The (In)effectiveness of Targeted Payroll Tax Reductions*

Alessandra Fenizia, Nicholas Li, and Luca Citino

January 1, 2024

Abstract

This paper studies the cost-effectiveness of payroll taxes for stimulating labor demand. It uses rich administrative data to study the effects of an Italian reform that raised social security contributions for apprenticeship contracts but granted a substantial discount for firms with 9 employees or less. The discount does not increase demand for apprenticeship contracts. Instead, it subsidizes inframarginal hiring. This reform is not cost-effective. Point estimates imply that each million euros of foregone social security contributions supports the employment of 347 apprentices for one month and no permanent contracts (these estimates are not statistically different from zero).

1 Introduction

Policymakers often turn to targeted payroll tax reductions to combat high unemployment rates among the young, the low-skilled, and the long-term unemployed (OECD, 2003, 2011). However, critics argue that targeted payroll tax reductions are not cost-effective because they subsidize inframarginal employment that would exist even in the absence of the reform. While the evidence on their effectiveness is mixed, four recent, prominent studies argue that these policies increase the employment of targeted workers (Benzarti and Harju, 2021a; Cahuc et al., 2019; Saez et al., 2019, 2021).

*Alessandra Fenizia: afenizia@gwu.edu. Nicholas Li: nicholas.li@gwu.edu. Luca Citino: Luca.Citino@bancaditalia.it. We thank David Card and Jörn-Steffen Pischke for their guidance and support during the early stages of this project. We also thank Fabrizio Balassone, Francesco D'Amuri, Andrew Goodman-Bacon, Steven Hamilton, Hilary Hoynes, Elira Kuka, Francesca Lotti, Matteo Paradisi, Enrico Rubolino, Vincenzo Scrutinio, Bryan Stuart, Roberto Torrini, Eliana Viviano, and Valeria Zurla for the useful discussion and comments. We thank Vincenzo Pezone for generously sharing the data on the minimum wages. The realization of this project was possible thanks to the "VisitInps Scholars" program. We are very grateful to the staff of Direzione Centrale Studi e Ricerche for their invaluable support with the data and the institutional setting. The opinions expressed in this paper are those of the authors alone and do not necessarily reflect the views of the Bank of Italy or INPS.

Measuring the effectiveness of payroll tax reductions in boosting the hiring of marginalized groups has two empirical requirements. First, the incentives must be targeted toward marginalized workers. Second, the variation in incentives across firms must be unrelated to unobservable determinants of firm composition. Owing to the lack of policy variation that satisfies both requirements simultaneously, there has been little direct observation of whether and how firms capture the surplus associated with targeted payroll tax reductions. Cross-regional studies do not directly target marginalized workers (Bennmarker et al., 2009; Benzarti and Harju, 2021a; Bohm and Lind, 1993; Korkeamaki and Uusitalo, 2006), and national studies that compare targeted and untargeted workers do not have variation in incentives across firms (Bozio et al., 2020; Egebark and Kaunitz, 2013; Huttunen et al., 2013; Saez et al., 2019, 2012, 2021; Rubolino, 2021).

This paper analyzes changes in firm behavior in response to a targeted, temporary reduction in payroll taxes using the confidential matched employer-employee dataset collected by the Italian Social Security Institute (INPS) between 2003 and 2009. In contrast to other studies, the policy targets specific workers at specific firms. In 2007, Italian firms with at most 9 employees were given relief from increases to required social security contributions (SSCs) for apprenticeship contracts. The relief was equivalent to roughly two months of earnings per apprentice, 8% of the earnings for a typical 20-month apprenticeship, and phased out over time. As a result of the policy, SSCs change discontinuously with firm size, providing a clean empirical setting to study the impact of a relative reduction in payroll taxes on firm outcomes. Our identification strategy exploits the differential change in SSCs for firms above and below the 9-employee threshold in a difference-in-discontinuity design. Concretely, the reduced-form differences-in-discontinuity effects measure “intention-to-treat” effects using narrow variation in a neighborhood of the policy threshold. We use the policy variation as an instrument to measure jobs supported per unit of lost tax revenue.

