Tax Policies and Substitution between Skilled and Unskilled Labor: Inference from a Wage Deduction Tax Policy in China

Yi Lu [*] Tsinghua

Wei Lyu [†] DUFE

Jianguo Wang [‡] NUS

Wanyu Yang [§] DUFE

This version: March 2020

Abstract

This paper explores a kinked wage deduction tax policy in China to analyze firms' responses to tax policies in the employment structure adjustment margin. A stylized theoretical model estimates that the substitution elasticity between skilled and unskilled labor is approximately 1.18, and that firms tend to over-report the unskilled labor by 5.54% to 7.91%. Welfare analysis shows that the manipulation of the reported unskilled labor eases the negative effect of the wage deduction policy on GDP, whereas the adjustment of labor magnifies the effect. An application of this framework to the payroll tax using the tax schedule and reforms in the U.S. shows a significant role of employment adjustment in transmitting the effects of tax incidences.

Keywords: bunching, tax-induced employment distortion, elasticity of substitution between skilled and unskilled labor, employment structure adjustment

JEL Classification:

[*]School of Economics and Management, Tsinghua University, Beijing, 100084, China (luyi@sem.tsinghua.edu.cn)

[†]Institute of Economic and Social Development, Dongbei University of Finance and Economics, Dalian, 116023, China (weilv008@163.com)

[‡]School of Economics, National University of Singapore, 117570, Singapore (wangjianguo@u.nus.edu)

[§]Institute for Advanced Economic Research, Dongbei University of Finance and Economics, 116025, China (wanyu.yang@dufe.edu.cn)

1 Introduction

Tax policies, critical to economic development, are at the top of the agenda of policy makers. To understand how tax policies work, it is important to analyze the responses of firms. Recent studies have studied the responses of investment (e.g., House and Shapiro, 2008; Yagan, 2015; Zwick and Mahon, 2017; Ohrn, 2018), wage and employment (e.g., Serrato and Zidar, 2016; Fuest et al., 2018; Saez et al., 2019), pricing (e.g., Benzarti et al., 2019), and capital labor ratio (e.g., Benzarti and Harju, 2018). In this paper, we investigate the adjustment margin of the employment structure, i.e., the substitution between skilled and unskilled labor.

Specifically, we intend to quantify the role of employment structure adjustment in transmitting the impacts of tax instruments. Key parameters in the evaluation are the substitution elasticities between skilled labor and unskilled labor and between labor and other production factors, such as capital. To this end, we study a tax policy in China that generates a kink point in firms' budget sets. By applying the bunching method to the universe of all manufacturing firms, we are able to solicit the substitution elasticities among production factors (i.e., skilled labor, unskilled labor, and capital). To illustrate the role of employment adjustment, we first investigate how firms respond to our focal tax policy. Then, we evaluate the effect of payroll tax, a widely-used tax instrument on the employment. As China does not apply the payroll tax, we consider the tax schedule in the U.S., and investigate the consequences of two on-going policy debates regarding changes in the social security payroll tax.

Our empirical analysis explores a wage deduction tax policy in China. Specifically, after China adopted a reform and opening up strategy in the late 1970s, it used a "dual-track" corporate income tax scheme for domestic enterprises (DEs) and foreign-invested enterprises (FIEs). Under this scheme, FIEs could fully deduct wage bills from taxable corporate income, whereas DEs could claim up to only a statutory province-level wage limit. This nonlinear wage deduction schedule introduces a kink point in firms' budget sets, affecting firms' choices of production inputs.

To estimate the substitution elasticities among inputs, we apply a bunching method to the first wave of economic census in 2004, which covers basic information and balance sheets for all manufacturing firms. The graphical results clearly demonstrate a sharp jump in the density distribution of the average per worker monthly wage for DEs around the wage deduction limit point, suggesting firms' tax responses to the policy. We do not find the same pattern of bunching for FIEs that can fully deduct wages or for DEs using the second wave of economic census in 2008, when the wage deduction limit policy was abolished. These observations confirm that the bunching of affected DEs to the policy is not spurious.

To gain insight into how firms responded to the wage deduction limit, we follow the methodology of Diamond and Persson (2017), which is based on the premise that the rela-

tionship between the variables of interest and the average per worker monthly wage would be smooth in the absence of a deduction limit. Hence, the relationship characterized in the nonmanipulation regions can be used to construct a counterfactual relationship for the manipulation regions around the kink point. We find that firms did not relabel part of their labor costs as unemployment insurance, employment benefit, and administration costs (which are fully deductible from taxable income) in response to the wage deduction limit. Instead, firms decreased their capital input and substituted skilled labor with unskilled labor. This decline in the labor ratio may come from the substitution of skilled with unskilled labor and the manipulation of the total employment level.

We then construct a theoretical model to quantify the role of employment adjustment. Specifically, firms differ in their relative importance of skilled and unskilled labor to output production and optimally choose their inputs given the tax policies. By comparing firm behavior under the wage deduction limit and that under a counterfactual tax policy when firms can fully deduct wages, we can derive the mass of firms that choose to bunch at the deduction limit in the former scenario. A crucial part of the model estimation is the estimation of the density distribution of the average monthly wage per worker under the counterfactual scenario. To this end, we use two methods: (1) a parametric approach following Chetty et al. (2011) to estimate the counterfactual density from the observed density; and (2) a nonparametric approach using groups of firms that were not subject to the wage deduction limit of 960 RMB as the counterfactuals and combining a difference-in-differences (DD) spirit. We estimate the substitution elasticity between skilled and unskilled labor $\hat{\sigma}$ to be around 1.1548 to 2.5926, which implies that when the wage ratio of skilled over unskilled labor increases by 1 percent, firms would decrease the ratio of skilled over unskilled labor by approximately 1.1548 to 2.5926 percent.

To alleviate the concern that firms may manipulate their reporting of employment (e.g., adding "ghost workers" to inflate the employment) in response to the policy, we extend our model to incorporate this possibility. Specifically, we allow firms to over-report the number of unskilled laborers by including a new parameter (i.e., the booking manipulation degree) in the model. To simultaneously pin down the two parameters of interests (i.e., the labor substitution elasticity and the booking manipulation degree), we explore two wage deduction limits implemented in different regions (i.e., 800 RMB and 960 RMB) to construct two moment equations. The estimation results show that firms reported 3.20% to 8.52% more nonexistent unskilled labor to claim the wage deduction from taxable income. The elasticity of substitution remains robust, ranging around 1.1803 to 1.2821.

Based on the substitution elasticity and the density distribution of firm heterogeneity estimated from the nonparametric DD approach and the extended model with the booking manipulation, we then quantify the role of employment structure adjustment underlying the

effect of the tax incidence. Specifically, we find that the wage deduction limit policy causes the aggregate GDP to decline by approximately 0.2872 percent. However, when firms are not allowed to change their employment structure, GDP increases by 0.8863 percent instead with the wage deduction limit. The intuition for the opposite effects is as follows. Firms respond to the imposed wage deduction limit by manipulating the reported employment level and substituting skilled labor with unskilled labor. The booking manipulation reduces corporate tax and hence stimulates production, whereas the labor substitution decrease firms' productivity and hence reduce total production. When firms are able to adjust their employment structure, the adjustment margin dominates the manipulation margin and hence leads to a decrease in GDP. However, when the labor substitution is not allowed, the manipulation eases the effect of tax incidence and hence GDP increases.

We take the social security payroll tax in the U.S. as an example to further illustrate the role of employment adjustment in transmitting the effects of tax incidences. Specifically, to shed light on the proposal by the H.R. 860 bill to have an annual increase in the payroll tax rate, we conduct a counterfactual mimicking this setting. With flexible employment adjustment, the increase in payroll tax rate increases total payroll tax revenue but reduces firms' capital and employment inputs and, hence, GDP. These results are consistent with the findings by Benzarti and Harju (2018), who employ a discontinuity in the average payroll tax rate and show that a higher tax rate decreases both capital and labor inputs and, hence, the sales of firms. The prohibition of employment adjustment makes firms more responsive to payroll tax increases and generates an additional 3.2 percent decline in GDP. The mitigating role of employment structure adjustment in the payroll tax setting is understandable. The increase in payroll tax rate makes firms use more skilled labor to replace unskilled labor, which reduces the negative shock of the tax incidence.

The maximum taxable earnings is adjusted annually in the U.S. according to the change in wage index. To illustrate how firms respond to the change in the maximum taxable earnings, we study a counterfactual with the same setup. The estimation results show that the increase in maximum taxable earnings decreases the ratio of skilled to unskilled labor as skilled labor becomes relatively expensive; as a result, GDP declines. However, total employment, capital, and GDP decline less when firms cannot adjust their employment structure. The intuition behind this result is that the increase in the maximum taxable earnings leads to a substitution from skilled labor to unskilled one, which amplifies the negative shock of the tax incidence.

Although the estimated absolute changes in input factors, GDP, and tax revenues may not provide useful insights on the welfare analysis due to the different market contexts between China and U.S. the estimated percentage changes in these variables when firms are not allowed to adjust the ratio of skilled to unskilled labor would shed some lights on the

mitigating role of employment structure adjustment in the effect of tax incidences.

Literature review. Our work is related to the literature on tax policies that target specific groups of workers and hence change firms' employment decisions. Examples include the work opportunity tax credit policy and the Indian employment credit policy in U.S., the apprenticeship job creation tax credit policy in Canada, and the payroll tax cuts policy for young workers in Sweden. Specifically, Katz (1998) analyzes the U.S. wage subsidy policies for the disadvantaged in U.S.; Huttunen et al. (2013) examine the effect of payroll tax cuts for older workers in Finland; Elias (2015) investigates the payroll tax cut policy for old and young workers in Spain; and Saez et al. (2019) study the payroll tax cut policies for young workers in Sweden. The departure of our study is that we investigate a tax policy designed without the intention to benefit specific groups of workers but indirectly distorting firms' employment preference. Meanwhile, we quantify the role of employment adjustment in transmitting the effects of the tax incidence.

This paper also contributes to the literature on the substitution between skilled and unskilled labor. Most previous studies have attempted to explain the rapid increase in the college premium in the U.S. and have therefore concentrated on estimating the substitution elasticity between college and high school labor (Katz and Murphy, 1992; Heckman et al., 1998; Krusell et al., 2000 ; Card and Lemieux, 2001; Acemoglu, 2002; Autor et al., 2008); see Katz and Autor (1999) for a review of this literature. One exception is Angrist (1995), who investigates the substitution between workers with sixteen years of schooling and those with less than twelve years of schooling in Palestine. Our work focuses on investigating a large emerging economy, i.e., China, and estimating the elasticity of substitution between high school and non-high school graduates, which is of more importance than that between college and non-college labor in the developing country setting.

The present study is related to recent literature that estimates behavioral responses to discontinuities in incentives created by kinked or notched policies.[1] The recent literature has applied the bunching methodology to explore various types of behavioral responses of firms, including splitting responses of large firms to a value-added tax (VAT) threshold (Onji, 2009), sales size adjustment in response to a VAT threshold (Harju et al., 2016; Liu et al., 2017), employment size adjustment in response to thresholds in labor laws and accounting and legal rules (Gourio and Roys, 2014; Garicano et al., 2016), substituting between labor and capital in response to a minimum wage rule (Harasztosi and Lindner, 2019), and R&D investment adjustment in response to a threshold associated with corporate tax cuts (Chen et al., 2019). This paper contributes to the literature by employing the methodology to

[1]The methodology based on kinked policies is developed by Saez (2010) and Chetty et al. (2011), while the methodology based on notched policies is developed by Kleven and Waseem (2013); see Kleven (2016) for a review.

study the substituting responses of firms between skilled and unskilled labor.

This paper is organized as follows. In Section 2, we describe the policy background and the data. In Section 3, we present the bunching evidence that motivates this paper and the reduced-form evidence of how firms respond to the wage deduction limit policy. In Section 4, we develop theoretical models to illustrate the behavioral responses of firms and present results for the model estimation. In Section 5, we conduct welfare analysis and quantify the role of employment structure adjustment in transmitting the distortion effect of the wage deduction limit policy and payroll tax incidence. Section 6 concludes the paper.

2 Background

2.1 Wage Deduction Limit Policy

In 1978, China began to implement economic reforms. However, the lack of capital and technology hampered economic development. To attract foreign investment and introduce new technology, the government adopted a "dual-track" CIT scheme for DEs and FIEs. Specifically, all DEs were governed by the Provisional Regulations of the People's Republic of China on Corporate Income Tax (promulgated in 1993), whereas FIEs were governed by the Income Tax Law of the People's Republic of China for Enterprises with Foreign Investment and Foreign Enterprises (promulgated in 1991). This dual-track CIT scheme lasted for approximately 15 years. In 2008, to provide a fair environment for all firms, China abolished the dual-track tax scheme and replaced it with a universal tax law—the Law of the People's Republic of China on Corporate Income Tax.

