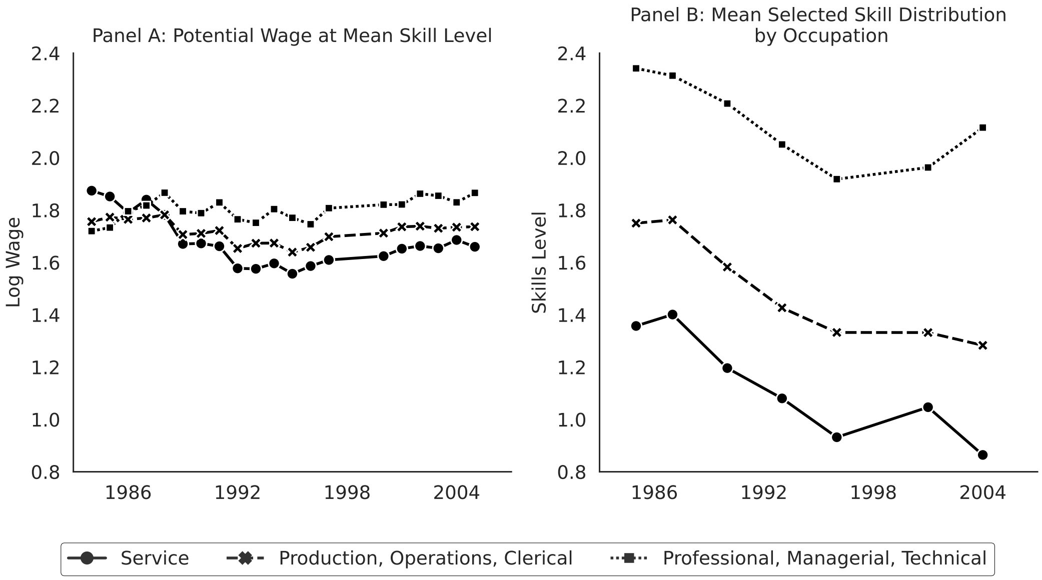
	1985	1987	1990	1993	1996	2001	2004
$\hat{\sigma}_Y$	0.49	0.56	0.53	0.52	0.56	0.55	0.51
$\hat{\sigma}_Y^{RA}$	0.52	0.56	0.51	0.51	0.55	0.54	0.49
Difference	-5.80 %	-1.55 %	1.87 %	4.15 %	2.97 %	1.67 %	4.12 %

Decomposing Trends in Wage Inequality

The comparison to randomly allocating workers to occupations illustrates the importance of occupational sorting to wage inequality. However, this does not directly answer the question of what changes during the period of study led to the increase wage inequality. In the framework of this model, there are three factors that can explain changes in wage inequality: (1) changes in the wage schedules for each occupation, (2) changes in the way workers sort into each occupation by skill level, and (3) changes in the marginal distribution of workers' skills. In this section, I present a several counterfactual wage series derived from the model estimates which isolate the contributions of change in each of these components to change in wage inequality.

I begin by examining the evidence that changes in the relative prices of labor in occupations led to increasing wage inequality. In order to isolate changes in labor prices from changes in the skills of workers who select into occupations, I compare the potential log wages of a worker in each occupation at a fixed skill level. Panel (A) of Figure 10, shows how the potential log wage of a worker at the mean skill level for each period evolved over time. The differences between the log potential wages in each occupation show how much wage inequality is induced by occupational choice compared to a situation in which skill prices are equal across occupations. To put the scale of these differences in perspective, recall that the difference between the mean observed log wages in professional and services occupations displayed in Panel (A) of Figure 2 is on the order of .6 log points. As displayed in Panel (A) of Figure 10, the overall magnitude of differences in potential log wages between occupations holding skills fixed at the mean skill level are on the order of .1 to .2 log points. Differences in log wages between the three occupations increased over the late 1980s and early 1990s and are generally stable afterward. On the other hand, there is limited evidence that the prices of production occupations fell relative to service occupations, as predicted by the theory of routine biased technological change.

Figure 10: Trends in Potential Wages and Skills



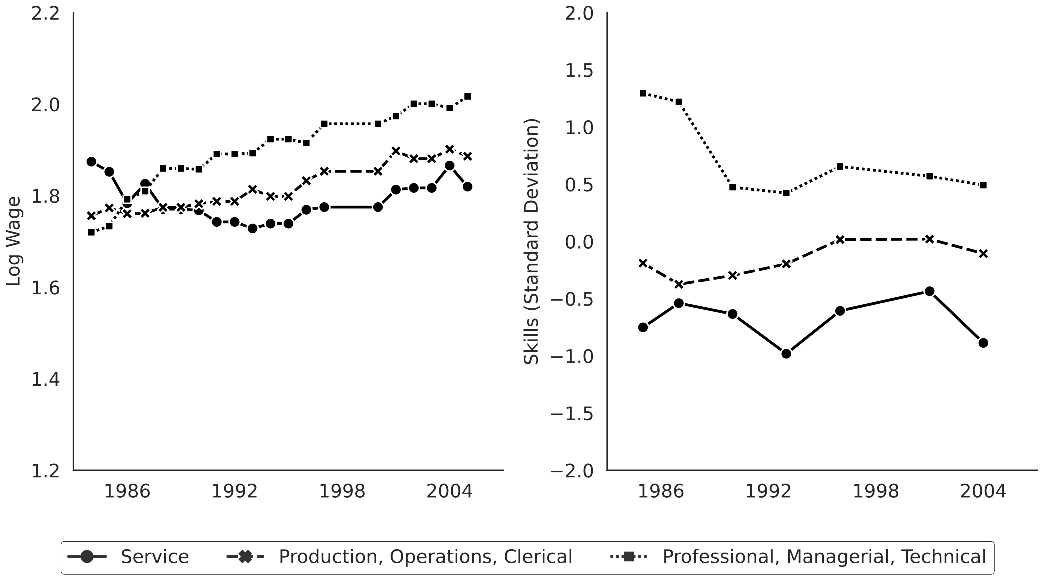
Differences in wages between occupations that are not explained by the prices of labor must be explained by differences in the skills of workers who select into each occupation. To quantify this, I next consider the opposite counterfactual excerise: rather than fixing skills to isolate the effect of switching occupations, I instead fix the occupation to isolate the effect of switching between conditional skill distributions. Panel