The design provides a strong first stage—we find discontinuous effects on subsidy take-up and SSCs. At the same time, we provide ample evidence supporting the validity of the reduced form effects and, correspondingly, the instrument’s exclusion restriction. First, we find no evidence of manipulation in firm size (the running variable) at the threshold, mitigating concerns that the reform generates costly firm-size distortions observed in other settings (Garicano et al., 2016; Caicedo et al., 2020). Second, there are no pre-trends in the estimated discontinuity in the outcomes. Third, there are no estimated effects on observed firm characteristics, industry composition, or geographical composition; ergo, our results are not confounded by comparing observably different firms over time. Fourth, the design is robust to secular rotations in the relationship between firm outcomes and firm size (rotations can be generated, for example, by macroeconomic trends or mean reversion).

We find that the SSC discount targeted to small firms does not increase the demand for apprenticeship contracts. Instead, the policy primarily subsidized inframarginal firms (i.e., those who did not change their behavior in response to the reform). We also find that the policy did not increase the rate at which existing apprentices were given permanent contracts (i.e., transformations). Critics of payroll tax cuts argue that these policies may have a limited impact on employment if firms can adjust wages. We find that firms do not adjust apprentices' earnings in response to the reform. (We, however, lack the statistical power to precisely quantify the tax incidence). This is consistent with wage rigidities imposed by the Collective Bargaining Agreements (CBAs).

Because our policy generates variation across firms, we can examine whether the reform led to undesirable and unintended firm responses. In contrast to previous studies (Cappellari et al., 2012), we find no substitution toward or away from apprentices to other contract types. Firms did not opportunistically re-label existing contracts, did not churn through more apprentices, and did not hire lower-quality workers. Eligible firms did not limit their hiring of permanent employees to remain eligible for the tax discount. We show that the null effects are unlikely to be driven by the size of the subsidy, low salience or awareness of the policy, the firm's incentives to maintain eligibility, or the subsidy's temporary nature. Instead, our null results are consistent with inelastic demand for apprenticeship contracts. The policy's only reduced-form effect is on the first-stage outcome—social security contributions.

Finally, we formally measure the cost-effectiveness of the policy. The policy is more cost-effective if the employment effects are relatively large and/or the jobs last longer; it is less cost-effective if the SSCs cuts primarily subsidize inframarginal employment (i.e., contracts that would have existed even in the absence of the policy) and/or the jobs have short duration. Using the research design as an instrument for tax expenditure, we estimate that each million euros of foregone social security contributions supports the employment of 347 apprentices for one month and no additional open-ended positions (the estimates are not statistically different from zero). IV estimates of jobs per unit of foregone revenue ignore potential savings in unemployment and welfare benefits; however, because the policy had no appreciable impact on employment, these savings are likely negligible. We conclude that this policy is not cost-effective. Direct hiring of apprentices or permanent workers at their prevailing wage would be cheaper by several orders of magnitude.

This paper studies labor demand, a tradition as old as economics itself (Smith, 1776; Marx, 1910). Because of their direct policy relevance, economists have extensively documented the impact of policies that change labor costs on workers and more recently on firms (Benzarti and Harju, 2021a; Card et al., 2018; Cahuc et al., 2019; Katz, 1998; Neumark, 2013; Saez et al., 2019, 2021; Levy Yeyati et al., 2019; Zurla, 2021). Because of their ubiquity, economists have paid specific attention to payroll tax cuts (Bohm and Lind, 1993; Bennmarker et al., 2009; Benzarti and Harju, 2021a; Egebark and Kaunitz, 2013; Huttunen et al., 2013; Korkeamaki and Uusitalo, 2006; Saez et al., 2012, 2019, 2021; Rubolino, 2021).

We contribute to the literatures on wage subsidies and payroll taxes in three ways. First, our empirical setting is conducive to credibly identifying the causal effects of targeted payroll tax reductions. Identification strategies require (1) a (quasi-random) set of firms to be treated and (2) incentives targeted at the same specific workers. Cross-region policies satisfy the former, generating variation across firms, but the workers may also be different (Bennmarker et al., 2009; Benzarti and Harju, 2021a; Bohm and Lind, 1993; Korkeamaki and Uusitalo, 2006). National policies may target similar workers, but all firms are treated (Bozio et al., 2020; Egebark and Kaunitz, 2013; Huttunen et al., 2013; Saez et al., 2019, 2012, 2021; Rubolino, 2021). Our setting satisfies both requirements: the policy we study targets specific workers at specific firms. The setting lends itself to a clean identification strategy that makes clear comparisons, provides transparent testable assumptions, and is robust to broad categories of confounding factors. Together, the setting and identification strategy complement our rich administrative data, which allow us to credibly document rich patterns of complementing or substituting hiring (or lack thereof) in the face of targeted subsidies.