During the dual-track CIT scheme period, FIEs could fully deduct wage bills from taxable corporate income, but DEs could claim only up to a statutory wage limit.[2] Specifically, when the average per worker monthly wage was less than or equal to the statutory wage limit, DEs could fully deduct their wage bills; otherwise, they could deduct only wage bills from taxable income for the CIT up to the imposed limit for each employee. The historic changes in the monthly wage deduction limit levels are summarized in Table 1. The limit was initially set at 500 RMB by the Chinese State Administration of Taxation (SAT) in 1994 and was increased to 550 RMB in 1996 and further to 800 RMB in 2000.[3] The local taxation bureaus were given the authority to inflate the limit up to 120% by the SAT, i.e., up to 600 RMB in 1994, 660 RMB in 1996, and 960 RMB in 2000, respectively. In July 2006, the SAT set the limit

[2]DEs and FIEs were also charged different CIT rates. Specifically, DEs were subject to CIT rates of 18%, 27%, and 33% for taxable income less than 30,000 RMB, between 30,000 and 100,000 RMB, and more than 100,000 RMB, respectively. However, FIEs were subject to CIT at reduced tax rates ranging from 15% to 33%, conditional on the location and the main business.

[3]1 U.S. dollar was traded at 8.28 RMB from 1994 to July 2005, after which China adopted a managed floating exchange rate system.

at 1600 RMB and canceled the adjustment power of local bureaus. After the dual-track CIT scheme was abolished in 2008, DEs were also allowed to deduct full wage bills.

[Insert Table 1 Here]

Figure 1 shows the distribution of wage deduction limits across regions in 2004, our sample period. Specifically, three municipalities (Beijing, Shanghai, and Tianjin), a subprovincial city (Qingdao), and six provinces (Anhui, Fujian, Guangdong, Hunan, Jiangsu, and Zhejiang) adjusted the limit to 960 RMB. One municipality (Chongqing), all other subprovincial cities, and 15 provinces (Gansu, Guangxi, Guizhou, Hebei, Henan, Hubei, Hunan, Jiangxi, Ningxia, Shaanxi, Shandong, Shanxi, Sichuan, Xinjiang, and Yunnan) implemented the regulated 800 RMB limit. Heilongjiang, Jilin, Liaoning, and Tibet adopted a wage deduction limit of 1200 RMB, due to the large-scale development strategy in west and north-east China. For three provinces (Hainan, Inner Mongolia, and Qinghai), the implemented limits were not documented.

[Insert Figure 1 Here]

2.2 Data and Variables

Our analysis is based mainly on the first Economic Census conducted by China's National Bureau of Statistics (NBS) in 2004, covering all firms in the secondary and tertiary sectors in China. Basic information for each firm includes the location, industry, ownership, and employment by education. Meanwhile, the data also have firms' balance sheets, which record firm output, capital, revenue, profit, total wage bill, total employee benefit, administrative cost, tax, etc.

In this paper, we focus on the tax avoidance behavior of domestic manufacturing firms who were subject to the wage deduction policy in 2004. Specifically, we exclude the three regions without information about wage deduction limits (i.e., Hainan, Inner Mongolia, and Qinghai) and four specially treated regions (i.e., Heilongjiang, Jilin, Liaoning, and Tibet). Meanwhile, one complication with the wage deduction limit of 800 RMB is that it coincided with the monthly exemption threshold for individual income tax during our sample period.[4] To this end, we focus on regions with a wage deduction limit of 960 RMB and use regions with a wage deduction limit of 800 RMB when constructing counterfactual densities (after properly controlling for the individual income tax effect). Our analysis sample contains

[4]According to the law of individual income tax, a certain level of income can be exempted from monthly taxable income. Specifically, the basic exemption was set at 800 RMB in September 1980 and was increased to 1600 RMB in January 2006, 2000 RMB in March 2008, 3500 RMB in September 2011, and 5000 RMB in October 2010.

757,333 manufacturing firms, of which 665,299 are DEs and the rest are FIEs. Given that FIEs are not subject to our focal policy, we use them as the counterfactual control group in the placebo test and the identification.

With the basic information and balance sheet information, we calculate the average monthly wage per worker by dividing the total wage bill by 12 months and by total employment. To investigate the effect of employment distortion, we classify firms' employment into two categories based on education level, i.e., skilled labor and unskilled labor. Specifically, skilled labor includes workers with a high school education or above, whereas unskilled labor includes workers with a junior secondary school education or below. For the measurement of capital, we use total capital. We calculate unemployment insurance per worker, employee benefit per worker, and administrative cost per worker as the ratios of the firm's total corresponding spending to total employment.

Table 2 shows the summary statistics for the DEs in our analysis sample. The average per worker monthly wage was approximately 838 RMB on average. The average number of employees was approximately 44, and the skilled over unskilled labor ratio was, on average, 0.98. The logarithm of total capital was, on average, 7.21. Meanwhile, on average, DEs spent 315 RMB, 301 RMB, and 2,589 RMB per worker on unemployment insurance, employee benefit, and administrative cost, respectively. Moreover, the distributions for employment, the labor ratio, and the spending per worker on unemployment insurance, employee benefits, and administrative cost were heavily right skewed.

[Insert Table 2 Here]

3 Reduced-Form Evidence

According to our focal policy, the tax schedule becomes nonlinear with the wage deduction limit. Therefore, DEs with an average monthly wage per worker higher than the deduction limit may have incentive to manipulate their average monthly wage per worker downwards to avoid paying higher taxes. This potential wage manipulation provides us with an identification opportunity to study the behavioral responses of firms to tax policies. Before we lay out our quantification framework in the next section, we first examine whether firms comply with the policy, i.e., the manipulation of the average monthly wage per worker. We then use the methodology developed by Diamond and Persson (2017) to present evidence of how firms respond to the wage deduction limit policy.

3.1 Bunching Evidence

Figure 2 presents the density distribution of the average monthly wage per worker for DEs with 960 RMB as the wage deduction limit. The solid curve plots the observed density, with the rounding patterns around multiples of 100 RMB for monthly wage and multiples for 500 RMB and 1,000 RMB for annual wage removed. Clearly, a sharp jump in the distribution occurs around the kink point of 960 RMB, implying DEs' strong responses to the wage deduction limit in the implemented nonlinear tax schedule.

[Insert Figures 2 Here]

One concern with Figure 2 is whether the bunching around the kink point reflects firms' behavioral responses to the wage deduction limit or a spurious correlation due to other unobserved factors. To alleviate this concern, we conduct several placebo tests; that is, we assess whether there is any bunching behavior at the location when the wage deduction limit policy is not in effect. First, we explore the distribution pattern for FIEs, who can fully deduct wage bills. Figure 3A shows no spike around the corresponding kink point of 960 RMB, which is consistent with the linear tax schedule for FIEs. Second, we study the distribution pattern of DEs located in regions with an 800 RMB limit in 2004. As the 960 RMB point was not the threshold for the wage deduction in these regions, we should not find any bunching at 960 RMB if only our focal policy was in place. As shown in Figure 3B, there is no jump at 960 RMB in the observed distribution of the average monthly wage per worker for DEs with 800 RMB as the wage deduction limit. Third, we use the second Economic Census conducted by China's NBS in 2008, in which all firms (including DEs) were allowed to fully deduct total wage bills from their taxable corporate income. These data are expected to have no bunching at 960 RMB for DEs, which is confirmed in Figure 3C. Combined, these results reinforce our findings that firms respond to the wage deduction limit policy by manipulating their average monthly wage per worker. In the following sections, we investigate how firms manipulate and the economic consequences of such manipulation.

[Insert Figures 3A-3C Here]

3.2 Estimation Framework

To investigate how firms respond to the wage deduction limit policy, we follow the methodology of Diamond and Persson (2017), which is based on the premise that the relationship between the variables of interest and the average per worker monthly wage would be smooth when no wage deduction limit is imposed. Hence, we can use the domain of firms not responding to the policy to estimate the relationship between variables of interest and

the average per worker monthly wage. Then, by applying this estimated relationship to the domain of responding firms, we can back out the counterfactual values of the variables of interest. The differences between the counterfactual and observed values of responding firms represent the policy effect.

Specifically, we first group the studied DEs into average wage bins of 5 and estimate the counterfactual outcomes of interest in the manipulation region around the kink point $[w_{lower}, w_{upper}]$ by fitting a third-order polynomial to the data of firms outside the region:

$$y_j = \sum_{i=0}^{3} \theta_i (w_j)^i + \sum_{r \in R, 12R} \theta_r I_{\left\{\frac{w_j}{r} \in \mathbb{N}\right\}} + \sum_{r \in 12R'} \theta' I_{\left\{\frac{w_j}{r} \in \mathbb{N}\right\}} + \varepsilon_j, \tag{1}$$

where y_j is the average value of the outcome variable in wage bin j; w_j is the average per worker monthly wage relative to the kink in terms of wage bins; and ε_j denotes the error term. In addition, to contain the reference point effects of integer wage rates, we add $R = \{1000, 1100, 1200\}$ to control for monthly wage rounding and $12R = \{11K, 12K, 13K, 14K\}$ for yearly wage rounding to 1K multiples, where $\mathbb{N}$ is the set of natural numbers. See Kleven and Waseem (2013) for the same practice. In addition, as the kink point itself is a yearly wage rounding point (i.e., 11.5K), we add $12R' = \{12.5K, 13.5K, 14.5K\}$ to control for yearly wage rounding to 0.5K multiples, and use the average of these controls to estimate the scale of rounding at 11.5K.

With the estimated coefficients from equation (1), we calculate the counterfactual relationship between the average wage and variables of interest inside the manipulation region $[w_{lower}, w_{upper}]$ as $\hat{y}_j(w_j, \hat{\theta}) = \sum_{i=0}^{3} \hat{\theta}_i (w_j)^i + \sum_{r \in R, 12R} \hat{\theta}_r I_{\{\frac{w_j}{r} \in \mathbb{N}\}} + + \sum_{r \in 12R'} \hat{\theta}' I_{\left\{\frac{w_j}{r} \in \mathbb{N}\right\}}$.

Second, we compute the average counterfactual values of the outcome variables inside the region $[w_{lower}, w_{upper}]$ with the following equation

$$\begin{aligned}
&E(y_j(w_j, \theta) | w_j \in [w_{lower}, w_{upper}], \text{no wage deduction limit}) \\
&= \int_{w_{lower}}^{w_{upper}} E(y_j(w_j, \theta) | \text{no wage deduction limit}) \\
&\times \frac{Pr(w_j | \text{no wage deduction limit})}{\int_{w_{lower}}^{w_{upper}} Pr(w_s | \text{no wage deduction limit}) dw_s} dw_j \\
&= \int_{w_{lower}}^{w_{upper}} \hat{y}_j(w_j, \hat{\theta}) \times \frac{\widehat{c}(w_j)}{\int_{w_{lower}}^{w_{upper}} \widehat{c}(w_s) dw_s} dw_j,
\end{aligned}$$

where $\widehat{c}(w_j)$ represents the counterfactual number of firms in wage bin j when no deduction limit is imposed.

Third, we estimate the intention to treat (IIT) effect as follows

$$\begin{aligned} ITT &= E(y_j(w_j,\theta)|w_j \in [w_{lower}, w_{upper}], \text{with wage deduction limit}) \\ &\quad - E(y_j(w_j,\theta)|w_j \in [w_{lower}, w_{upper}], \text{no wage deduction limit}) \\ &= \frac{\sum_{w_j \in [w_{lower}, w_{upper}]} y_j(w_j,\theta)}{N_{w_j \in [w_{lower}, w_{upper}]}} - \int_{w_{lower}}^{w_{upper}} \hat{y}_j(w_j,\hat{\theta}) \times \frac{\widehat{c}(w_j)}{\int_{w_{lower}}^{w_{upper}} \widehat{c}(w_s) dw_s} dw_j, \end{aligned} \tag{2}$$

where $N_{w_j \in [w_{lower}, w_{upper}]}$ denotes the total number of firms with the observed average wage ranging from w_{lower} to w_{upper}.

A crucial element to obtain the ITT estimates is the counterfactual density of the average per worker monthly wage under the linear tax schedule when DEs can fully deduct wage bills from taxable income. To this end, we first follow the empirical framework of Chetty et al. (2011) to estimate the counterfactual density from the observed density. The estimation methodology relies on the assumptions that the density of the average wage would be smooth in the absence of the wage deduction limit and that firms with an average wage much larger than the deduction limit face a very high adjustment cost and hence would not respond to the imposed wage deduction limit. With these assumptions, we estimate the counterfactual density by excluding observations in the region around the kink point and fitting a polynomial to the observed counts in each bin with the condition that the excess bunching mass equals the missing mass. Specifically, the estimation equation is

$$c(w_j) = \sum_{i=0}^{q} \beta_i (w_j)^i + \sum_{r \in R, 12R} \rho_r I_{\{\frac{w_j}{r} \in \mathbb{N}\}} + \sum_{r \in 12R'} \rho' I_{\left\{\frac{w_j}{r} \in \mathbb{N}\right\}} + \sum_{i=w_{lower}}^{w_{upper}} \gamma_i I_{\{w_j = i\}} + \epsilon_j, \tag{3}$$

where $c(w_j)$ is the number of firms in wage bin j; q is the order of the polynomial; and $[w_{lower}, w_{upper}]$ denotes the width of the excluded region around the kink point (i.e., the fraction of firms choosing to bunch at the kink point). Following Diamond and Persson (2017), we choose the values of q, w_{lower}, and w_{upper} based on a 5-fold cross-validation method. To address the problem of reference point effects, we add controls for integer wage levels, as in equation (1).