(B) of Figure 10 shows the evolution of potential log wages in the service occupation at the mean skill level of workers who choose each of the three occupations. This provides a measure of how much wage inequality is induced by differences in the skill levels of workers who choose each occupation.

The magnitude of differences in the counterfactual wages in this exercise are significantly larger than the differences in skill prices, and are similar in magnitude to the differences in the mean observed log wages in Panel (A) of Figure 2. This result indicates that in each time period, most of the difference in observed wages between occupations is due to sorting by skill level, rather than differences in skill prices. The differences in these counterfactual wages were relatively stable up to the mid 1990s, at which point the gap between the mean skills of professional workers started to increase relative to workers in the other two occupations.

Together, these results indicate that there were two distinct periods in this sample. Up to the middle of the 1990s, differences in skill prices between occupations widened; after that, these gaps were relatively stable, but sorting by skill level increased. This result has a possible interpretation in light of the evidence presented in Beaudry, Green, and Sand (2016) that there was a structural shift in the adoption of new information technology around the turn of the century. In particular, the first period can be understood as a time of rapid technological change, during which changes to the production process led to changes in productivity of labor in different occupations. In the second period, the prices of labor are more stable, but as the workforce adapted to the new technology.

Figure 11: Mean of Potential Log Wage by Occupation Holding Initial Marginal Distribution of Skills Fixed



In the analysis so far, I have considered two counterfactuals in which either the occupational choice or the quantile of the skill distribution is held fixed during a given time period. However, the skill distribution changed over time under the model estimates, so the temporal changes shown in Figure 10 include changes in the skill distribution. The potential wages in Panel (A) of Figure 10 are evaluted at the mean skill level in each period, and the counterfactual log wages in Panel (B) of Figure 10 are evaluated at the mean of the conditional skill distributions for each period.

In order to isolate the effect of changes in workers' sorting patterns from changes in the distribution of skills, I next consider a counterfactual in which the marginal distribution of skills is held fixed at the initial period. Panel (A) of Figure 11 shows the potential wages in each occupation at the mean skill level in the initial period. Holding the skill level fixed in this analysis does not significantly change the picture. The major difference is that holding skills fixed, wages do not decline across all occupations.

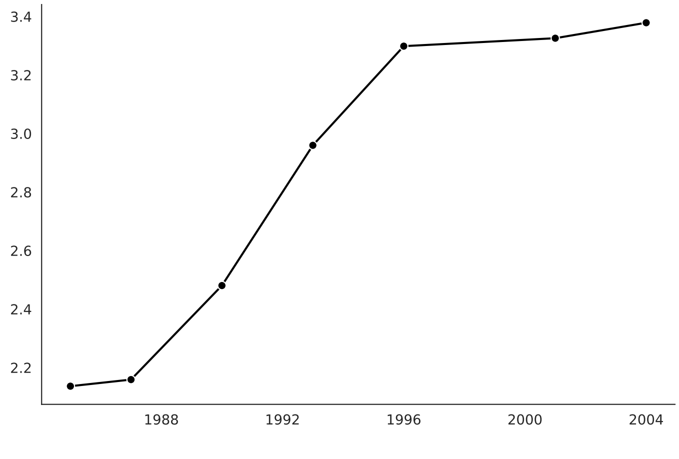
Panel (B) of Figure 11 shows the potential log wage in the service occupation holding the marginal distribution of skills fixed. This is the distribution of skills that would arise if the marginal distribution of skills remained the same over time, but the probability that workers choose each occupation change as estimated by the model. Formally, this is the distribution,

$$\hat{F}_{X;0}\hat{F}_{D|X;t}(d|x)$$

where $\hat{F}_{X,D;t}$ is the distribution of skills and occupational choices estimated for period t.

Recall that in the previous section, I compared the wage inequality that would arise if workers where randomly allocated to occupations to wage inequality under the estimated sorting pattern in each period. In that analysis, the marginal distribution of skills and skill prices were not fixed; under the model estimates, they change from period to period. The current counterfactual exercise, I fix the other two contributors to wage inequality (the marginal distribution of skills, and skill prices) and consider the impact of changes in sorting to wage inequality alone. The results in Panel (B) of Figure 11 show that holding these factors fixed, sorting did not appear to be a significant driver of rising inequality. When holding the marginal distribution of skills fixed, the differences between the mean skills of professional workers and production workers actually narrow in the first half of the period and are relatively stable afterward.