Our paper provides two new perspectives on the literature documenting the effects of wage subsidies and payroll tax cuts. Limitations notwithstanding, the evidence on the policies' effectiveness is mixed. However, four recent, prominent studies argue that wage subsidies (1) increase the employment of targeted workers (Saez et al., 2019, 2021) and (2) are particularly effective when temporary and during recessions (Cahuc et al., 2019; Benzarti and Harju, 2021a). Our paper provides a direct and transparent test for these hypotheses. Instead, we find that temporary subsidies during a recession—the best-case scenario for effectiveness—primarily subsidize inframarginal employment and have no effects on labor demand.

We also note that other studies have generally evaluated wage subsidies' effectiveness based on employment effects. Along this dimension, our reduced-form findings stand in stark contrast with recent estimates and add a precise zero to the collection of previous results. However, we argue that wage subsidies should instead be evaluated based on their cost-effectiveness. Examining differing policies across different countries is inherently difficult, but normalizing employment effects against fiscal costs offers a unified way of comparing results across studies. As such, we provide disciplined IV estimates for jobs supported per unit revenue, our target parameter of interest.

Our policy has two features that should make it relatively more cost-effective than comparable estimates: (1) take-up is not automatic (so fewer inframarginal "always-taker" firms are subsidized) and (2) the subsidy is phased-out over time. Surveying the literature on wage subsidies, evaluations with large employment effects often come at enormous costs. Our IV approach for jobs supported per unit revenue is novel relative to the literature, but taking back-of-the-envelope estimates from other studies at face value, we find that the cost of generating higher employment with wage subsidies is exceedingly high in those settings. We conclude that, from the perspective of cost-effectiveness, the literature on wage subsidies and payroll taxes is much more negative than previously suggested.

Finally, our paper also contributes to the literature that critically examines difference-in-differences designs and what researchers can learn from parallel pre-trends (Borusyak et al., 2021; De Chaisemartin and D'haultfœuille, 2023; Goodman-Bacon, 2021; Rambachan and Roth, 2023; Roth and Sant'Anna, 2023; Roth et al., 2023; Sun and Abraham, 2021). Researchers often evaluate the impact of policies that apply to units above a given threshold by comparing the outcome of units just above the threshold with those just below using a difference in difference design (Benzarti et al., 2020; Bozio et al., 2017; Cahuc et al., 2019; Goos and Konings, 2007; Saez et al., 2019). Our paper illustrates the perils of discretizing continuous treatment variables in a difference-in-differences setting. While this seems like a transparent and reasonable design, we show that it can lead to misleading conclusions even in the presence of parallel pre-trends. (In fact, a previous version of this paper applied such a design and drew such conclusions.) Such a design does not distinguish level shifts at the threshold with rotations of the conditional expectation function, leading to potentially spurious effects. Our design is robust to such rotations.

The paper is structured as follows. Section 2 describes the institutional setting. Section 3 presents the data. Section 4 develops the empirical strategy and discusses the identifying assumptions. Sections 5 probes the validity of our design and discusses the main results. Section 6 evaluates the cost-effectiveness of the policy. Finally, Section 7 concludes.

2 Institutional Background

This section describes the legal framework for apprenticeship contracts in Italy and the policy variation we exploit in our empirical analysis.

2.1 Apprenticeship Contracts in Italy

Apprenticeships are labor contracts that allow workers to earn a professional qualification and a salary in exchange for labor services (Snell, 1996; Ryan, 2012). In Italy, the law mandates at least 120 hours of training per year, of which at least 80 hours are devoted to occupation-specific training and 40 hours to general training (e.g., job safety regulation, psychology of labor, and teamwork).[1] The collective bargaining agreements (CBAs) regulate the content of the training. The training combines on-the-job training and formal education, which can be offered either by the firm or by government-funded third parties (see Albanese et al., 2017 for more details).[2]

During the period of study, only private-sector workers aged 18–29 are eligible to work as apprentices. Newly hired apprentices go through a short probationary period (maximum two months), after which they can be laid off only for just cause. Apprenticeship contracts can last at most six years. Although the apprentices nominally have permanent contracts, at the end of the contract, the firm can decide whether to convert the apprenticeship contract into an open-ended contract or to let the worker go at no additional cost.