After obtaining the estimated coefficients from equation (3), we calculate the counterfactual density distribution as $\hat{c}(w_j) = \sum_{i=0}^{q} \hat{\beta}_i (w_j)^i + \sum_{r \in R, 12R} \hat{\rho}_r I_{\{\frac{w_j}{r} \in \mathbb{N}\}} + \sum_{r \in 12R'} \hat{\rho'} I_{\left\{\frac{w_j}{r} \in \mathbb{N}\right\}}$. The excess mass of bunching is estimated as $\widehat{B} = \sum_{w_{lower}}^{\bar{w}} (c(w_j) - \widehat{c}(w_j))$, the missing mass is $\widehat{M} = \sum_{\bar{w}+1}^{w_{upper}} (\widehat{c}(w_j) - c(w_j))$, and the normalized bunching mass is defined as $\hat{b} = \widehat{B} / (\sum_{w_j \in [-4,5]} \hat{c}(w_j)/10)$. The red dotted curve in Figure 2 plots the estimated counterfactual density. Specifically, we set the excluded region as [930, 1070], normalize the average wage with respect to the kink point 960 RMB, employ a third-order polynomial, and control for the wage rounding points.

The above polynomial fitting approach to construct a counterfactual density relies on two assumptions: (1) the proper formulation of the polynomial function; and (2) the non-manipulation region is a good counterfactual for the manipulation region. As a robustness, we use an alternative approach to construct the counterfactual density. Specifically, in our research setting, we have groups of firms that were not subject to the wage deduction limit of 960 RMB, and their density distribution can be used to construct a counterfactual density for firms subject to the policy. First, FIEs can fully deduct wage bills, and as shown in Figure 3A, there is no significant jump at 960 RMB for these firms. Hence, to improve the comparability, we use FIEs located in the regions with a wage deduction limit of 960 RMB to construct a counterfactual density for DEs located in these regions. Specifically, we choose an arbitrary adjustment degree d, construct a density distribution by multiplying the density of FIEs by d, and calculate the squared sum of differences between the constructed density and the observed density of DEs in all wage bins outside the exclusion region. The counterfactual density is chosen as the one minimizing the squared sum of differences.

Second, DEs located in the regions with a wage deduction limit of 800 RMB were not subject to the cutoff of 960 RMB and can be used to construct a counterfactual density distribution. We follow a similar procedure as before and select the counterfactual density as the one minimizing the difference between the constructed density and the observed density in all wage bins outside the exclusion region.

However, one may be concerned that FIEs are different from DEs and that regions are different, and hence, the density estimated in the above two samples of firms may not represent a good counterfactual. To improve the comparability, we use a method in the spirit of DD analysis; that is, we first construct the difference of the density distribution around 960 RMB between DEs located in regions with a wage deduction limit of 960 RMB and DEs in regions with a wage deduction limit of 800 RMB and then compare the result with the corresponding difference of FIEs between these regions. This double difference can help us control for the differential distributions between FIEs and DEs and also those between different regions.

With the estimated counterfactual density (from two different approaches), we can obtain the ITT estimates from equation (2). Standard errors are estimated using a parametric bootstrap procedure. Specifically, following Chetty et al. (2011), we redraw the estimated vector of errors ε_j in equation (1) with replacement to generate a new sample and calculate a new ITT estimate. We repeat this procedure 200 times and obtain the standard error of the estimate as the standard deviation of the 200 new estimates.

3.3 Manipulation Evidence

The bunching evidence in Figure 2 suggests that DEs respond to the wage deduction limit policy by adjusting their average monthly wage per worker. We then investigate what variables firms manipulate. One possible manipulation is that firms relabel part of labor wage as unemployment insurance, employee benefit, or administration cost, which is deductible from CIT taxable income. Hence, the effective payments that workers receive do not change. To investigate this potential manipulation, we estimate the ITT effects of the imposed wage deduction limit on firm expenditure on unemployment insurance, employee benefit, and administration cost: the results are summarized in columns (1)-(3) of Table 3. All the three ITT estimates are statistically insignificant, regardless of the approach employed to construct the counterfactual density, suggesting that firms did not respond to the limit by relabeling.

[Insert Table 3 Here]

Second, there is some anecdotal evidence that firms inflate employment in response to the wage deduction limit by adding *ghost workers*; that is, workers who exist on the books but not in reality..[5] This helps to lower the average monthly wage per worker to achieve the threshold for tax reduction. Without data on the number of ghost workers, it is difficult to directly examine this potential manipulation. Instead, we hypothesize that if firms responded in this way, the real firm operation would not be significantly affected; hence, key performance indicators would not change. To this end, we examine the policy effect on firm capital: the results are reported in column (4) of Table 3. The ITT estimates for capital are about -0.3079 to -0.7980 and are statistically significant at 1% level, which suggests that firms did respond to the wage deduction limit policy by decreasing firm capital, alleviating the concern that firms only manipulated the books.

Third, firms could change their employment structure to lower the average monthly wage per worker. Specifically, they can substitute skilled labor with unskilled labor, as the latter receives lower wages. In column (5) of Table 3, we investigate the policy effect on employment structure, that is, the ratio of skilled over unskilled labor. We find estimated coefficients ranging from -0.2608 to -0.6001, which are statistically significant at 1% level.

In summary, we find that the wage deduction limit policy changed firms' capital and reported employment structure. However, without direct information, we cannot fully rule out the possibility that firms may manipulate the booking of employment. Hence, the coefficient from the reduced form estimation may indicate an upper bound for the real

[5]For example, in 2005, tax auditors in Henan province found one mining company manipulating the employment level by comparing its monthly reported number of employees and checking the payment summaries. The company finally confessed the tax evasion and was fined 50 percent the amount of tax evaded, totally 390,076.5 RMB.

response to the policy in the margin of employment structure adjustment. In the following sections, when we quanify firm adjustment in a response to tax policies, we will incorporate the manipulation of employment booking and identify the margin of employment structure adjustment.

4 Quantification Model

To capture the behavioral responses by firms to the wage deduction limit policy and quantify the role of employment adjustment, we develop a stylized model of employment decisions by firms. We first do not consider that firms can inflate the booking number of workers and focus on the optimal adjustment of the employment structure. In section 4.5, we will extend the model to incorporate the possibility that firms can manipulate the booking by adding some ghost workers.

4.1 Setup

We assume a monopolistic competitive market with the representative consumer utility function being

$$U = \left(\int_{j \in J} q_j^{\beta} dj \right)^{\frac{1}{\beta}}, \tag{4}$$

where U denotes the utility level; q_j denotes the consumption units of firm j's product; J denotes the full set of firms in the market; and $\beta \in (0,1)$ is the substitution parameter.

Hence, for firm j, the demand function is $p_j = q_j^{-(1-\beta)} P Q^{1-\beta}$, where $P = \left(\int_{j \in J} p_j^{-\frac{\beta}{1-\beta}} dj \right)^{-\frac{1-\beta}{\beta}}$ denotes the price index and $Q = \left(\int_{j \in J} q_j^{\beta} dj \right)^{\frac{1}{\beta}}$ denotes the quantity index.

The production function of firm j is assumed to be

$$q_j = K_j^{\alpha_s} \left[(\lambda_j H_j^{\frac{\sigma-1}{\sigma}} + (1-\lambda_j) L_j^{\frac{\sigma-1}{\sigma}})^{\frac{\sigma}{\sigma-1}} \right]^{1-\alpha_s}, \tag{5}$$

where K_j, H_j, and L_j denote capital, skilled labor, and unskilled labor, respectively; α_s denotes the share of capital in the total output in sector s; and λ_j is a factor augmenting technology term of skilled labor that varies across firms to generate firm heterogeneity. σ is our parameter of interest, capturing the elasticity of substitution between skilled and unskilled labor.[6]

[6]Based on the CES-in-CD production functional form, the ratio of capital and the combined labor inputs (i.e., $CL_j = (\lambda_j H_j^{\frac{\sigma-1}{\sigma}} + (1-\lambda_j) L_j^{\frac{\sigma-1}{\sigma}})^{\frac{\sigma}{\sigma-1}}$) is constant at $\alpha_s/(1-\alpha_s)$. Hence, the complementarity between capital and skilled or unskilled labor depends on the consequential change in the combined labor input. The reduced form empirical results in section 3.3 shows that firms responded to the tax incidence by substituting

4.2 Optimal Decision

We start with the counterfactual linear tax schedule, in which firms can fully deduct wage bills. The profit function is written as

$$(\pi_j)_{ct} = (1-\tau)\left(p_j q_j - w_H H_j - w_L L_j - rK_j\right), \tag{6}$$

where w_H, w_L, and r denote the input prices of skilled labor, unskilled labor, and capital, respectively;[7] and τ denotes the corporate tax rate. Additionally, $w_j^r \equiv \frac{w_H H_j + w_L L_j}{H_j + L_j}$ denotes the average per worker monthly wage level.

Maximizing the profit function, we obtain the optimal decision as

$$\frac{(H_j)^*_{ct}}{(L_j)^*_{ct}} = \left[\frac{w_L}{w_H}\frac{\lambda_j}{1-\lambda_j}\right]^{\sigma}, \tag{7}$$

and

$$(w_j^r)^*_{ct} = \frac{w_H\left[\frac{w_L}{w_H}\frac{\lambda_j}{1-\lambda_j}\right]^{\sigma} + w_L}{\left[\frac{w_L}{w_H}\frac{\lambda_j}{1-\lambda_j}\right]^{\sigma} + 1}. \tag{8}$$

Next, we consider the implemented nonlinear tax schedule under which firms can only deduct wage bills up to a preset limit. Let $\bar{w}$ denote the average monthly wage deduction limit. The deductible wage bill for firm j is

$$DC_j = min\{\bar{w}, w_j^r\}N_j^r, \tag{9}$$

where $N_j^r = H_j + L_j$ denotes the reported total employment.

Hence, the profit function becomes

$$\begin{aligned}\pi_j = {} & p_j q_j - (w_H H_j + w_L L_j + rK_j) - \tau(p_j q_j - rK_j - DC_j) \\ & - C \times I_{\left\{\frac{(H_j)^*_{ct}}{(L_j)^*_{ct}} > D\right\}} \times (L_j + H_j - (L_j)^*_{ct} - (H_j)^*_{ct}).\end{aligned} \tag{10}$$

To capture that firms with high ratios of skilled to unskilled labor have large costs of manipulating the employment structure, we add a fixed cost of adjustment $C \times I_{\left\{\frac{(H_j)^*_{ct}}{(L_j)^*_{ct}} > D\right\}} \times (L_j + H_j - (L_j)^*_{ct} - (H_j)^*_{ct})$, where $C = \tau\bar{w}$; and $D > (\bar{w} - w_L)/(w_H - \bar{w})$.

skilled with unskilled labor and cutting down capital level. This implies that the combined labor input declines, i.e., $\Delta CL_j = dCL_j/dH_j \times \Delta H_j + dCL_j/dL_j \times \Delta L_j < 0$. Hence, we can have $\lambda_j/(1-\lambda_j) > -(\Delta L_j/\Delta H_j) \times (H_j/L_j)^{\frac{1}{\sigma}}$

[7] There are totally 158,044 DEs with the average per worker monthly wage between 860 RMB and 1360 RMB in the bunching range. These affected firms account for about 11% of the total firms, and hence, shall not have significant influence on the input market. Therefore, in the paper, we treat the input prices as given.