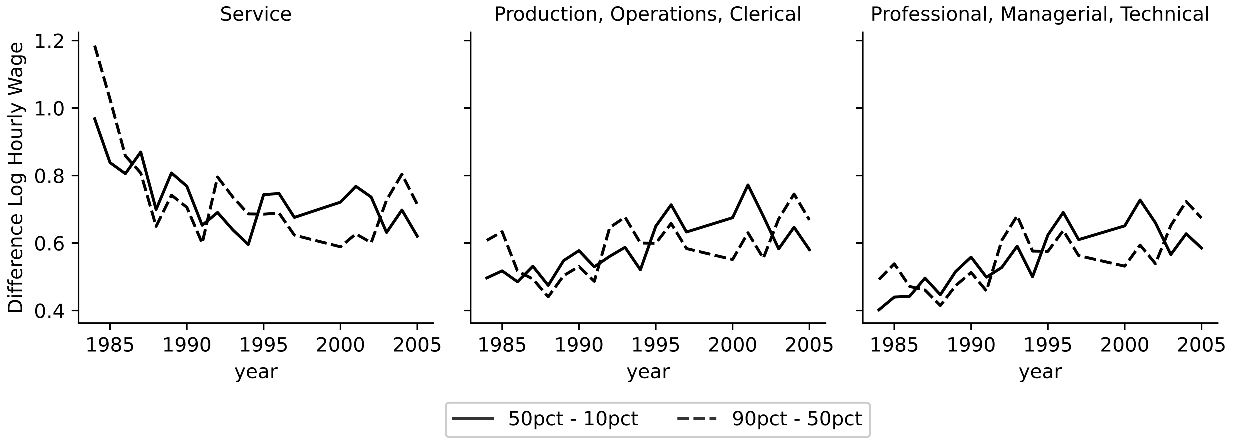
Figure 12: Difference between the 90th and 10th Percentile of the Distribution of Skills



What explains the apparent discrepency in these results and the findings in the previous sections? The key observation is that rising inequality in the marginal distribution of skills can lead to increasing inequality between the skills of workers who select each occupation even if workers do not sort more conditional on their skill level. If there are comparatively more workers at a high skill level, this can shift the distribution of skills among professional workers up if high skill workers are more likely to choose professional occupations.[5]

The model estimates of the marginal distribution of skills confirm that the distribution of skills did become more disperse during this period. Figure 12, shows that the difference between the 90th and 10th percentiles of the marginal skill distribution increased significantly at the beginning of this period. The increasing dispersion throughout the first half of the 1990s helps to explain the differences in patterns observed in Panel (B) of Figure 10 and Panel (B) of Figure 11. Though changes in sorting conditional on skills by itself would have led to decrease in the gap between the skills of professional and production workers over this period, the sharp increase dispersion in the skill distribution resulted in relative stability in this gap.

Figure 13: Upper and Lower Tail Inequality in Potential Wages



Returning to the three factors affecting changes in wage inequality, these results suggest that up the 1990s, increasing wage inequality was driven largely by increasing dispersion of the marginal distribution of skills with a smaller contribution from the relative prices across occupations. After accounting for these factors, changes to sorting patterns did not increase inequality more. In the period after the mid-1990s, the overall dispersion of the marginal wage distribution began to level off and differences in skill prices also stabilized. This mirrors the trend observed in raw wage data as shown in Panel (A) of Figure 1.

The analysis up to this point has focused on explaining rising wage inequality during the first half of the period. Recall that the other major trend during this period is the divergence between inequality in the upper and lower tails of the wage distribution. As show in in Panel (B) of Figure 4, inequality in the upper half of the wage distribution increased while inequality in the lower half of the distribution decreased. The Routine-Biased Technological Change (RBTC) hypothesis attempts to explain this by a decline in relative

[5] To make this more concrete, consider a simplified setting where there are two skill levels, $X \in \{0,1\}$ and two occupations $\{L,H\}$ in two period, $t \in \{0,1\}$. Suppose the probabilities of selecting the "high skill" occupation, H, decreases from $P_0(D_0 = H|X = x) = (x+1)/3$ in period 0, to $P_1(D_1 = H|X = x) = (x+1+K)/(3+2K)$ in period 1 for some $K > 0$. Notice that $1/2 < P_1(D_1 = W|X = 1) < P_0(D_0 = H|X = 1) = 2/3$. Suppose also that the maginal distribution of X changes from $P_0(X = 1) = 1/2$, to $P_1(X = 1) = 2/3$. Then after applying Bayes rule, we find that $E_0(X|D_0 = H) = 2/3$, and $E_1(X|D_1 = H) = (4+2K)/(5+3K) \in (2/3, 4/5)$. This illustrates how even decreasing the probability of choosing the high-skill occupation conditional on having a high skill level, the overall level of inequality increase if the marginal distribution of skills becomes more right skewed.

prices for production occupations. Under the parameter estimates from my model, there is limited evidence of a decline in prices for labor in production occupations. In Panel (B) of Figure 11, the wage of a production worker at the mean skill level increased relative to service workers over much of the period.

An alternative explanation for the divergence of inequality in the lower and upper tails of the log wage distribution is change within the skills of workers in each occupation. Figure 13 presents the upper and lower tail inequality in the skills of workers who select into each occupation. Upper and lower tail wage inequality decline in service occupations while both increase in the other two occupations. This presents a different explanation for changes wage polarization: since service workers have lower wages overall, decreasing wages inequality among service workers led to a decrease in inequality at the lower tail of the wage distribution. This indicates that change in the composition of workers in the service sector was an important factor in wage polarization.

7 Conclusion

In the debates surrounding increasing wage inequality and polarization in recent history, it has been widely acknowledged that workers' skills are not fully captured by simple measures such as education and work experience. In this paper, I develop an approach to identifying and estimating the changes in the unobserved skill distribution over time, and how workers of different skill levels have been reallocated across occupations. Using this approach, I document several features of the changing skill distribution over time which are new to the literature.

First, I find that sorting into occupations by skill level explains the majority of wage differences between occupations. Workers sort into occupations largely by skill level: the lowest skill workers tend to select service occupations, which have the lowest returns to skills and the highest skill workers tend to select professional occupations, which have the highest return to skills. However, in contrast to common theoretical predictions, this sorting does not appear to be based on comparative advantage. In fact, the estimation results suggest that most workers who are not working in professional occupations could increase their wage by switching occupations.