In return for the training they provide, firms can pay apprentices up to 2 levels below the pay grade negotiated by the CBAs for their qualification. Firms also benefit from lower employer SSCs (starting from 2007, employer SSCs on apprenticeship contracts amount to 10% of the apprentice earnings, while employer SSCs on open-ended contracts amount to 27% of the worker's earnings). Firms also pay lower SSCs in the first year in which the apprenticeship contract is converted into an open-ended contract (Law 56/1987). The law imposes limitations on the use of these contracts. Firms cannot employ more apprentices than regular workers. An exception to this rule is that firms with at most three workers can employ up to three apprentices.

Table A.2 reports the summary statistics for the apprentices at baseline (January 2006). The typical apprentice is male (58.5%), 22.5 years old, earns 994.3 euros per month, and has 3.7 years of experience. The vast majority of apprentices are native (89%) and have had at least one previous job in the private sector (98.3%).

2.2 The 2007 Budget Bill

The 2007 Budget Bill (Law n.296/2006), passed on December 27, 2006 and effective on January 1, 2007, increased employers' SSCs on all apprenticeship contracts to finance the

[1] During the study period, there were three types of apprenticeship contracts – two of which are quite rare. In this section, we describe the "apprenticeship for job qualification" (*apprendistato professionalizzante*), which covers approximately 95% of apprentices in the country (D'Arcangelo et al., 2019).

[2] In Italy, there is a debate about whether such training is really valuable for workers. Tiraboschi (2014) argues that "although a number of legal provisions establish compulsory training during the apprenticeship, the reality is often very distant from the ideal apprenticeship model, and this tool becomes a mere instrument of exploitation of a flexible and cheaper labor force". For evidence on the returns to apprenticeship contracts on Italian young workers, refer to Citino (2020).

introduction of paid sick leave for apprentices.[3] This bill introduced a discontinuity in SSCs for apprentices for firms above 9 employees. The discontinuity in the change in labor costs provides a clean empirical setting to study the impact of a change in payroll taxes on firm outcomes.

Figure 1 shows how SSCs changed in response to the 2007 Budget Bill. For illustrative purposes, we compute the SSCs for the average apprenticeship contract at baseline (yearly earnings: 12,000 euros). Before 2007, all firms paid 2.85 euros per apprenticeship contract per week.[4] This amounted to roughly 148 euros per year (green triangles in Figure 1). Starting from January 1, 2007, the employer SSCs changed discontinuously for firms above and below the 9-employee threshold. Under the new regime, firms with more than 9 employees paid 10% of the apprentice's earnings in social contributions (≈ 1,200 euros per year, hollow blue circles in Figure 1). Firms with 9 employees or less paid 1.5% of the apprentice's earnings in the first year of the contract (≈ 180 euros), 3% in the second year (≈ 360 euros), and 10% in all the following years (≈ 1,200 euros, orange circles in Figure 1). Appendix Table A.1 reports the implied changes in SSCs for the average apprenticeship contract at baseline (yearly earnings: 12,000 euros). The savings amount two roughly two months of earnings per apprentice, 8% of a typical 20-month apprenticeship.

The eligibility for reduced SSCs was based on the *policy-relevant* firm size, total full-time equivalent employment minus apprentices, temporary agency workers, workers on leave, and workers with an on-the-job training contract. Our rich administrative data allow us to follow this definition closely (see Appendix B for more details). The increase in SSCs applied to both existing apprenticeship contracts and those signed after January 1, 2007. For contracts signed after January 1, 2007, eligibility was determined by the firm size at the time of hiring. For pre-existing contracts, the eligibility was determined based on the average firm size in 2006.

The discount was not applied automatically to firms meeting the eligibility criteria. To claim the discount, firms had to flag a box when filing their monthly report to the Italian Social Security Agency. No other pre-existing or concurrent policy was discontinuous at nine employees.[5]

The top panel of Figure 2 illustrates the relationship between the share of firms claiming the subsidy in January 2007 (take-up rate) and the policy-relevant firm size in the same

[3] The 2007 Budget Bill did not affect the SSCs paid by the apprentices. Because SSCs represent a large share of payroll taxes in Italy, we use these two terms interchangeably.