For firms with $(w_j)^*_{ct} \leq \bar{w}$ (i.e., $\lambda_j \leq 1-1/\left[1+\left(\frac{\bar{w}-w_L}{w_H-\bar{w}}\right)^{\frac{1}{\sigma}}\frac{w_H}{w_L}\right] \equiv \lambda_1$), we have $(H_j)^*_{ct}/(L_j)^*_{ct} \leq (\bar{w}-w_L)/(w_H-\bar{w}) < D$. Hence, profit functions (6) and (10) are the same. Consequently, firms are unaffected by the imposed wage deduction limit and choose the same optimal solution as that under the counterfactual linear tax schedule, i.e.,

$$\frac{(H_j)^*}{(L_j)^*} = \frac{(H_j)^*_{ct}}{(L_j)^*_{ct}}, \tag{11}$$

and

$$(w_j^r)^* = (w_j^r)^*_{ct}. \tag{12}$$

Define $\bar{\lambda} \equiv \frac{w_H D^{\frac{1}{\sigma}}}{w_H D^{\frac{1}{\sigma}}+w_L}$. Given that $D > (\bar{w}-w_L)/(w_H-\bar{w})$, we have $\lambda_1 < \bar{\lambda}$. Hence, for firms with $(w_j)^*_{ct} > \bar{w}$ (i.e., $\lambda_j > \lambda_1$), when $\lambda_1 < \lambda_j \leq \bar{\lambda}$, we have $(H_j)^*_{ct}/(L_j)^*_{ct} \leq D$, and the profit function becomes

$$\pi_j = (1-\tau)\left(p_j q_j - w_H H_j - w_L L_j - rK_j\right) - I_{\{w_j^r > \bar{w}\}}\tau(w_j^r - \bar{w})N_j^r. \tag{13}$$

In this scenario, firms have two options for the average monthly wage rate. First, firms can set it at $\bar{w}$. Second, firms can solve the maximization of the new profit function (13), which generates

$$\frac{(H_j)^*}{(L_j)^*} = \left[\frac{w_L-\tau\bar{w}}{w_H-\tau\bar{w}}\frac{\lambda_j}{1-\lambda_j}\right]^{\sigma}, \tag{14}$$

and

$$(w_j^r)^* = \frac{w_H\left[\frac{w_L-\tau\bar{w}}{w_H-\tau\bar{w}}\frac{\lambda_j}{1-\lambda_j}\right]^{\sigma} + w_L}{\left[\frac{w_L-\tau\bar{w}}{w_H-\tau\bar{w}}\frac{\lambda_j}{1-\lambda_j}\right]^{\sigma} + 1}. \tag{15}$$

Comparing the profits from these two options, we have when $\lambda_1 < \lambda \leq 1-1/\left[1+\left(\frac{\bar{w}-w_L}{w_H-\bar{w}}\right)^{\frac{1}{\sigma}}\frac{w_H-\tau\bar{w}}{w_L-\tau\bar{w}}\right] \equiv \lambda_2 < \bar{\lambda}$, the firms are affected by the wage deduction limit and choose to bunch at the corner solution

$$\frac{(H_j)^*}{(L_j)^*} = \frac{\bar{w}-w_L}{w_H-\bar{w}}, \tag{16}$$

and

$$(w_j^r)^* = \bar{w}. \tag{17}$$

When $\lambda_2 < \lambda \leq \bar{\lambda}$, firms are affected but do not set the average monthly wage per worker at $\bar{w}$. Instead, they choose the new optimal solution described in equations (14) and (15).

Finally, for firms with $\lambda_j > \lambda_3 \equiv \bar{\lambda}$, we have that $(H_j)^*_{ct}/(L_j)^*_{ct} > D$; hence, the profit

function is given by

$$\pi_j = (1-\tau)\left(p_j q_j - w_H H_j - w_L L_j - rK_j\right) - \tau(w_j^r - \bar{w})N_j^r - C(L_j + H_j - (L_j)_{ct}^* - (H_j)_{ct}^*). \quad (18)$$

By maximizing this profit function, we obtain the optimal solution as

$$\frac{(H_j)^*}{(L_j)^*} = \frac{(H_j)_{ct}^*}{(L_j)_{ct}^*}, \quad (19)$$

and

$$(w_j^r)^* = (w_j^r)_{ct}^*. \quad (20)$$

In other words, in this scenario, the firms are unaffected by the kink introduced into the tax schedule.

In summary, the optimal choice of the average per worker monthly wage under the nonlinear tax schedule is

$$(w_j^r)^* = \begin{cases} (w_j^r)_{ct}^* & \text{if } \lambda_j \leq \lambda_1 \\ \bar{w} & \text{if } \lambda_j \in (\lambda_1, \lambda_2] \\ \dfrac{w_H\left[\frac{w_L - \tau\bar{w}}{w_H - \tau\bar{w}}\frac{\lambda_j}{1-\lambda_j}\right]^{\sigma} + w_L}{\left[\frac{w_L - \tau\bar{w}}{w_H - \tau\bar{w}}\frac{\lambda_j}{1-\lambda_j}\right]^{\sigma} + 1} & \text{if } \lambda_j \in (\lambda_2, \lambda_3] \\ (w_j^r)_{ct}^* & \text{if } \lambda_j > \lambda_3 \end{cases}. \quad (21)$$

Figure 4 compares firms' optimal wage choices under the linear and nonlinear tax schedules.

[Insert Figure 4 Here]

4.3 Implications for Bunching

A firm who bunches at $\bar{w}$ with the lowest λ_j (i.e., $\lambda_j = \lambda_1$) chooses the same skilled to unskilled labor ratio and the same reported average per worker monthly wage under the implemented nonlinear policy as those under the counterfactual linear tax schedule, that is $\frac{(H_j)^*}{(L_j)^*}|_{\lambda_j=\lambda_1} = \frac{(H_j)_{ct}^*}{(L_j)_{ct}^*}|_{\lambda_j=\lambda_1} = \frac{\bar{w}-w_L}{w_H-\bar{w}}$ and $(w_j^r)^*|_{\lambda_j=\lambda_1} = (w_j^r)_{ct}^*|_{\lambda_j=\lambda_1} = \bar{w}$. A firm who bunches at $\bar{w}$ with the highest λ_j (i.e., $\lambda_j = \lambda_2$) chooses $\frac{(H_j)^*}{(L_j)^*}|_{\lambda_j=\lambda_2} = \frac{\bar{w}-w_L}{w_H-\bar{w}}$ and $(w_j^r)^*|_{\lambda_j=\lambda_2} = \bar{w}$ under the implemented nonlinear policy and would choose $\frac{(H_j)_{ct}^*}{(L_j)_{ct}^*}|_{\lambda_j=\lambda_2}$ and $(w_j^r)_{ct}^*|_{\lambda_j=\lambda_2}$ under the counterfactual linear policy.

Hence, all firms with $\lambda_j \in [\lambda_1, \lambda_2]$ (or $(w_j^r)_{ct}^* \in [\bar{w}, \bar{w} + \Delta w]$) bunch at $\bar{w}$ under the

nonlinear tax schedule, where

$$\begin{aligned}\Delta w &\equiv (w_j^r)^*_{ct}|_{\lambda_j=\lambda_2} - \bar{w} \\ &= \frac{w_H \left[\frac{w_L}{w_H}\frac{w_H-\tau\bar{w}}{w_L-\tau\bar{w}}\right]^{\sigma}\frac{\bar{w}-w_L}{w_H-\bar{w}} + w_L}{\left[\frac{w_L}{w_H}\frac{w_H-\tau\bar{w}}{w_L-\tau\bar{w}}\right]^{\sigma}\frac{\bar{w}-w_L}{w_H-\bar{w}} + 1} - \bar{w} \\ &= \varphi(w_L, w_H, \bar{w}, \tau, \sigma).\end{aligned} \tag{22}$$

Therefore, the excess fraction of bunching is given by

$$B = \int_{\lambda_1}^{\lambda_2} f(\lambda_j)d\lambda_j = \int_{\bar{w}}^{\bar{w}+\Delta w} \hat{c}((w_j^r)^*_{ct})d(w_j^r)^*_{ct} \simeq \hat{c}(\bar{w})\Delta w,$$

where $\hat{c}((w_j^r)^*_{ct})$ denotes the density distribution of the reported average monthly wage under the counterfactual tax schedule; and the second approximation is based on the assumption (as in Saez (2010) and Chetty et al. (2011)) that the counterfactual density $\hat{c}((w_j^r)^*_{ct})$ is uniform around the deduction limit $\bar{w}$. We then have that $\Delta w = B/\hat{c}(\bar{w})$. Hence, we can estimate σ as a function of observable parameters $(w_L, w_H, \bar{w}, \tau, r)$ and the empirically estimable variable $\widehat{\Delta w}$.

In addition, by comparing the ratio H/L for the firms affected by the nonlinear tax schedule, we obtain that

$$\frac{(H_j)^*_{ct}}{(L_j)^*_{ct}} = \left[\frac{w_L}{w_H}\frac{\lambda_j}{1-\lambda_j}\right]^{\sigma} > \frac{(H_j)^*}{(L_j)^*} = \left[\frac{w_L-\tau\bar{w}}{w_H-\tau\bar{w}}\frac{\lambda_j}{1-\lambda_j}\right]^{\sigma}.$$

Thus, the introduction of the wage deduction limit reduces the relative employment of skilled labor, which is consistent with the reduced-form evidence in Section 3.3.

4.4 Estimation

We estimate the previous model using the 2004 Economic Census. Specifically, we set the interest rate $r = 5.58\%$ in 2004 according to the World Bank. Meanwhile, we estimate wage rates w_L and w_H from the census data using the following two conditions:

$$w_L * \sum_j L_j + w_H * \sum_j H_j = \sum_j \text{total\_wage\_bill}_j \tag{23}$$

and

$$Median\left(\frac{H_j}{L_j}\right) = \frac{\bar{w}-w_L}{w_H-\bar{w}} \tag{24}$$

for firms bunching around the kink point 960 RMB. We obtain $w_L = 867.65$ and $w_H = 1252.44$. For the corporate tax rate τ, we use the mean effective tax rate of the studied DEs. The counterfactual density distribution of the average monthly wage per worker (that is, the distribution in the scenario when DEs can fully deduct wage bills from their taxable corporate income) is estimated using both the parametric estimation à la (Chetty et al. (2011)) and the nonparametric differencing estimation approach elaborated in Section 3.2.

With the estimated counterfactual density distribution and values of key parameters $(w_L, w_H, \bar{w}, \tau, r)$, we can estimate $\widehat{\Delta w}$ and hence the elasticity $\hat{\sigma}$ from equation (22). The results are summarized in panel A of Table 4. The first row presents the results with the counterfactual density constructed using the empirical framework of Chetty et al. (2011)). The last three rows present the results with the counterfactual density estimated by the nonparametric differencing approaches. Results show that $\widehat{\Delta w}$ is around 12.5880 to 27.3043 and the elasticity $\hat{\sigma}$ ranges from 1.1548 to 2.4002. This implies that when the relative wage of skilled to unskilled labor increases by 1%, firms decrease the ratio of skilled to unskilled labor by about 1.1548% to 2.4002%.

[Insert Table 4 Here]

In the baseline, we estimate w_L and w_H using the model moments. To assess the sensitivity of our results to wage rates, we calculate wage rates from the 2005 Chinese Mini Population Census. Specifically, we use the mean wage rates of employees with education lower than senior secondary school and those with education higher than junior secondary school to calculate w_L and w_H, respectively: we obtain $w_L = 860$ and $w_H = 1250$. The results obtained using the wage rates calculated from the 2005 Mini Population Census are reported in panel B of Table 4, replicating the analyses in panel A: $\hat{\sigma}$ ranges from 1.2344 to 2.5926.

Comparison with the existing literature. The literature on the estimation of the elasticity of substitution between skilled and unskilled labor can be traced back to Katz and Murphy (1992). Specifically, they estimate the substitution elasticity between college and high school labor as 1.41 using the U.S. March Current Population Surveys (CPS) data from 1967-1987. Katz and Autor (1999) include demand shifts as controls and find that the elasticity ranges from 1 to 3. Krusell et al. (2000) use an alternative definition of skilled labor as workers with at least some college education and obtain a moderately higher elasticity estimate of 1.89. Card and Lemieux (2001) include an aggregate relative supply index and age-group specific relative supplies of college workers in the model of Katz and Murphy (1992) and find that the elasticity of substitution between college and high school labor is approximately 2.2 to 2.5 in both the U.S. and the U.K. Autor et al. (2008) use the 1968-2005 CPS data

to re-estimate the elasticity with the model of Katz and Murphy (1992) and show that the elasticity of substitution between college and high school equivalents is 2.43 when no trend break in the annual growth rate in 1992 is added.

The estimated elasticity in this paper is comparable to the estimates from the U.S. One difference from most previous studies is that our work focuses on the substitution between the high school and non-high school graduates. One study similar to ours is by Angrist (1995), who uses Palestinian data and finds the elasticity of substitution between workers with sixteen years of schooling and those with less than twelve years of schooling to be approximately 2.

4.5 An Extension with the Booking Manipulation

In response to the wage deduction limit, firms may manipulate their reported employment level without effectively changing their operation. This manipulation behavior is likely to occur when tax administration and enforcement are weak or the penalties for tax evasion are not tough. In the studied period in China, tax auditing was not strict despite of significant penalties.[8] Typically, tax auditors verify the reported employment level by looking at firms' employee rosters, labor attendance sheets, payment summaries, social security payment records, etc. Hence, there are some room for firms to fabricate relevant documents and payment records to manipulate the employment level.

In Section 3.3, we present some reduced-form evidence that firms did have real responses to the policy, e.g., an increase in capital. However, this may not fully rule out the role of the booking manipulation in generating the bunching response and, hence, possibly bias our estimates in the previous analysis. To this end, in this section, we extend our aforementioned theoretical model to incorporate the possibility that firms may inflate the reporting of employment.

Specifically, we change N_j^r in equation (9) to $N_j^r = H_j + L_j + L_j^m$. Without loss of generality and for simplicity of the calculation, we define $L_j^m = L_j \times \eta_j$ as firm j's manipulated number of unskilled employment, where η_j is the manipulation degree variable. Additionally, $w_j^r = \frac{w_H H_j + w_L L_j}{N_j^r}$.