This finding adds to a growing literature showing that many labor market choices appear to be driven by nonpecuniary considerations. In contrast to the classic analysis of the Roy model showing that if workers select an occupation to maximize their income, this reduces overall wage inequality, the observed sorting patterns increases wage inequality relative to a baseline of random selection into occupations.

Second, I find that the distribution of workers' skills has become more disperse over time, and that the diverging patterns of wage inequality in the upper and lower tails of the wage distribution were due primarily to changes in the skills of workers within the production and service occupations. While the literature on routine-biased technological change focused on the hypothesis that changes in wages and employment in the production sector drove wage polarization, I find that falling lower-tail inequality in the service occupations account for most of the overall pattern of wage polarization.

Methodologically, this paper demonstrates that employment and wage data alone can reveal features about workers' skills which are not captured well by other measures. One major avenue for future research would be to push this methodology further to reveal clusters of occupations that use the same kinds of skills and how that has changed over time. Most research on the transferability of skills across occupations has relied on external measurements of the task content of occupations, such as the O*NET, which collects data on the requirements and tasks involved in different occupations.

Labor force data is an important complement to these datasets: the wages of occupations switchers provide direct evidence on how similar workers' skills are in different occupations, and occupational choice patterns reveal similarities in amenities in occupations and workers' preferences for them. The contributions of this paper to the analysis of the identification and computation of this model provide a foundation to estimate specifications of this model with higher-dimensional skills and finer-grained classifications of occupations, which can address these questions.

References

Abowd, John M., Francis Kramarz, and David N. Margolis. 1999. “High Wage Workers and High Wage Firms”. *Econometrica* 67 (2): 251–333.

Acemoglu, Daron., and David Autor. 2011. “Skills, Tasks and Technologies: Implications for Employment and Earnings”. Ed. by Orley. Ashenfelter and David Card. *Handbook of Labor Economics* 4:1043–171.

Altonji, Joseph G., Prashant. Bharadwaj, and Fabian Lange. 2012. “Changes in the Characteristics of American Youth: Implications for Adult Outcomes”. *Journal of Labor Economics* 30 (4): 783–828.

Arcidiacono, Peter, et al. 2020. “Ex ante returns and occupational choice”. *Journal of Political Economy* 128 (12): 4475–4522.

Autor, D.H., and D. Dorn. 2013. “The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market”. *American Economic Review* 103 (5): 1553–1597.

Autor, D.H., et al. 2006. “The Polarization of the U.S. Labor Market”. *American Economic Review* 96 (2): 189–194.

Autor, David .H., Lawrence F. Katz, and Alan B. Krueger. 1998. “Computing inequality: Have computers changed the labor market?” *Quarterly Journal of Economics* 113 (4): 1169–1213.

Autor, David H., and Michael J. Handel. 2013. “Putting Tasks to the Test: Human Capital, Job Tasks, and Wages”. *Journal of Labor Economics* 31 (S1): S59–S96.

Autor, David H., Lawerence F. Katz, and Melissa S. Kearney. 2008. “Trends in U.S. Wage Inequality: Revising the Revisionists”. *Review of Economics and Statistics* 90 (2): 300–323.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. “The Skill Content of Recent Technological Change: An Empirical Exploration”. *Quarterly Journal of Economics* 118 (4): 1279–1333.

Beaudry, P., D.A. Green, and B.M. Sand. 2016. “The Great Reversal in the Demand for Skill and Cognitive Tasks”. *Journal of Labor Economics* 34 (S1): S199–S247.

Böhm, Michael J. 2020. “The price of polarization: Estimating task prices under routine-biased technical change”. *Quantitative Economics* 11 (2): 761–799.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2019. “A Distributional Framework for Matched Employer Employee Data”. *Econometrica* 87 (3): 699–739.

Card, David., Jörg Heining, and Patrick Kline. 2013. “Workplace Heterogeneity and the Rise of West German Wage Inequality”. *Quarterly Journal of Economics* 128 (3).

Carneiro, Pedro, Karsten T. Hansen, and James J. Heckman. 2003. “Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice”. *International Economic Review* 44 (2): 361–422.

Carr, Michael D, Robert A Moffitt, and Emily E Wiemers. 2020. “Reconciling trends in volatility: Evidence from the sipp survey and administrative data”, NBER Working Paper, no. w27672.

Cavaglia, Chiara, and Ben Etheridge. 2020. “Job polarization and the declining quality of knowledge workers: Evidence from the UK and Germany”. *Labour Economics* 66:101884.

Chen, Xiaohong. 2007. “Large Sample Sieve Estimation of Semi-Nonparametric Models”. Chap. 76 in *Handbook of Econometrics*, 1st ed., ed. by J.J. Heckman and E.E. Leamer, vol. 6B. Elsevier.

Chernozhukov, Victor, et al. 2013. “Average and Quantile Effects in Nonseparable Panel Models”. *Econometrica* 81 (2): 535–580.

Cortes, Guido M. 2016. “Where Have the Middle-Wage Workers Gone? A Study of Polarization Using Panel Data”. *Journal of Labor Economics* 34 (1): 63–105.

Costinot, Arnaud, and Jonathan Vogel. 2010. “Matching and inequality in the world economy”. *Journal of Political Economy* 118 (4): 747–786.

D’Haultfoeuille, X. 2011. “On the completeness condition in nonparametric instrumental problems”. Cited By 29, *Econometric Theory* 27 (3): 460–471.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin. 1977. “Maximum likelihood from incomplete data via the EM algorithm”. *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1): 1–22.