[4] The weekly contribution for contracts eligible for occupational injury insurance was 0.1 euros higher.

[5] Consistent with the absence of other policies, Figure 5 shows the cumulative density function of policy-relevant firm size before (orange lines) and after the reform (green lines). There is no discontinuity at the 9-employee threshold, and the distribution of policy-relevant firm size remains stable over time.

month (green circles). The take-up rate is approximately 2% for firms below the 9-employee threshold, sharply decreases around nine employees, and converges to 0.4% for firms above the threshold. These relatively low monthly take-up rates reflect primarily the fact that 75% of firms do not hire apprentices at all and relatively few firms hire apprentices in any given *month* (Table 1). Because 80% of eligible firms who hire apprentices receive the discount, it is unlikely that the low take up rates are primarily driven by firms begin unaware of the policy. Similarly, because policy-relevant firm size is not a function of the number of apprentices, the relatively low take-up rates are unlikely to be a byproduct of firms choosing not to hire apprentices to maintain eligibility.

This figure highlights two important facts. First, there is no appreciable discontinuity at the threshold. This is partly due to mismeasurement in policy-relevant firm size *at the time of hiring*: we measure policy-relevant firm size over the course of the month, but eligibility is determined instantaneously.[6] Second, the take-up rate does not drop to zero past the threshold. Some firms receive the payroll tax reduction despite being ineligible. As we discuss in Section 4.2, two-way non-compliance will lead intention-to-treat (ITT) estimates to be smaller than the treatment effect on the treated.

3 Data

In this section, we describe the data that form the basis of our empirical analysis and how we construct our sample.

3.1 Data and Sample Selection

Social Security Records. Our main source of data is the confidential matched employer-employee dataset collected by the Italian Social Security Institute (*Istituto Nazionale di Previdenza Sociale*—INPS hereafter) known as UNIEMENS. These data originate from the reports that firms have to file monthly with INPS. These data cover the universe of all private non-agricultural firms with at least one employee from 1983 to today. Firms are identified by a unique tax number and workers are identified by their social security number. As for firms, these data include location, detailed industry codes, juridical status, and opening and closing dates. For each job spell, we observe the beginning and end dates, earnings net of SSCs, detailed information about whether the contract is covered by specific policies,

[6] We also do not directly observe very rare on-the-job training contracts and workers on temporary leave (see Appendix B). As a result, we may overestimate policy-relevant firm size for some firms, but because these arrangements are exceedingly rare, these sources of measurement error are likely to be small.

part- versus full-time status, coarse occupation categories (apprentice, blue-collar, white-collar, or manager), and worker's demographic information. The social security records also contain detailed information on employer SSCs and whether the firm received the SSC discount. The UNIEMENS database does not contain information on self-employed workers, the unemployed, the informally employed, or public-sector workers.

In our analysis, we primarily use data between January 2003 and December 2009. However, we also use the full length of our panel to construct the complete workers' histories (e.g., previous employment status and previous earnings) and their contract duration.

Sample selection: We restrict our sample to firms with policy-relevant firms size between 3 and 15 employees. This yields a sample of 857,587 firms and 24,532,943 firm-month observations. Our sample is skewed toward small firms by construction. However, because in Italy, a disproportionate share of apprenticeship contracts take place at small firms, this is a good setting to study how changes in SSCs impact apprenticeship contracts. Moreover, given that most Italian firms are very small, our sample includes the vast majority of firms in the country.

The next section describes the characteristics of our sample.

3.2 Descriptive Statistics

Table 1 displays the summary statistics for firm characteristics at baseline (i.e., in January 2006). Column 1 reports the characteristics for the full sample; columns 2 and 3 display the statistics for firms that hire apprentices and firms that ever take up the subsidy, respectively.