With this change, the profit function becomes

$$\begin{aligned}\pi_j = p_j q_j - (w_H H_j + w_L L_j + rK_j) - \tau(p_j q_j - rK_j - DC_j) - c(L_j, \eta_j) \\ - C \times I_{\left[\frac{(H_j)^*_{ct}}{(L_j)^*_{ct}} > D\right]} \times (L_j + L_j^m + H_j - (L_j)^*_{ct} - (L_j^m)^*_{ct} - (H_j)^*_{ct}),\end{aligned} \tag{25}$$

[8] According to the law of the People's Republic of China on the administration of tax collection, tax payers who forged accounting books or overstated expenses would be fined not less than 50 percent but not more than five times the amount of tax he fails to pay or underpays.

where $c(L_j, \eta_j) = \frac{c}{2}\eta_j^2 L_j$ denotes the cost of the booking manipulation[9] and $(L_j^m)_{ct}^*$ denotes firm j's optimal choice of the manipulated number of unskilled employment under the counterfactual linear tax schedule.

The optimal choices under the implemented nonlinear tax schedule can then be derived as

$$\left((w_j^r)^*, \eta_j^*\right) = \begin{cases} \left((w_j^r)_{ct}^*, 0\right) & \text{if } \lambda_j \leq \lambda_1' \\ \left(\bar{w}, [0, \frac{\tau\bar{w}}{c}]\right) & \text{if } \lambda_j \in (\lambda_1', \lambda_2'] \\ \left(\frac{w_H \left[\frac{w_L - \tau\bar{w} - \frac{\tau^2\bar{w}^2}{2c}}{w_H - \tau\bar{w}} \frac{\lambda_j}{1-\lambda_j}\right]^\sigma + w_L}{\left[\frac{w_L - \tau\bar{w} - \frac{\tau^2\bar{w}^2}{2c}}{w_H - \tau\bar{w}} \frac{\lambda_j}{1-\lambda_j}\right]^\sigma + 1 + \frac{\tau\bar{w}}{c}}, \frac{\tau\bar{w}}{c} \right) & \text{if } \lambda_j \in (\lambda_2', \lambda_3'] \\ \left((w_j^r)_{ct}^*, 0\right) & \text{if } \lambda_j > \lambda_3'. \end{cases} \tag{26}$$

where $\lambda_1' \equiv 1 - 1/\left(1 + \left[\frac{\bar{w} - w_L}{w_H - \bar{w}}\right]^{\frac{1}{\sigma}} \frac{w_H}{w_L}\right)$; $\lambda_2' \equiv 1 - 1/\left(1 + \left[\frac{\bar{w}(1+\frac{\tau\bar{w}}{c}) - w_L}{w_H - \bar{w}}\right]^{\frac{1}{\sigma}} \frac{w_H - \tau\bar{w}}{w_L - \tau\bar{w} - \frac{\tau^2\bar{w}^2}{2c}}\right)$; and $\lambda_3' \equiv \frac{w_H D^{\frac{1}{\sigma}}}{w_H D^{\frac{1}{\sigma}} + w_L}$. The comparison of firms' optimal choices of the average per worker monthly wage and the manipulation degree are presented in Figure 5.

[Insert Figure 5 Here]

The bunching firm with the lowest $\lambda_j = \lambda_1'$ chooses the same $(w_j^r)^*|_{\lambda_j = \lambda_1'} = (w_j^r)_{ct}^*|_{\lambda_j = \lambda_1'} = \bar{w}$ under both the implemented nonlinear and the counterfactual linear tax schedules. The optimal solution of the bunching firm with the highest $\lambda_j = \lambda_2'$ under the counterfactual linear tax schedule is

$$(w_j^r)_{ct}^*|_{\lambda_j = \lambda_2'} = \frac{w_H \left[\frac{w_L}{w_H} \frac{\lambda_2'}{1-\lambda_2'}\right]^\sigma + w_L}{\left[\frac{w_L}{w_H} \frac{\lambda_2'}{1-\lambda_2'}\right]^\sigma + 1}. \tag{27}$$

Hence, all firms with $\lambda_j \in [\lambda_1', \lambda_2']$ (or the counterfactual reported average monthly wage $(w_j^r)_{ct}^* \in [\bar{w}, \bar{w} + \Delta w']$) bunch around $\bar{w}$ under the nonlinear tax schedule, where

$$\Delta w' = \varphi(w_L, w_H, \bar{w}, \tau, c, \sigma). \tag{28}$$

[9]The cost of manipulation might come from fabrication of relevant documents, social security payments to be paid, monetary penalties incurred if caught cheating, etc. Since the forging costs were much smaller compared with the potential tax reduction, some firms did take risk to manipulate the reported employment level and hence reduced the employment adjustment. As a result, the costs of manipulation affected firms' employment decision indirectly. However, as social security payments and administration costs were deductible from the tax payments, manipulation costs were unlikely to affect the firms' employment decisions too much directly.

The excess fraction of bunching is

$$B = \int_{\lambda_1'}^{\lambda_2'} f(\lambda_j) d\lambda_j = \int_{\bar{w}}^{\bar{w}+\Delta w'} \hat{c}((w_j^r)_{ct}^*) d(w_j^r)_{ct}^* \simeq \hat{c}(\bar{w}) \Delta w', \tag{29}$$

Combining equations (28) and (29), we have two unknown parameters c and σ, which require two empirical moments to pin down simultaneously. The bunching around 960 RMB of DEs located in regions with a wage deduction limit of 960 RMB provides one moment. For the other, we resort to DEs located in regions with a wage deduction limit of 800 RMB.

The estimation results of σ and $\eta = \frac{\tau \bar{w}}{c}$ are presented in Table 5. The baseline analysis relies on Chetty et al.'s (2011) method to estimate the counterfactual density distribution of the average monthly wage per worker. Results are reported in the first two rows for DEs in regions with a wage deduction limit of 960 RMB and in regions with a wage deduction limit of 800 RMB, respectively. $\hat{\eta} = 5.95\%$ for the former regions and $\hat{\eta} = 4.48\%$ for the latter regions, implying that affected firms with a 960 RMB limit and an 800 RMB limit reported 5.95% and 4.48% more unskilled labor as ghost workers, respectively.[10] Incorporating the possibility of the booking manipulation, we estimate the substitution elasticity as $\hat{\sigma} = 1.2607$, approximately 31% smaller than the baseline estimate without the booking manipulation in Table 4.

[Insert Table 5 Here]

However, the bunching around 800 RMB may reflect not only firms' responses to the deduction limit threshold of CIT but also the responses to the exemption threshold of individual income tax, as these two thresholds coincide in our studied period. Hence, in the rest of the table, we use the alternative estimation methods of the counterfactual density distribution as discussed in Section 3.2 and mimic the analyses in Table 4. Specifically, as the control groups for the counterfactual, we employ the corresponding FIEs in the same regions, the DEs at the 960 RMB kink point in regions with an 800 RMB deduction limit, and the DD design using both of these two samples, respectively. The elasticity estimates $\hat{\sigma}$ and manipulation estimates $\hat{\eta}$ remain stable, ranging from 1.1803 to 1.2821 and 3.20% to 8.52%, respectively.

[10] Given the not very strict tax auditing environment and the relatively high penalties, the estimated booking manipulation falls within a reasonable range.

5 Counterfactual Analysis

In this section, we conduct counterfactuals to quantify the role of employment adjustment in the distortion effect of tax incidences and investigate the welfare consequences of the wage deduction limit policy and payroll tax incidence.

5.1 Role of Employment Adjustment

To understand the role of employment adjustment in transmitting the tax incidence, we conduct several counterfactuals in this subsection. Specifically, we first compute the GDP outcome for the counterfactual when there is no wage deduction limit policy (referred to as benchmark). Next, we calculate the same outcome for our observed scenario; that is, the one with the wage deduction limit policy in place. The difference between these two generates the consequences of our focal policy. Finally, we consider another counterfactual in which the focal policy was in effect but firms cannot adjust their employment structures (i.e., the employment structure was fixed at the benchmark). Comparing the change from the benchmark to the observed scenario with the change from the benchmark to the counterfactual without employment adjustment, we can then quantify the role of employment adjustment in the effect of tax incidence.

Specifically, under the benchmark scenario (i.e., a linear tax schedule in which all firms can fully deduct wage bills), all firms' profit functions are presented by equation (6). We can solve, for each firm, the optimal K, H, and L as functions of λ (i.e., $K^B(\lambda_j)$, $H^B(\lambda_j)$, and $L^B(\lambda_j)$). Consequently, the baseline aggregate GDP for regions with a wage deduction limit of 960 RMB is

$$GDP^B = \left[\int_{\lambda_j} \left(K_j^B(\lambda_j)^{\alpha_s} \left[(\lambda_j H_j^B(\lambda_j)^{\frac{\sigma-1}{\sigma}} + (1-\lambda_j) L_j^B(\lambda_j)^{\frac{\sigma-1}{\sigma}})^{\frac{\sigma}{\sigma-1}} \right]^{1-\alpha_s} \right)^{\beta} d\lambda_j \right]^{\frac{1}{\beta}}, \quad (30)$$

Next, under the observed scenario, DEs are subject to the wage deduction limit, and their profit functions are presented by equation (10), whereas FIEs' profit functions are captured by equation (6). Hence, we can solve a new set of K, H, and L as functions of λ_j. The aggregate GDP for regions with a 960 RMB deduction limit is calculated as

$$GDP = \left[\int_{\lambda_j} \left(K_j(\lambda_j)^{\alpha_s} \left[(\lambda_j H_j(\lambda_j)^{\frac{\sigma-1}{\sigma}} + (1-\lambda_j) L_j(\lambda_j)^{\frac{\sigma-1}{\sigma}})^{\frac{\sigma}{\sigma-1}} \right]^{1-\alpha_s} \right)^{\beta} d\lambda_j \right]^{\frac{1}{\beta}}. \quad (31)$$

α_s is computed by dividing the aggregate capital cost by the aggregate total input costs

in the domestic and foreign firm set s, i.e.,

$$\alpha_s = \frac{(1-\tau)rK_s}{(1-\tau)rK_s + p_H H_s + p_L L_s}.$$

We set $r = 5.58\%$; $\tau_j = 0.18,\ 0.27,\ or\ 0.33$ for DEs conditional on the firms' taxable income; $\tau_j = 0.15,\ 0.24,\ or\ 0.33$ for FIEs conditional on the firms' registration location and main business; and $w_L = 867.65$ and $w_H = 1252.44$. The distribution of λ_j can be inversely calculated from the estimated counterfactual distribution and the corresponding elasticity $\sigma = 1.1803$ estimated from the extended model in Section 4.5, given equation (8). β is set at 0.75, a central value in the range of estimates used in the previous literature (for a review, see Head and Mayer, 2014).

Panel A of Table 6 presents the percentage change from the benchmark to the observed scenario, i.e., $\left(GDP^O - GDP^B\right)/GDP^B$. As shown in column (8), the wage deduction limit policy caused aggregate GDP to decline by approximately 0.2872 percent, which accounts for 21.82 billion RMB for regions with a 960 RMB limit with reference to the published provincial and city-level GDPs in 2004.[11] As illustrated in the model, firms responded to the policy by manipulating the reported employment level and substituting skilled labor with unskilled labor. The manipulation allowed tax reduction and hence stimulated production. However, the labor substitution reduced firm productivity, which dominated the stimulation effect of the booking manipulation and leaded to a decrease in aggregate GDP.

[Insert Table 6 Here]

Finally, to understand the role of employment adjustment, we consider a counterfactual where the focal policy was in place but firms were not allowed to adjust their employment structure. Specifically, DEs maximize the profit function captured by equation (10) with the constraint that

$$\frac{(H_j)^*}{(L_j)^*} = \frac{(H_j)^*_{ct}}{(L_j)^*_{ct}}. \tag{32}$$

We can then solve the optimal K, H, and L as functions of λ for each firm and calculate GDP using equation (31). Similarly, we calculate the percentage change from the benchmark to the concerned counterfactual using the same parameters and distribution of λ.

The results are presented in panel B of Table 6. GDP increased by 0.8863 percent. One explanation might be that without the ability of labor substitution, firms enjoyed reduction in corporate taxes by manipulating the reported employment levels and hence expanded production. Therefore, the manipulation margin eases the effect of tax incidence, whereas

[11] According to the China Statistical Yearbook (2005) and provincial statistical yearbooks, the aggregate GDP for regions with a 960 RMB limit is approximately 7,599 billion RMB.

the adjustment of employment structure magnifies the effect.