Fox, Jeremy T, Kyoo il Kim, and Chenyu Yang. 2016. “A simple nonparametric approach to estimating the distribution of random coefficients in structural models”. *Journal of Econometrics* 195 (2): 236–254.

Freyberger, Joachim. 2018. “Non-Parametric Panel Data Models with Interactive Fixed Effects”. *Review of Economic Studies* 85 (3): 1824–1851.

Gottschalk, Peter, David A Green, and Benjamin M Sand. 2015. “Taking selection to task: Bounds on trends in occupational task prices for the US, 1984-2013”. *Working Paper.*

Gu, Jiaying, and Roger Koenker. 2017. “Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective”. *Journal of Business & Economic Statistics* 35 (1): 1–16.

Heckman, James J., and Bo E. Honore. 1990. “The Empirical Content of the Roy Model”. *Econometrica* 58 (5).

Heckman, James J., and Jose Scheinkman. 1987. "The Importance of Bundling in a Gorman-Lancaster Model of Earnings". *Review of Economic Studies* 54 (2): 243–255.

Heckman, James J., and Guilherme L. Sedlacek. 1985. "Heterogeneiety, Aggregation, and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market". *Journal of Political Economy*, no. 6.

– . 1990. "Self-selection and the distribution of hourly wages". *Journal of Labor Economics* 8 (1): S329–S363.

Heckman, James J., Jora. Stixrud, and Sergio. Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior". *Journal of Labor Economics* 24 (3): 411–482.

Hu, Y. 2017. "The Econometrics of Unobservables: Applications of Measurement Error Models in Empirical Industrial Organization and Labor Economics". *Journal of Econometrics* 200 (2): 154–168.

Hu, Yingyao, Robert Moffitt, and Yuya Sasaki. 2019. "Semiparametric estimation of the canonical permanent-transitory model of earnings dynamics". *Quantitative Economics* 10 (4): 1495–1536.

Hu, Yingyao, and Susanne M. Schennach. 2008. "Instrumental Variable Treatment of Nonclassical Measurement Error Models". *Econometrica* 76 (1): 195–216.

Jung, Jaewon, and Jean Mercenier. 2014. "Routinization-Biased Technical Change and Globalization: Understanding Labor Market Polarization". *Economic Inquiry* 52 (4): 1446–1465.

Kennan, John, and James Walker. 2011. "The Effect of Expected Income on Individual Migration Decisions". *Econometrica* 79 (1): 211–251.

Koenker, Roger, and Ivan Mizera. 2014. "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules". *Journal of the American Statistical Association* 109 (506): 674–685.

Lemieux, T. 2006. "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?" *American Economic Review* 96 (3): 461–498.

Lindsay, Bruce G. 1995. "Mixture Models: Theory, Geometry and Applications". *NSF-CBMS Regional Conference Series in Probability and Statistics* 5:i–163.

Moffitt, Robert, et al. 2022. "Reconciling Trends in US Male Earnings Volatility: Results from Survey and Administrative Data". *Journal of Business & Economic Statistics*: 1–19.

Moffitt, Robert A, and Peter Gottschalk. 2012. "Trends in the transitory variance of male earnings methods and evidence". *Journal of Human Resources* 47 (1): 204–236.

Mourifie, Ismael, Marc Henry, and Romuald Meango. 2018. "Sharp bounds and testability of a Roy model of STEM major choices". *Available at SSRN 2043117*.

Rosen, Sherwin. 1978. "Substitution and division of labour". *Economica* 45 (179): 235–250.

Shin, Donggyun, and Gary Solon. 2011. "Trends in men's earnings volatility: What does the Panel Study of Income Dynamics show?" *Journal of public Economics* 95 (7-8): 973–982.

Sorkin, Isaac. 2018. “Ranking firms using revealed preference”. *The Quarterly Journal of Economics* 133 (3): 1331–1393.

Teulings, Coen N. 1995. “The wage distribution in a model of the assignment of skills to jobs”. *Journal of political Economy* 103 (2): 280–315.

Wiswall, Matthew, and Basit Zafar. 2018. “Preference for the workplace, investment in human capital, and gender”. *The Quarterly Journal of Economics* 133 (1): 457–507.

A Appendix A: Identification Proofs

A.1 Proof of Theorem 1

The proof of Theorem 1 follows the approach developed in Hu and Schennach (2008) applied to panel data models in Freyberger (2018). I first establish a general lemma using high level conditions which is similar to the approach in Hu and Schennach (2008), and then in the main proof of Theorem 1, I follow an approach similar to Freyberger (2018), extending the result to allow for all parameters to depend on the choices and using a higher level condition on the functional form of the factor structure. The approach in this proof also differs in establishing the equivalence class of factor structures in the identified set rather than imposing normalizing assumptions along with the other assumptions to establish point identification. This makes it possible to then analyze the set of parameters that are point identified.

I will use the following notation. An integral operator L on an $\mathcal{L}^2(\mathcal{X},\mu)$ with a kernel f is defined by:

$$[Lg](y) = \int f(y,x)g(x)d\mu_x(x)$$

It’s adjoint, L^* is defined by:

$$[L^*h](x) = \int f(y,x)h(y)d\mu_y(y)$$

Lemma (A.1)**.** Let Y_0 and X be random variables with support $\mathcal{Y}_0 = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{Y}_3$ and $\mathcal{X}$. And let Y_j be the projection of Y onto $\mathcal{Y}_j$. For $j = 0, 1, 2, 3$, assume Y_j is absolutely continuous with respect to the measure μ_j and let f_{Y_j} be its density.