The average firm in our sample is a Limited Liability Company (LLC) established in the early 90s and employs 7 workers. Full-time equivalent employment is marginally lower than the average number of employees, reflecting that most of the workers are employed full-time. 93% (=6.63/7.10) of the employees have a permanent contract, while only 6% (=0.429/7.10) of them have temporary contracts. While apprenticeship contracts are nominally permanent contracts, they are not widespread and make up for approximately 6% (=0.427/7.10) of the contract types in our sample. The fraction of apprentices contracts in our sample is higher than the one in the overall economy, reflecting the fact that this contract type is more popular among small firms. Apprenticeship contracts are less prevalent in Italy than in Germany or Austria. In Italy, apprentices make up 1.8% of the labor force, compared to 3.6% in Germany and Austria, and 0.2% in the US (G20-OECD-EC Conference, 2014).

Apprenticeship contracts last on average about 20 months, and apprentices experience a substantial amount of turnover. The average firm hires 0.030 apprentices and separates from 0.015 apprentices each month. Two out of a hundred apprentices are offered a full-time position at the end of their contract ("transformations" hereafter).

While some firms use this contract type unsparingly, 75% (=1-99,311/398,412) of firms in our sample do not use it at all. Firms that have at least an apprentice in January 2006 (column 2) are marginally larger and use more apprenticeship contracts than the average firm in our sample (column 1), but look very similar in terms of hiring and separation flows of other contract types. Firms that take up the SSCs relief (column 3) are by construction marginally smaller and have more apprentices than the average firm in the sample (column 1), but do not appear to be selected on other dimensions.

Appendix Table A.3 has the same structure as Table 1 and report the industry composition of firms on our sample. Apprentice contracts are used predominantly in manufacturing, transportation, warehouse, construction, trading, and services (column 1). Firms that hire apprentices (column 2) are more likely to be in manufacturing than the average firm in our sample (column 1), and less likely to be in agriculture, services, or in public administration, education and health, reflecting the fact that public sector workers are not eligible for this contract type.

4 Measuring the Cost-Effectiveness of Employment Support

This paper's target parameter of interest is the number of jobs supported per unit of revenue. This section first discusses its identification requirements. It then explains how the differences-in-discontinuity design satisfies those requirements by (1) exploiting discontinuous eligibility in tax breaks generated by the policy while (2) remaining robust to secular changes in the relationship between apprentice hiring and firm size.

4.1 Jobs Supported per Unit Revenue

Let R_i denote firm i's tax payments. Taking as given firms' labor demand responses, a social planner chooses tax parameters that balance increased employment of type j, L^*_{ij}, against their fiscal costs, targeting an optimal number of jobs supported per unit of lost revenue, $-\frac{\partial L^*_{ij}}{\partial R_i}$.[7] The textbook model of labor supply and demand in competitive equilibrium generates an exclusion restriction: tax parameters $\boldsymbol{\tau}$ affect labor demand only because it

[7]The minus sign reflects measuring jobs per unit lost (instead of increased) revenue. Formally, if the planner has preferences governed by utility $v\left(\boldsymbol{L}^*_i, R_i\right)$ where v is increasing in both arguments, the planner chooses $\boldsymbol{\tau}$ so that jobs support per lost revenue equals the marginal rate of substitution between revenue and labor, $\frac{\partial L^*_{ij}}{\partial R_i} = -\frac{\partial v/\partial R_i}{\partial v/\partial L^*_{ij}}$.

affects firms' tax liability R and thus their marginal costs.[8] Thus, an ideal approach to recover the average jobs supported per unit lost revenue, $\gamma = -\mathbf{E}\left[\frac{\partial L_{ij}^*}{\partial R_i}\right]$, would be to use instruments derived from an experiment that randomly assigns tax parameters to different firms. The Wald estimator for jobs supported per unit revenue is given by,

$$\gamma^{\text{Wald}} = -\frac{\mathbf{E}\left[L_{ij}^*(\boldsymbol{\tau}_1) - L_{ij}^*(\boldsymbol{\tau}_0)\right]}{\mathbf{E}\left[R_i(\boldsymbol{\tau}_1) - R_i(\boldsymbol{\tau}_0)\right]}, \tag{1}$$

where $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_0$ are taxes in treatment and control regimes, respectively. The key identification challenge is generating variation in $\boldsymbol{\tau}$, holding constant firm characteristics. In the next subsection, we show how our differences-in-discontinuity design measures such variation.