5.2 Social Security Payroll Tax

With the estimated fundamentals (i.e., σ and the distribution of λ), we can illustrate the welfare consequences of other tax policies and, in particular, the role of employment adjustment in transmitting the effects of tax incidences. Specifically, we investigate a widely-used tax instrument on employment; that is, the social security payroll tax as an example. However, as China does not apply the payroll tax, we consider the tax schedule in the U.S. and its intended tax reforms. The U.S. payroll tax requires that employers and employees each pay social security tax at a specified rate up to a taxable maximum. In 2020, the tax rate is set at 6.2%, and the tax applies to the first US$137,700 in earnings. To apply our framework to the tax setting, we set the tax rate as $\tau_{payroll} = 0.062$ and the maximum taxable earnings as $\bar{w}_{payroll} = 960RMB$, the threshold of the previous setting. Consequently, the profit function changes to

$$\pi_j = p_j q_j - w_H H_j - w_L L_j - rK_j - \tau_{payroll}\bar{w}_{payroll}H_j - \tau_{payroll}w_L L_j. \tag{33}$$

The optimal labor ratio is

$$\frac{H_j}{L_j} = \left(\frac{\lambda}{1-\lambda}\frac{w_L + \tau_{payroll}w_L}{w_H + \tau_{payroll}\bar{w}_{payroll}}\right)^{\sigma}, \tag{34}$$

which is a function of the two policy parameters, $\tau_{payroll}$ and $\bar{w}_{payroll}$. Hence, firms could respond to the change in policy parameters by adjusting their employment structure.

There are two debating policy changes of payroll tax in the U.S. First, there is an ongoing debate in the U.S. about whether to increase $\tau_{payroll}$; specifically, the H.R. 860 bill proposes an annual increase of 0.0005. To understand the possible consequence of this rate increase, we conduct a counterfactual with the tax rate $\tau_{payroll}$ increased to 0.0625 and the maximum taxable earnings $\bar{w}_{payroll}$ unchanged. To do so, we set the parameters at the extended model levels (as shown in Table 6). Specifically, $r = 5.58\%$, $\beta = 0.75$, $w_L = 867.65$ and $w_H = 1252.44$; $\sigma = 1.1803$ and $\eta = 0.0791$ estimated from the extended model and the nonparametric DD approach; the distribution of λ_j is inversely estimated from equation (8).

The results are summarized in panel A of Table 7. With flexible employment adjustment, the increase in payroll tax rate increases total payroll tax revenue by 0.6464 percent. However, it also reduces GDP by 0.1569 percent: firms decrease capital by 0.1216 percent and total employment by 0.1646 percent on average. These results are consistent with the findings of Benzarti and Harju (2018), who employ a discontinuity in the average payroll tax rate and show that a higher tax rate decreases both capital and labor inputs and, hence, the

sales of firms.

[Insert Table 7 Here]

In addition, with the tax rate increased, skilled labor becomes relatively less expensive. As a result, firms substitute unskilled labor with skilled labor, leading to an average increase in the ratio of skilled to unskilled of 0.0124 percent.

To understand the role of employment adjustment, we consider a counterfactual with $\tau_{payroll}$ increased by 0.0005, $\bar{w}_{payroll}$ fixed, and no employment adjustment. We find that firms become more responsive; that is, they reduce employment and capital more on average (i.e., by 0.1678 percent and 0.1247 percent, respectively). As a result, GDP decreases by 0.1619 percent, more than the decline in the setting with flexible employment adjustment. The increase in payroll tax rate makes firms use more skilled labor to replace unskilled labor, which alleviates the negative shock of the tax incidence. While our research findings may not directly apply to the policy debate in the U.S. (due to the differences in the choice of the maximum taxable earnings and the application of the estimated elasticity of labor substitution), the estimated percentage changes in input factors, GDP, and tax revenue may shed some light on the mitigating role of employment structure adjustment in the payroll tax setting.

Second, the maximum taxable earnings $\bar{w}_{payroll}$ is adjusted annually according to the change in wage index. To illustrate the effect of $\bar{w}_{payroll}$, we study a counterfactual with $\bar{w}_{payroll}$ increased by 1 percent (=9.6 RMB) and the tax rate $\tau_{payroll}$ fixed. The estimation results are presented in panel B of Table 7. The increase of $\bar{w}_{payroll}$ makes skilled labor relatively more expensive; therefore, firms decrease the ratio of skilled to unskilled labor by 0.0535 percent on average. Meanwhile, firms decrease total employment and capital by 0.0916 percent and 0.0743 percent on average, respectively. As a result, GDP declines by 0.1219 percent. On the other hand, the payroll tax revenue increases by 0.5083 percent, reflecting that the increase in maximum taxable earnings dominates the decline in tax base.

Moreover, when firms cannot adjust their employment structure, the total employment, capital, and GDP decline less, but tax revenue increases more after an increase in $\bar{w}_{payroll}$. The amplifying effect of the employment adjustment with the increase in maximum taxable earnings operates in the opposite direction to that with the increase in the payroll tax rate, as the former induces to a substitution from skilled labor to unskilled labor, whereas the latter induces the reverse substitution.

6 Conclusions

In this paper, we analyze the responses of firms to tax policies in the employment structure adjustment margin by exploring a kinked wage deduction tax policy in China. According

to this policy, domestic firms in China can deduct their per worker monthly wage payments from taxable income only up to a certain statutory limit and thus have incentives to adjust their employment structures.

Our analysis is based on the first wave of the Economic Census of China in 2004. We first present clear evidence of bunching in the distribution of the firm-level average per worker monthly wage around the deduction limit. This pattern is observed only for firms subject to the limit when the limit is imposed, indicating that the bunching is due to the firms' responses to the kink. We then apply a reduced-form bunching method to estimate how firms respond to the policy and find that firms did not relabel part of the labor cost as other deductible terms. Instead, firms decreased capital input and the ratio of skilled to unskilled labor by around 0.31 to 0.80 percent and 0.26 to 0.60 percent, respectively. The decline in the labor ratio may come from both the labor substitution and the booking manipulation. By constructing theoretical models, we quantify the role of employment structure adjustment in transmitting the distortion effect of the tax incidence. We estimate the substitution elasticity between skilled and unskilled labor to be approximately 1.1803 and find that firms reported 5.54% to 7.91% more unskilled labor as ghost workers to achieve a greater reduction of wage payments. Welfare analysis shows that the nonlinear wage deduction policy decreased GDP by 0.2872 percent, with the booking manipulation margin easing the effect of tax incidence whereas the adjustment of employment structure amplifying the effect.

Finally, we apply our framework to the social security payroll tax setting in the U.S. The results show that a 0.05 percentage point increase in the payroll tax rate decreases GDP by 0.1569 percent, with the employment structure adjustment alleviating the decline, whereas a 1 percent increase in the maximum taxable earning reduces GDP by 0.1219 percent, with the employment structure adjustment amplifying the decrease. This result may shed light on the ongoing debate in the U.S. about whether to increase the payroll tax rate annually and provide policymakers some advice on whether to promote labor market flexibility to alleviate the distortion effect of such an annual increase.

References

Acemoglu, D. (2002). Technical Change, Inequality, and the Labor Market. *Journal of Economic Literature*, 40(1):7–72.

Angrist, J. D. (1995). The Economic Returns to Schooling in the West Bank and Gaza Strip. *The American Economic Review*, 85(5):1065–1087.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). Trends in U.S. Wage Inequality: Revising the Revisionists. *The Review of Economics and Statistics*, 90(2):300–323.

Benzarti, Y., Carloni, D., Harju, J., and Kosonen, T. (2019). What Goes Up May Not Come Down: Asymmetric Incidence of Value-Added Taxes.

Benzarti, Y. and Harju, J. (2018). Are Taxes Turning Humans into Machines? Using Payroll Tax Variation to Estimate the Capital-Labor Elasticity of Substitution.

Card, D. and Lemieux, T. (2001). Can Falling Supply Explain the Rising Return to College for Younger Men? A Cohort-Based Analysis. *The Quarterly Journal of Economics*, 116(2):705–746.

Chen, Z., Liu, Z., Serrato, J. C. S., and Xu, D. Y. (2019). Notching R&D Investment with Corporate Income Tax Cuts in China.

Chetty, R., Friedman, J. N., Olsen, T., and Pistaferri, L. (2011). Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records. *The Quarterly Journal of Economics*, 126(2):749–804.

China Statistical Yearbook (2005). China Statistics Press [in Chinese].

Diamond, R. and Persson, P. (2017). The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests.

Elias, F. (2015). Labor Demand Elasticities Over the Life Cycle: Evidence from Spain's Payroll Tax Reforms. pages 1–71.

Fuest, C., Peichl, A., and Siegloch, S. (2018). Do Higher Corporate Taxes Reduce Wages? Micro Evidence from Germany. *American Economic Review*, 108(2):393–418.

Garicano, L., Lelarge, C., and Van Reenen, J. (2016). Firm Size Distortions and the Productivity Distribution: Evidence from France. *The American Economic Review*, 106(11):3439–3479.

Gourio, F. and Roys, N. (2014). Size-Dependent Regulations, Firm Size Distribution, and Reallocation. *Quantitative Economics*, 5(2):377–416.

Harasztosi, P. and Lindner, A. (2019). Who Pays for the Minimum Wage? *American Economic Review*, 109(8):2693–2727.

Harju, J., Matikka, T., and Rauhanen, T. (2016). The Effects of Size-Based Regulation on Small Firms: Evidence from VAT Threshold.

Head, K. and Mayer, T. (2014). Gravity Equations: Workhorse, Toolkit, and Cookbook. In Gopinath, G., Helpman, E., and Rogoff, K., editors, *Handbook of International Economics*, volume 4, chapter 3, pages 131–195. North-Holland, Poland.

Heckman, J. J., Lochner, L., and Taber, C. (1998). Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents. *Review of Economic Dynamics*, 1(1):1–58.

House, C. L. and Shapiro, M. D. (2008). Temporary Investment Tax Incentives: Theory with Evidence from Bonus Depreciation. *American Economic Review*, 98(3):737–768.

Huttunen, K., Pirttilä, J., and Uusitalo, R. (2013). The Employment Effects of Low-Wage Subsidies. *Journal of Public Economics*, 97(1):49–60.

Katz, L. F. (1998). Wage Subsidies for the Disadvantaged. In Freeman, R. B. and Gottschalk, P., editors, *Generating Jobs*, pages 21–53. Russell Sage Foundation, New York.

Katz, L. F. and Autor, D. H. (1999). Changes in the Wage Structure and Earnings Inequality. In Ashenfeher, O. and Card, D., editors, *Handbook of Labor Economics*, volume 3, chapter 26, pages 1463–1555. Elsevier Science B.V.

Katz, L. F. and Murphy, K. M. (1992). Changes in Relative Wages, 1963-1987: Supply and Demand Factors. *The Quarterly Journal of Economics*, 107(1):35–78.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8(1):435–464.

Kleven, H. J. and Waseem, M. (2013). Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan. *The Quarterly Journal of Economics*, 128(2):669–723.

Krusell, P., Ohanian, L. E., Ríos-Rull, J. V., and Violante, G. L. (2000). Capital-Skill Complementarity and Inequality: A Macroeconomic Analysis. *Econometrica*, 68(5):1029–1053.

Liu, L., Lockwood, B., and Almunia, M. (2017). VAT Notches, Voluntary Registration , and Bunching: Theory and UK Evidence.

Ohrn, E. (2018). The Effect of Corporate Taxation on Investment and Financial Policy: Evidence from the DPAD. *American Economic Journal: Economic Policy*, 10(2):272–301.

Onji, K. (2009). The Response of Firms to Eligibility Thresholds: Evidence from the Japanese Value-Added Tax. *Journal of Public Economics*, 93:766–775.

Saez, E. (2010). Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy.*, 2(3):180–212.

Saez, E., Schoefer, B., and Seim, D. (2019). Payroll Taxes, Firm Behavior, and Rent Sharing: Evidence from a Young Workers' Tax Cut in Sweden. *The American Economic Review*, 109(5):1717–1763.

Serrato, J. C. S. and Zidar, O. (2016). Who Benefits from State Corporate Tax Cuts? A Local Labor Markets Approach with Heterogeneous Firms. *American Economic Review*, 106(9):2582–2624.

Yagan, D. (2015). Capital Tax Reform and the Real Economy: The Effects of the 2003 Dividend Tax Cut. *American Economic Review*, 105(12):3531–3563.

Zwick, E. and Mahon, J. (2017). Tax Policy and Heterogeneous Investment Behavior. *American Economic Review*, 107(1):217–248.

Figures and Tables

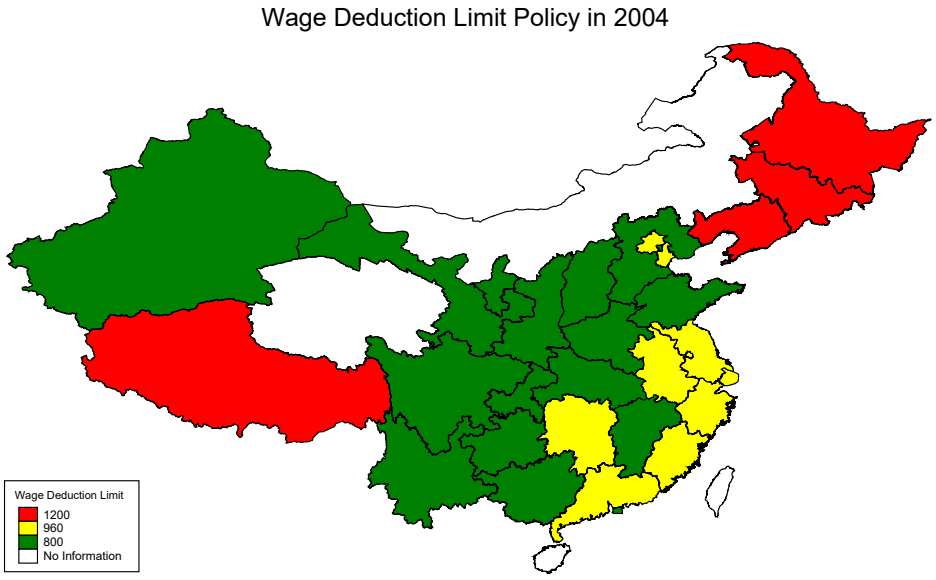


Figure 1

Wage Deduction Limit Policy in 2004

Notes: This figure shows the distribution of implemented wage deduction limits across all municipalities and provinces in China in 2004.