Suppose that for each $s \in S$, there exist integral operators, $L^s_j : \mathcal{L}(\mathcal{X}) \mapsto \mathcal{L}(\mathcal{Y}_j)$ with associated kernels, $f^s_j : \mathcal{Y}_j \times \mathcal{X} \mapsto \mathbb{R}$, for each $j = 0, 1, 2, 3$, and a function $f^s : \mathcal{X} \mapsto \mathbb{R}$, which satisfy the following properties,

1. f^s_j are strictly positive and bounded, $f^s_0(y,x) = \prod_{j=1}^3 f^s_j(y_j,x)$ with $\int f_j(y_j,x)d\mu_j(y_j) = 1$ for all $x \in \mathcal{X}$
2. L^s_1 and $(L^s_3)^*$ are invertible
3. For all $x \in \mathcal{X}$: $\int\int |f_2(y,x) - f_2(y,x')|d\mu_2(y)d\mu_X(x') > 0$
4. f^s is strictly positive on and bounded on $\mathcal{X}$ and $L^s_j f^s = f_{Y_j}$ for $j = 0, 1, 2, 3$

then for $s, s' \in S$ there is a bijective function $\pi : \mathcal{X} \mapsto \mathcal{X}$ such that $f^{s'}_j(y|x) = f^s_j(y|\pi(x))$, and $f^{s'}$ is the density with respect to μ_X for the probability measure, $\bar{\mu}(B) = \int_{\mathcal{X}} \mathbf{1}(\pi(x) \in B)f^s(x)d\mu_x(x)$.

Proof. The proof will make use of the following linear operators. First, for any $y \in \mathcal{Y}_2$, the operators $M(y_2) : \mathcal{L}(\mathcal{X}) \mapsto \mathcal{L}(\mathcal{X})$, $\bar{M} : \mathcal{L}(\mathcal{X}) \mapsto \mathcal{L}(\mathcal{X})$ are defined by,

$$\begin{aligned}
[M(y_2)g](y_1) &= \int f_{Y_0}(y_1, y_2, y_3) g(y_3) d\mu_3(y_3) \\
[\bar{M}g](y_1) &= \int \int f_{Y_0}(y_1, y_2, y_3) g(y_3) d\mu_2(y_2) d\mu_3(y_3)
\end{aligned}$$

Next, for $s \in S$, $j = 1, 2, 3$, and $y \in \mathcal{Y}_j$, the operator, $D_j^s(y) : \mathcal{L}(\mathcal{X}) \mapsto \mathcal{L}(\mathcal{X})$ is defined by,

$$[D_j^s(y)g](x) = f_j^s(y, x) g(x)$$

And the operator $H^s : \mathcal{L}(\mathcal{X}) \mapsto \mathcal{L}(\mathcal{X})$ is defined by,

$$[H^s g](x) = f^s(x) g(x)$$

First I show that for any $s \in S$,

$$\begin{aligned}
M(y_2) &= L_1^s D_2^s(y_2) H^s (L_3^s)^* \\
\bar{M} &= L_1^s H^s (L_3^s)^*
\end{aligned}$$

The first is obtained from the following derivation,

$$\begin{aligned}
[M(y_2)g](y_1) &= \int f_{Y_0}(y_1, y_2, y_3) g(y_3) \mu_3(y_3) \\
&= \int L_0^s f^s(y_1, y_2, y_3) g(y_3) d\mu_3(y_3) \\
&= \int \int f_1^s(y_1, x) f_2^s(y_2, x) f_3^s(y_3, x) f^s(x) d\mu_x(x) g(y_3) d\mu_3(y_3) \\
&= \int f_1^s(y_1, x) f_2^s(y_2, x) f^s(x) \int f_3^s(y_3, x) g(y_3) d\mu_3(y_3) d\mu_x(x) \\
&= \int f_1^s(y_1, x) f_2^s(y_2, x) f^s(x) [(L_3^s)^* g](x) d\mu_X(x) \\
&= \int f_1^s(y_1, x) D_2^s(y_2) [(L_3^s)^* g](x) d\mu_X(x) \\
&= L_1^s D_2^s(y_2) H^s (L_3^s)^* g](y_1)
\end{aligned}$$

The second line follows from condition 4, the third line applies the definition of the operator L_0^s and condition 1, and the fourth line is an application of Fubini's theorem. The rest are applications of the definitions of the linear operators. The decomposition of $\bar{M}$ follows analogously.

H^s is invertible because, under condition 4, we can define $(H^s)^{-1}$ by,

$$[(H^s)^{-1} g](x) = 1/f^s(s)$$

which can be verified to be an inverse as follows,

$$[H^s (H^s)^{-1} g](x) = H^s(1/f^s(x)) g(s) = f^s(x)/f^s(x) g(x) = g(x)$$

Combining this with condition 2, this implies that $\bar{M}$ is invertible and for each $y_2 \in \mathcal{Y}_2$,

$$M(y_2) M^{-1} = L_1^s D(y_2; f_2^s) (L_1^s)^{-1} \tag{6}$$

Decompositions of this form, which are analogous to an eigenvalue decomposition, are studied in Hu and Schennach (2008) and Freyberger (2018). Hu and Schennach (2008) shows that under condition 5, this decomposition is unique up to the three potential indeterminancies. These three indeterminancies can be decribed by showing how to satisfy an alternative function, $\bar{f}_j$, which would satisfy (6).

1. Scale: For any positive function $a : \mathcal{X} \mapsto \mathbb{R}_+$, $\bar{f}_1(y,x) = a(x)f_1^s(y,x)$
2. Eigenvalue degeneracy: If there is a subset $\bar{\mathcal{X}} \subseteq \mathcal{X}$, $f_2^s(y_2,x) = f_2^s(y_2,x')$ for all $x, x' \in \bar{\mathcal{X}}$, $y_2 \in \mathcal{Y}_2$, then there is a function $b : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ such that, $\bar{f}_1(x) = \int_{\bar{X}} b(x,x')f_1^s(x')dx'$.
3. Ordering: For one-to-one function $\pi : \mathcal{X} \mapsto \mathcal{X}$, set, $\bar{f}_j(x) = f_j(\pi(x))$ for $j = 1, 2, 3$.