Instrumental variables estimation provides a disciplined way of estimating cost-effectiveness. There are two practical and theoretical reasons to prefer jobs supported per unit revenue $-\frac{\partial L_{ij}^*}{\partial R_i}$ to its reciprocal cousin, the cost per job, $-\frac{\partial R_i}{\partial L_{ij}^*}$. First, randomly assigned tax parameters would not be a valid instrument to recover the latter: the exclusion restriction does not hold because tax parameters directly affect tax revenue even with no causal labor demand response. Second, violations of exclusion notwithstanding, there may be no "first-stage" relationship even in a large-scale randomized experiment if labor demand is the same between treatment and control firms. The "divide-by-zero" problem would generate infinite implied cost per job. Perversely, the experiments with the greatest precision would tend to be those with the greatest denominator and thus the largest first stage; in other words, the most precise estimates would come from the most cost-effective policies.

4.2 Empirical Specification: Difference-in-Discontinuity

The incentives generated by the law suggest comparing firms with policy-relevant firm size (Z_{it}) above and below the eligibility threshold of 9. Define $T_{it} = \mathbf{1}[Z_{it} > 9]$. Firm size is not randomly assigned, so firms of different sizes differ in dimensions other than program eligibility.

The discontinuity in eligibility suggests a regression discontinuity (RD) design that compares firms in close proximity to the threshold. Consider a standard regression discontinuity model in a single cross-section of the data (omitting the time index for parsimony),

$$Y_i = a_1 + a_2 T_i + g_1 Z_i + g_2 Z_i \times T_i + u_i, \tag{2}$$

[8] In a static, frictionless, neoclassical labor demand model, firm i hires vector $\boldsymbol{L}_i$ workers of different types until their marginal revenue product equals their marginal costs, $\boldsymbol{L}_i^* : \underbrace{\boldsymbol{MRP}(\boldsymbol{L}_i^*)}_{\text{marginal benefit}} = \underbrace{\boldsymbol{w} + \partial_{\boldsymbol{L}} R(\boldsymbol{w}, \boldsymbol{L}_i^*; \boldsymbol{\tau})}_{\text{marginal cost}}$, where $\boldsymbol{w}$ denotes a vector of workers' respective market wages. Note that $\boldsymbol{\tau}$ only enters the expression as a parameter of the tax liability.

where the estimated intercept shift at the discontinuity a_2 captures the causal effect of the policy. Given this design, the Wald estimator for cost effectiveness in equation 1 for firms at the discontinuity is the effect of the discontinuity on labor demand (a_2^L) divided by the effect of the discontinuity on (minus) tax payments ($-a_2^R$), the ratio of reduced-form coefficients, $\hat{\gamma} = -\frac{a_2^L}{a_2^R} = -\frac{\mathbf{E}\left[L_{ij}^*(\boldsymbol{\tau}_1) - L_{ij}^*(\boldsymbol{\tau}_0) | Z_{it}=9\right]}{\mathbf{E}[R_i(\boldsymbol{\tau}_1) - R_i(\boldsymbol{\tau}_0) | Z_{it}=9]}$.

The standard practice for regression discontinuity is to estimate local linear regressions in a small neighborhood around the discontinuity. This is infeasible in our setting for three reasons. First, local linear regression requires that the density of the running variable is continuous. In our setting, firm size bunches at round numbers (Section 5.2.2). Second, most firms have only a few employees, constraining estimation to a very narrow bandwidth. Finally, we observe the running variable at the monthly level, but eligibility is defined instantaneously, leading to measurement error in the running variable. This measurement error would tend to smooth the conditional expectation function, attenuating the RD estimates toward zero.

We address these challenges using a difference-in-discontinuities approach. Re-introducing the time index, our estimated discontinuities come from normalizing period-specific discontinuity estimates to the baseline period, January 2006:

$$\begin{aligned} Y_{it} &= a_{1t} + a_{2t}T_{it} + g_{1t}Z_{it} + g_{2t}Z_{it} \times T_{it} + u_{it} \quad \forall t \\ b_t &\equiv a_{2t} - a_{2,\text{Jan 2006}}. \end{aligned} \tag{3}$$

Estimated discontinuities at baseline reflect non-linearities in the relationship between the outcome and policy-relevant firm size. The reduced-form effects are given by b_t, the *changes* in the estimated discontinuity at the threshold. We cluster the errors at the firm level.[9] In Section 6, we estimate the Wald ratio estimator of cost-effectiveness averaging over time periods post-reform, $\hat{\gamma} = -\frac{\mathbf{E}\left[b_t^L | t \geq \text{Jan 2007}\right]}{\mathbf{E}\left[b_t^R | t \geq \text{Jan 2007}\right]}$.