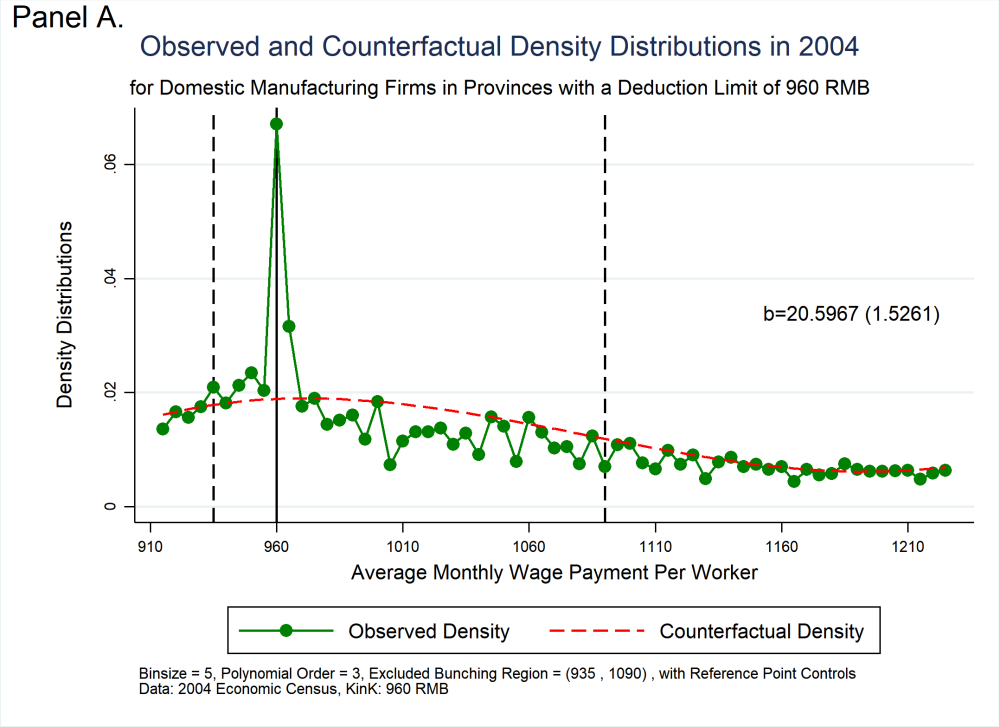
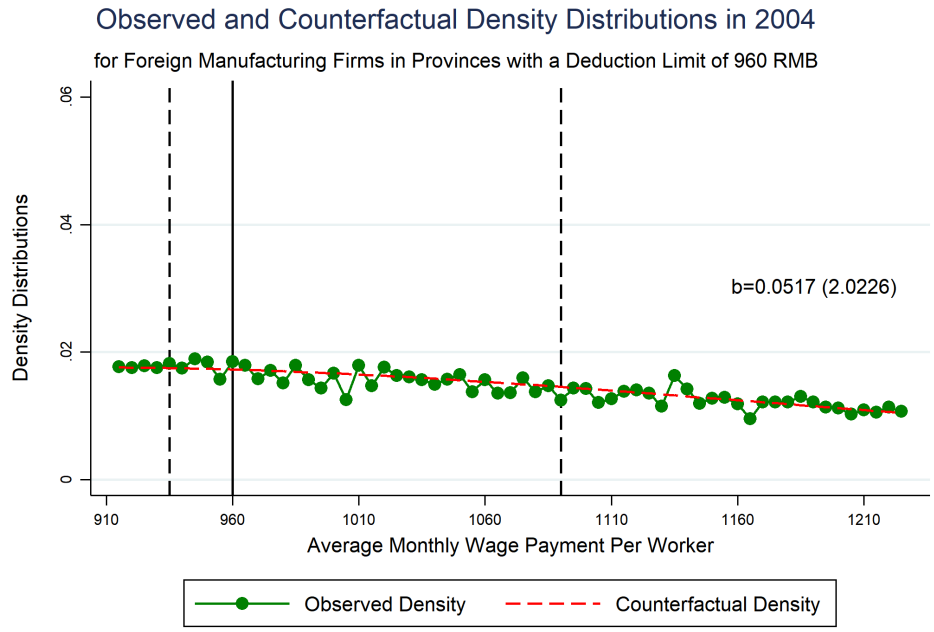


Figure 2

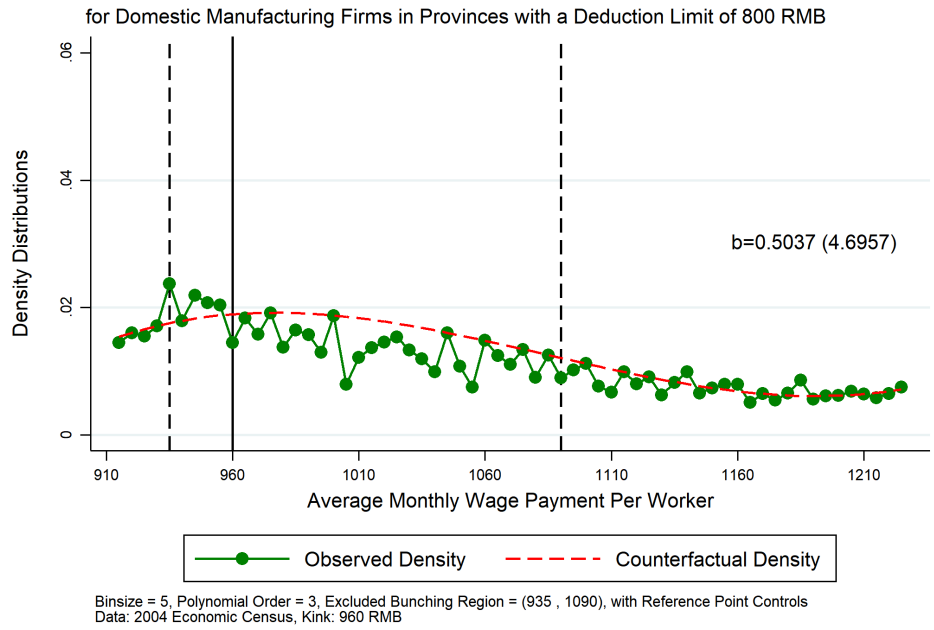
Density Distribution of Monthly Average Per Worker Wage for DEs with a Deduction Limit of 960 RMB

Notes: This figure shows the density distributions of the average per worker monthly wage around the kink point of 960 (demarcated by the vertical line) for domestic enterprises (DEs) located in the provinces with 960 RMB wage deduction limit. The solid curve displays the observed density in 5 RMB bins, with the rounding patterns around multiples of 100 RMB for monthly wage and 500 RMB and 1000 RMB for annual wage removed; the dashed curve displays the counterfactual density constructed by excluding a window of (930, 1070) centered on the kink point, controlling for multiples of 100 RMB for monthly wage and 500 RMB and 1000 RMB for annual wage and fitting a polynomial of third order to the observed distributions.

Panel A.



Panel B.



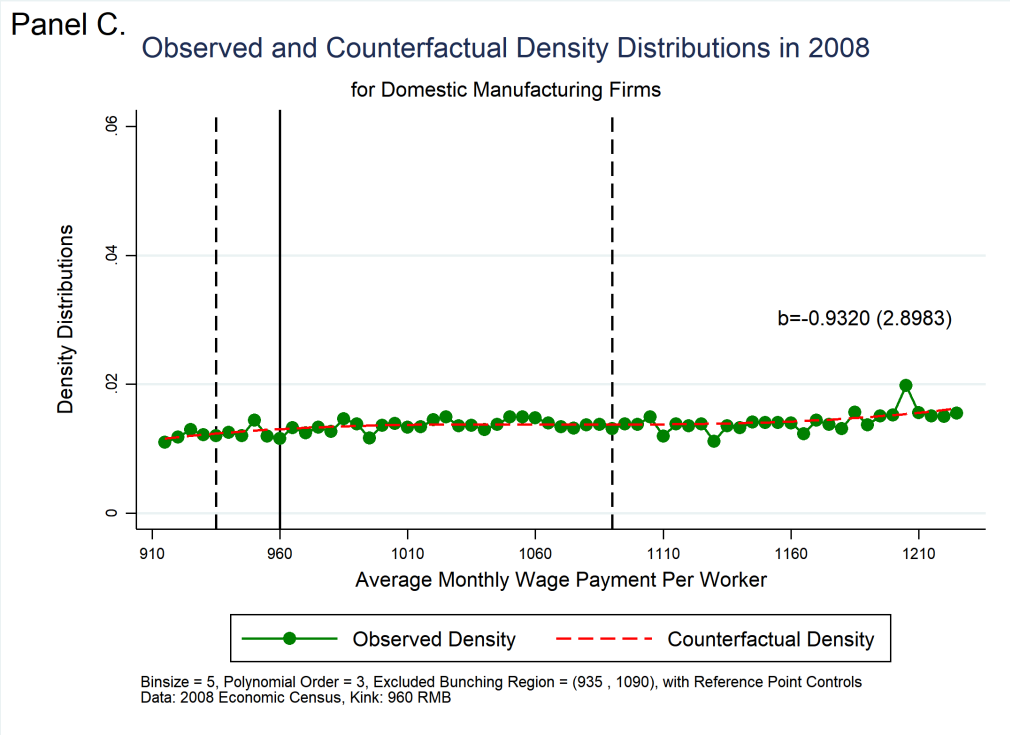


Figure 3

Density Distribution of Monthly Average Per Worker Wage for Control Groups

Notes: These figures show the density distributions of the average per worker monthly wage around the kink point of 960 (demarcated by the vertical line) for firms in the control groups. Panel A depicts the densities for foreign-invested enterprises (FIEs) located in the provinces with a 960 RMB wage deduction limit; Panel B shows those for domestic enterprises (DEs) subject to an 800 RMB wage deduction limit; Panel C plots those for all DEs in 2008 using the second Economic Census data. The solid curves display the observed densities in 5 RMB bins, with the rounding patterns around multiples of 100 RMB for monthly wage and 500 RMB and 1000 RMB for annual wage removed; the dashed curves display the counterfactual densities by excluding a window of (930, 1070) around the kink point, controlling for multiples of 100 RMB for monthly wage and 500 RMB and 1000 RMB for annual wage, and fitting a polynomial of third order to the observed distributions in panel A, B, and C, respectively.

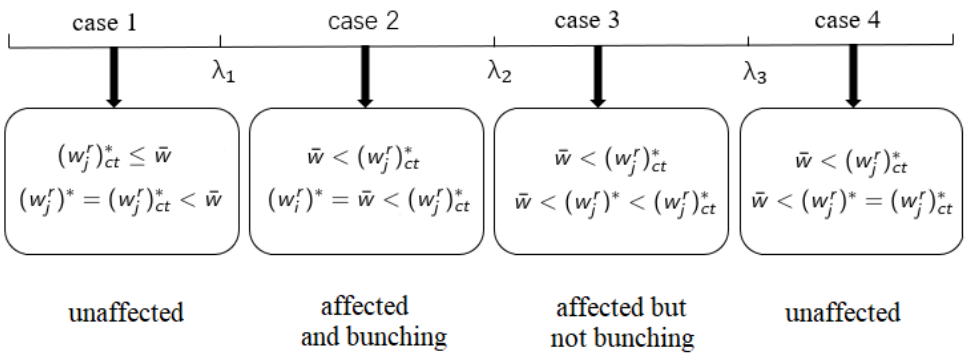


Figure 4

Firms' Responses under the Linear and Nonlinear Tax Schedules

Notes: This figure compares the firms' optimal choices of the average per worker monthly wage under the linear and nonlinear tax schedules.

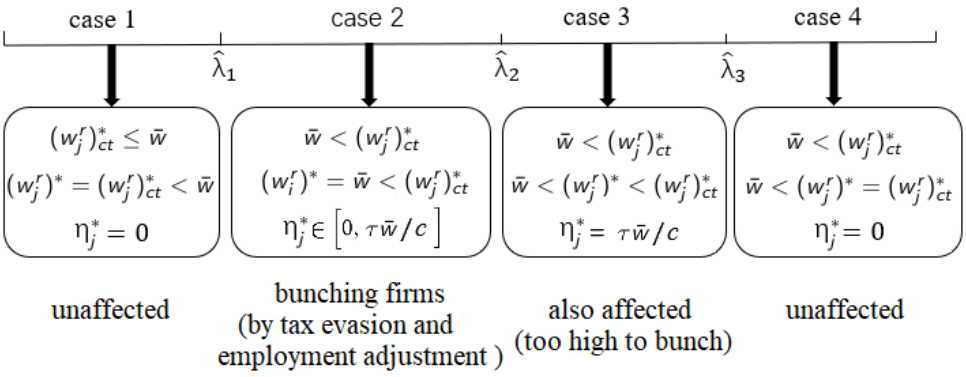


Figure 5

Firms' Responses under the Linear and Nonlinear Tax Schedules with Employment Manipulation

Notes: This figure compares the firms' optimal choices of the average per worker monthly wage and manipulation degree under the linear and nonlinear tax schedules.