As in Hu and Schennach (2008) and Freyberger (2018), the first indeterminancy is ruled out by condition 1, and the second indeterminancy is ruled out by condition 3.

This leaves the only third indeterminancy, which shows that if there exists $s' \in S$ such that $s' \neq s$ such that $f_j^{s'}(y,x) = f_j^s(y,\pi(x))$, for some bijective function $\pi : \mathcal{X} \mapsto \mathcal{X}$. Finally, recall that under condition 4, if $s \in S$, then $f_{Y_j}(y) = L_j^s f^s(y)$ for each $j = 0, 1, 2, 3$, and that f^s defines a probability density with respect to μ_x. For $s' \in S$,

$$f_{Y_j}(y) = \int f_j^s(y,\pi(x))f^{s'}(x)d\mu_x(x)$$

This probability density s' can be found by noticing that it is the density of a random variable $X^{s'} = \pi(X^s)$ where X^s is a random variable with the density f^s. Finally, since L_1^s is invertible, it is unique.

□

Proof of Theorem 1

For a sequence of choices partition the choices into $d = (d^1, d^2, d^3)$, where d^1 and d^3 are length $\bar{k}$ and d^2 is length $\bar{t} - 2\bar{k}$. Let the operator L_j to be an integral operator with the kernel $f_{Y^{t_j}(d^j)|X}$, and let L_0 be the integral operator with kernel $f_{Y^{\bar{t}}(d_{\bar{t}})|X}$. Under assumption M, $Y^{t_1}(d^1), Y^{t_2}(d^2), Y^{t_3}(d^3)$ are mutually independent conditional on X, and $L_j^s f_X = f_{Y_j}$ by construction. This establishes conditions 1 and 4 of lemma A.1, so it is sufficient then to establish that L_j are invertible for $j = 1, 2, 3$.

For each $j \in 1, 2, 3$, invertibility of L_j is established by verifying the conditions of Theorem 2.1(i) in D'Haultfoeuille (2011). $Y^{\bar{t}}(d^{\bar{t}})$ has the additive structure used in D'Haultfoeuille (2011), and we can verify the conditions of theorem 2.1(i) as follows:

- A.1: under assumption 1, $X \perp U^t(d)$ for all (t,d)
- A.2: under assumption 5, for each (t,d), $\mu^t(d,\cdot)$ is continuous, and under assumption 4, X is absolutely continuous so $\mu^t(d,X)$ is absolutely continuous. Furthermore under assumption 5, $\mu^t(d,\mathcal{X}) = R^{\bar{k}}$.
- A.3, A.4: under assumption 3, $U^t(d)$ is continuous and has a non-zero, infinitely differential characteristic function

This shows that $Y^{t_j}(d^j)$ is bounded complete for X, which is equivalent to invertibility of L_j operator on the space bounded integrable functions. Invertibility of L_3 combined with the assumption that the conditional densities are bounded, imply that L_3^* is invertible. Finally, invertibility of L_2 also establishes condition 3 in Lemma A.1.

This shows that for each sequence d, the conditional density, $f_{Y^t(d^t)|X}$, and conditional density of X, $f_{X|D}$ are identified up to a bijective map of the support of X, $\pi : \mathcal{X} \mapsto \mathcal{X}$. It is left to show that the same mapping π relates the full joint distribution of (Y, D) across all choices sequences.

Suppose (μ^0, F^0, G^0) and (μ, F, G) are in the identified set and for choice sequences $d \neq d'$, $F_{X,D}(x, d) = F^0_{X,D}(\pi(x), d)$ and $F_{X,D}(x, d) = F^0_{X,D}(\pi'(x), d')$. I will show this implies that $\pi = \pi'$. Partition $t = (t^1, t^2, t^3)$ and $d = (d^1, d^2, d^3)$, and $d' = (d'^1, d'^2, d'^3)$. Now consider the choice sequence $d'' = (d^1, d'^2, d'^3)$.

Applying the previous result for $t \in t^1$, $\mu_t(d_t, x) = \mu^0(d_t, \pi(x))$ and for $t \in (t^2, t^3)$, $\mu_t(d_t, x) = \mu^0(d_t, \pi'(x))$. But since there is a unique function $\tilde{\pi}$ for $\mu_t(d_t, x) = \mu^0_t(d_t, \tilde{\pi}(x))$ and $F_{X|D}(x|d'') = F_{X|D}(\tilde{\pi}(x)|d'')$, this implies $\pi = \pi'$.

A.2 Proof of Theorem 4

Define the following notation. Bolded variables are tuples[6] of length less than T, that is:

$$\mathbf{x} = (x_1, \ldots, x_S) : \text{ for some } S \leq T$$

For any $(\mathbf{t}, \mathbf{d})$, $\mathbf{y}(\mathbf{t}, \mathbf{d})$ is the $(n \times S)$ matrix with entries, $y_{it}(\mathbf{t}, \mathbf{d}) = \tilde{y}_{it}(\mathbf{t}, \mathbf{d}) - n^{-1} \sum_j \tilde{y}_{it}(\mathbf{t}, \mathbf{d})$, where for each $i \in [n]$, $t \in \mathbf{t}$, $\tilde{\mathbf{y}}_{it}(\mathbf{t}, \mathbf{d})$ is:

$$\tilde{y}_{it}(\mathbf{t}, \mathbf{d}) = \left(\prod_{s \in \mathbf{t}, d \in \mathbf{d}} \mathbf{1}(d_{is} = d) \right) y_{it}$$

For each $(t, d) \in [T] \times \mathcal{D}_t$, define the set:

$$A(t, d) = \left\{ (\mathbf{t}_1, \mathbf{t}_2, \mathbf{d}_1, \mathbf{d}_2) : |\mathbf{t}_1| = k, \{t\} \cup \mathbf{t}_1 \cup \mathbf{t}_2 = [T], d_t \in \mathcal{D}_t, \mathbf{d}_1 \in \mathcal{D}^{\mathbf{t}_1}, \mathbf{d}_2 \in \mathcal{D}^{\mathbf{t}_2} \right\}$$

This set gives all the ways to partition the set of time periods such that one partition is t and another has length k, and covers all possible choices sequences for each partition. Using this, for each $(t, d) \in [T] \times \mathcal{D}_t$, we have the following set of moment restrictions. For each $(\mathbf{t}_1, \mathbf{t}_2, \mathbf{d}_1, \mathbf{d}_2) \in A(t, d)$,

$$\mathbf{y}(\mathbf{t}_2, \mathbf{d}_2)^T \left(\mathbf{y}(t, d) - \mathbf{y}(\mathbf{t}_1, \mathbf{d}_1) \mathit{\Lambda}_{\mathbf{t}_1}(\mathbf{d}_1)^{-1} \lambda_t(d) \right) = 0$$

To simplify notation further, define the set,

$$\mathcal{Y}(t, d) = \{ (\mathbf{y}(\mathbf{t}_1, \mathbf{d}_1), \mathbf{y}(\mathbf{t}_2, \mathbf{d}_2), \mathit{\Lambda}_{\mathbf{t_1}}(\mathbf{d}_1) : (\mathbf{t}_1, \mathbf{t}_2, \mathbf{d}_1, \mathbf{d}_2) \in A(t, d) \}$$

Then for each (t, d), and $(\mathbf{y}_1, \mathbf{y}_2, \mathit{\Lambda}_1) \in \mathcal{Y}(t, d)$, we have,

$$\mathbf{y}_2^T y - \mathbf{y}_2^T \mathbf{y}_1 \mathit{\Lambda}_1^{-T} \lambda_t(d_t) = 0$$

Premultiplying this by $L(\mathbf{y}_1, \mathbf{y}_2) = \mathit{\Lambda}_1^{-T} (\mathbf{y}_1^T \mathbf{y}_2 (\mathbf{y}_2^T \mathbf{y}_2)^{-1} \mathbf{y}_2^T \mathbf{y}_2)^{-1} \mathbf{y}_1^T \mathbf{y}_2 (\mathbf{y}_2^T \mathbf{y}_2)^{-1}$, we get:

$$L(\mathbf{y_1}, \mathbf{y_2}) \mathbf{y}_2^T y - \mathbf{y}_2^T \mathbf{y}_1 \mathit{\Lambda}_1^{-T} \lambda_t(d_t) = \left(\mathit{\Lambda}_1^T \hat{m}(y, \mathbf{y}_1, \mathbf{y}_2) - \lambda \right) = 0$$

[6] I define a tuple here as an enumerated set. The union of two tuples as a union of the set with any arbitrary ordering. Two tuples are equal if the sets are equal, regardless of ordering.

where,

$$\hat{m}(y, \mathbf{y}_1, \mathbf{y}_2) = (\mathbf{y}_1^T \mathbf{y}_2 (\mathbf{y}_2^T \mathbf{y}_2)^{-1} \mathbf{y}_2^T \mathbf{y}_2)^{-1} \mathbf{y}_1^T \mathbf{y}_2 (\mathbf{y}_2^T \mathbf{y}_2)^{-1} \mathbf{y}_2^T y$$

which is simply a 2SLS estimator of Next, let ϕ_1, ϕ_0 be the $k \times T|\mathcal{D}|$ and $1 \times T|\mathcal{D}|$ matrices, such that, $\varLambda^T \phi_1^T = \varLambda_1^T$ and $\varLambda^T \phi_0^T = \lambda$.

Assuming that the relevant matrices are invertible, we can rewrite the moment conditions as:

$$\begin{aligned} & 0 = \varLambda^T \phi_1^T \hat{m}(y, \mathbf{y}_1, \mathbf{y}_2) - \varLambda^T \phi_0^T \\ \iff & 0 = (\hat{m}^T(y, \mathbf{y}_1, \mathbf{y}_2)\phi_1 - \phi_0)\varLambda \end{aligned}$$

Enumerating $\{(y^{(q)}, \mathbf{y}_1^{(q)}, \mathbf{y}_2^{(q)}, \phi_1^{(q)}, \phi_0^{(q)}) : q = 1, \ldots, Q\}$

$$\hat{M} = \begin{bmatrix} \hat{m}^T(y^{(1)}, \mathbf{y}_1^{(1)}, \mathbf{y}_2^{(1)})\phi_1^{(1)} - \phi_0^{(1)} \\ \vdots \\ \hat{m}^T(y^{(Q)}, \mathbf{y}_1^{(Q)}, \mathbf{y}_2^{(Q)})\phi_1^{(Q)} - \phi_0^{(Q)} \end{bmatrix}$$

Finally, let $\varLambda = A + \tilde{\varLambda}$, where A are a set of normalizations. Let $\hat{M}_0 = \hat{M}A$. Then we can write the system of equations as:

$$\begin{aligned} 0_{Q \times k} &= \hat{M}\tilde{\varLambda} - \hat{M}_0 \\ 0_{Qk \times 1} &= (I_k \otimes \hat{M})\text{vec}(\tilde{\varLambda}) - \text{vec}(\hat{M}_0) \end{aligned}$$