Table 1. Deduction Limit for the Average Per Worker Monthly Wage in Corporate Income Tax

Period	Domestic Enterprises		Foreign-Invested Enterprises
	National Deduction Limit	Allow for 20% inflation	Deduction limit
1994.5∼1995.12	500	Yes	Fully
1996.1∼1999.12	550	Yes	Fully
2000.1∼2006.6	800	Yes	Fully
2006.7∼2007.12	1600	No	Fully
2008.1∼Now	Fully	No	Fully

Note: This table summarizes deduction limits for the average per worker monthly wage in Corporate Income Tax (CIT) for all firms in China since May 1994.

Table 2. Summary Statistics for Domestic Enterprises in the Main Sample

Variable	(1) N	(2) mean	(3) P25	(4) P50	(5) P75
Average Wage Per Worker (Monthly, RMB)	629,759	837.5624	518.0180	727.2728	947.9167
Employment	665,299	44.4820	8	16	38
Skilled versus Unskilled Labor	587,054	0.9756	0.1111	0.3333	0.8750
Capital (ln)	626,182	7.2088	6.2538	7.0901	8.0790
Unemployment Insurance per Worker (RMB)	629,759	315.4075	0	0	62.7962
Employee Benefits per Worker (RMB)	660,333	301.0581	0	0	0
Administrative Cost per Worker (RMB)	660,333	2588.5920	0	0	0

Note: This table displays the summary statistics for the DEs in the main sample used in this paper, i.e., domestic manufacturing firms in provinces with a 960 RMB limit. The average monthly wage per worker is calculated by dividing the total wage bill by 12 months and by total employment. Workers are classified into two categories based on education level—the highest education levels of high school or above and junior secondary school or below. The skilled versus unskilled labor ratio is calculated as the ratio of total employment of these two groups. Total capital is used as the capital measurement. The variables unemployment insurance per worker, employee benefits per worker, and administrative cost per worker are calculated as the ratios of the firm's total corresponding spending to total employment.

Table 3. Reduced-Form Estimates of Manipulation

	Unemployment Insurance per Worker (1)	Employee Benefits per Worker (2)	Administrative Cost per Worker (3)	Capital (ln) (4)	H/L (5)
DE_960_2004	-0.1687	-0.0757	0.9998	-0.3079***	-0.6001***
	(0.1741)	(0.0772)	(0.9096)	(0.1033)	(0.0414)
DE_960_2004 - FIE_960_2004	0.0507	0.0287	-0.1040	-0.7061***	-0.5235***
	(0.0612)	(0.0257)	(0.6854)	(0.2873)	(0.0558)
DE_960_2004 - DE_800_2004	0.0405	0.0139	0.7676	-0.7980***	-0.2608***
	(0.0696)	(0.0332)	(0.7490)	(0.3554)	(0.0181)
(DE_960_2004 - DE_800_2004)	0.0549	0.0184	-0.1740	-0.6712***	-0.2707***
(FIE_960_2004 - FIE_800_2004)	(0.0623)	(0.0258)	(0.6620)	(0.2913)	(0.0134)

Note: This table shows the reduced-form estimates of firms' manipulation responses, with standard errors in parentheses. Due to data limitations, unemployment insurance is estimated over small domestic firms with annual sales revenue not exceeding 5 million, employee benefits and administrative cost are estimated over large domestic firms with annual sales revenue exceeding 5 million. The outcome variable capital is presented in natural logarithm form. The first two rows present the results using a parametric approach following Chetty et al. (2011) to construct the counterfactual density for Domestic Enterprises (DEs) with a 960 RMB limit. The third and fourth rows present the results using the density distribution of Foreign-Invested Enterprises (FIEs) located in the provinces with a 960 RMB limit to construct the counterfactual density. The fifth and sixth rows present the results using the density distribution of DEs in the provinces with an 800 RMB limit to construct the counterfactual density. The last two row present the results using the difference between the density distributions of FIEs located in provinces with 960 RMB and 800 RMB limits to construct the counterfactual for the difference between the densities of DEs with 960 RMB and 800 RMB limits. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table 4. Elasticity of Substitution between Skilled and Unskilled Labor

Samples	(1) Δw	(2) w_H	(3) w_L	(4) τ	(5) σ
Panel A: Δw based on 2004 Economic Census					
w_H and w_L from 2004 Economic Census					
DE_960_2004	20.5967	1252.44	867.65	0.31	1.8444
DE_960_2004 - FIE_960_2004	27.3043	1252.44	867.65	0.31	2.4002
DE_960_2004 - DE_800_2004	12.5880	1252.44	867.65	0.31	1.1548
(DE_960_2004 - DE_800_2004) (FIE_960_2004 - FIE_800_2004)	22.7656	1252.44	867.65	0.31	2.0261
Panel B: Δw based on 2004 Economic Census,					
w_H and w_L from 2005 Chinese Population Census					
DE_960_2004	20.5967	1250	860	0.31	1.9829
DE_960_2004 - FIE_960_2004	27.3043	1250	860	0.31	2.5926
DE_960_2004 - DE_800_2004	12.5880	1250	860	0.31	1.2344
(DE_960_2004 - DE_800_2004) (FIE_960_2004 - FIE_800_2004)	22.7656	1250	860	0.31	2.1816

Note: This table shows the estimates for the elasticity of substitution between skilled and unskilled labor for Domestic Enterprises (DEs) with a 960 RMB limit. Columns (1) presents the normalized bunching size. Columns (2) and (3) present the monthly wages for skilled and unskilled labor, respectively, and the values are calculated with equations (23) and (24) as conditions. Column (4) presents the tax rate, calculated as the mean effective tax rate of the corresponding group of firms. Column (5) presents the elasticity estimate. Panel A (B) displays the results using wage rates calculated from the 2004 economic census (2005 population census). The first row of each panel presents the results using a parametric approach following Chetty et al. (2011) to construct the counterfactual density for DEs with a 960 RMB limit. The second and third rows of each panel present the results using the density distributions of Foreign-Invested Enterprises (FIEs) located in the provinces with a 960 RMB limit and of Domestic Enterprises (DE) in the provinces with an 800 RMB limit to construct the counterfactual densities, respectively. The fourth row of each panel presents the results using the difference between the density distributions of FIEs located in provinces with 960 RMB and 800 RMB limits to construct the counterfactual for the difference between the densities of DEs with 960 RMB and 800 RMB limits.

Table 5. Elasticity of Substitution and Degree of Booking Manipulation

Samples	Δw (1)	w_H (2)	w_L (3)	τ (4)	σ (5)	η (6)
DE_960_2004	20.5967	1252.44	867.65	0.31	1.2607	5.95%
DE_800_2004	9.1000	1090.02	704.49	0.28	1.2607	4.48%
DE_960_2004 - FIE_960_2004	27.3043	1252.44	867.65	0.31	1.2821	8.52%
DE_800_2004 - FIE_800_2004	11.4180	1090.02	704.49	0.28	1.2821	5.37%
DE_960_2004 - DE_800_2004	12.5880	1252.44	867.65	0.31	1.2145	4.25%
DE_800_2004 - DE_960_2004	11.5062	1090.02	704.49	0.28	1.2145	3.20%
(DE_960_2004 - DE_800_2004) (FIE_960_2004 - FIE_800_2004)	22.7656	1252.44	867.65	0.31	1.1803	7.91%
(DE_800_2004 - DE_960_2004) - (FIE_800_2004 - FIE_960_2004)	14.9741	1090.02	704.49	0.28	1.1803	5.54%

Note: This table shows the estimates for the elasticity of substitution between skilled and unskilled labor and the degree of booking manipulation. Column (1) presents the normalized bunching size. Columns (2) and (3) present the monthly wages for skilled and unskilled labor, respectively, and the values are calculated from the 2004 economic census with equations (23) and (24) as conditions. Column (4) presents the tax rate, calculated as the mean effective tax rate of the corresponding group of firms. Column (5) presents the elasticity estimate, and column (6) presents the booking manipulation estimate. The first (second) row presents the results using a parametric approach following Chetty et al. (2011) to construct the counterfactual density for Domestic Enterprises (DEs) with a 960 (800) RMB limit. The third (fourth) row presents the results using the density distribution of Foreign-Invested Enterprises (FIEs) located in the provinces with a 960 (800) RMB limit to construct the counterfactual density. The fifth (sixth) row presents the results using the density distribution of DEs in the provinces with an 800 (960) RMB limit to construct the counterfactual density. The seventh (ninth) row presents the results using the difference between the density distributions of FIEs located in provinces with 960 RMB and 800 RMB (800 RMB and 960 RMB) limits to construct the counterfactual for the difference between the densities of DEs with 960 RMB and 800 RMB (800 RMB and 960 RMB) limits.

Table 6. Welfare Analysis with the Extended Model

β (1)	r (2)	w_h (3)	w_l (4)	σ (5)	η (6)	Percent of Difference in GDP (7)
Panel A. Full Adjustment of Employment Structure						
0.75	0.0558	1252.44	867.65	1.1803	7.91%	-0.2872%
Panel B. No Adjustment of Employment Structure						
0.75	0.0558	1252.44	867.65	1.1803	7.91%	0.8863%

Note: This table shows the welfare analysis results for the extended model. The counterfactual density is estimated from the extended model with the booking manipulation. We use the difference between the density distributions of Foreign-Invested Enterprises (FIEs) located in provinces with 960 RMB and 800 RMB limits to construct the counterfactual for the difference between the densities of Domestic Enterprises (DEs) with 960 RMB and 800 RMB limits. Panel A presents the results when firms can fully adjust their skilled to unskilled labor ratio; panel B presents the results when firms cannot adjust the employment structure. Column (1) presents the substitution parameter of the CES utility function in equation (4), and the value is set based on the central value in the range of estimates used in the previous literature [for a review, see Head and Mayer (2014)]. Column (2) presents the interest rate, which is set based on data from the World Bank. Columns (3) and (4) present the monthly wages for skilled and unskilled labor, respectively, which are calculated from the 2004 economic census with equations (23) and (24) as conditions. Column (5) presents the elasticity estimate from the nonparametric difference-in-differences (DD) approach. Column (6) presents the booking manipulation estimate. Column (7) presents the percent difference in GDP, calculated as the ratio of the difference between the GDP under the implemented nonlinear tax schedule and the GDP' under the counterfactual linear tax schedule over GDP'.

Table 7. Welfare Analysis of U.S. Social Security Tax

	$\tau_{payroll}$ (1)	Earning (2)	Maximum Taxable σ (3)	η (4)	Percent of Difference in Tax Revenue (5)	Percent of Difference in GDP (6)	Percent of Difference in Capital (7)	Percent of Difference in H/L (8)	Percent of Total Employment (9)
Panel A. Increase the Payroll Tax Rate by 0.01									
Full Adjustment	0.062	960	1.1803	7.91%	0.6464%	-0.1569%	-0.1216%	0.0124%	-0.1646%
No Adjustment	0.062	960	1.1803	7.91%	0.6418%	-0.1619%	-0.1247%	0	-0.1678%
Panel B. Increase the Maximum Taxable Earning by 5 RMB									
Full Adjustment	0.062	960	1.1803	7.91%	0.5083%	-0.1219%	-0.0743%	-0.0535%	-0.0916%
No Adjustment	0.062	960	1.1803	7.91%	0.5280%	-0.0999%	-0.0610%	0	-0.0778%

Note: This table shows the welfare analysis results of the U.S. social security payroll tax. The counterfactual density is estimated from the extended model with booking manipulation and constructed by using the difference between the density distributions of Foreign-Invested Enterprises (FIEs) located in provinces with 960 RMB and 800 RMB limits to construct the counterfactual for the difference between the densities of Domestic Enterprises (DEs) with 960 RMB and 800 RMB limits. Panel A presents the estimated changes when the payroll tax rate increases by 0.01 percent; panel B presents those estimates when the maximum taxable earnings threshold increases by 5 RMB. Column (1) presents the payroll tax rate, which is set based on the current rate. Column (2) presents the maximum taxable earnings level of payroll tax, which is set at the studied threshold of 960 RMB. Column (3) presents the elasticity estimate from the nonparametric difference-in-differences (DD) approach. Columns (4) presents the booking manipulation estimate. Columns (5) to (9) present the percent difference in payroll tax revenue, GDP, capital level, ratio of skilled to unskilled labor, and total employment, respectively, which are calculated as the ratio of the difference between the value under the specified payroll tax schedule and the value when the payroll tax rate or maximum taxable earnings increases.