To ensure that more weight comes from observations closest to the discontinuity, we follow the standard approach in the RD literature and weight observations according to a triangular kernel function (Calonico et al., 2014). To avoid estimated null results coming

[9]In order to cluster standard errors at the firm, we operationalize the difference-in-discontinuity approach

from measurement error, we exclude firms within firm-size 1 of the discontinuity.[10]

Our approach offers four key advantages. First, the approach yields pre-trend validity tests that mirror validity tests of difference-in-differences designs. Specifically, we perform a series of placebo tests by examining the difference-in-discontinuity coefficients b_t for $t <$ Jan 2007. Second, it ensures that our null results do not come from measurement error in the running variable. Ultimately, we show that our null results are unaffected by the inclusion or exclusion of firms nearest to the discontinuity (see Section 5.5). Third, our approach uses the longitudinal dimension of the panel data to bias-correct our estimates. Whereas the literature on RD has focused on removing bias by deleting data (i.e., estimating local quadratic regressions and restricting estimation to a narrow bandwidth (Calonico et al., 2014)), we subtract the bias generated by non-linearities and extrapolate to the discontinuity using the pre-period.[11] Finally, our approach is robust to rotations of the conditional expectation function that could lead to spurious inferences. Applying standard difference-

with a saturated, stacked regression model,

$$\begin{aligned}
Y_{it} = \; & \underbrace{a_{1,\text{Jan 2006}}}_{\text{Baseline intercept}} && + \underbrace{\sum_{t\neq \text{Jan 2006}} a_{1t}\Delta_t}_{\text{Time-varying intercepts}} \\
& + \underbrace{a_{2,\text{Jan 2006}} T_{it}}_{\text{Baseline discontinuity}} && + \underbrace{\sum_{t\neq \text{Jan 2006}} \boldsymbol{b}_t \left(\boldsymbol{T}_{it} \times \boldsymbol{\Delta}_t\right)}_{\textbf{Difference-in-discontinuities}} \\
& + \underbrace{g_{1,\text{Jan 2006}} Z_{it}}_{\text{Baseline slope below discontinuity}} && + \underbrace{\sum_{t\neq \text{Jan 2006}} g_{1t} \left(Z_{it} \times \Delta_t\right)}_{\text{Time-varying slope below discontinuity}} \\
& + \underbrace{g_{2,\text{Jan 2006}} Z_{it} \times T_{it}}_{\text{Baseline slope above discontinuity}} && + \underbrace{\sum_{t\neq \text{Jan 2006}} g_{2t} \left(Z_{it} \times T_{it} \times \Delta_t\right)}_{\text{Time-varying slope above discontinuity}} + u_{it},
\end{aligned} \tag{4}$$

where Δ_t are time dummies. The point estimates are identical to estimating separate regression models in each period and subtracting the baseline discontinuity from the measured discontinuity. See Appendix C that formalizes the identification assumptions.

[10]Our approach differs from standard applications of "donut-hole" RD for two reasons. First, the usual impetus for excluding data near the discontinuity in other settings is the manipulation of the running variable, but we find no evidence of manipulation, and our specification passes all tests of validity (Section 5.2). Second, our longitudinal data allows us to subtract the bias associated with extrapolation. Regardless, our (null) results do not appear to be driven by the inclusion (or exclusion) of data closest to the discontinuity (see Section 5.5).

[11]A technical literature has emerged to select a bandwidth that balances bias and precision while debiasing the estimates using controls for higher-order polynomials (Calonico et al., 2014). Calonico et al. (2014) Remark 7 notes that conventional point estimates from a quadratic regression specification coincide with their procedure that allows the point estimate and bias correction specifications to be fit on samples with differing bandwidths. As we mentioned above, the procedure proposed by Calonico et al. (2014) is not feasible in our setting. Bunching in the running variable at round numbers and the fact that most firms have only a few employees leaves a very narrow bandwidth to estimate the local linear regression.

in-differences specifications by discretizing a continuous treatment inadvertently generates an omitted variable bias by failing to control for changing slopes (i.e., restricting $g_{1t} = 0$ and $g_{2t} = 0$ in the regression discontinuity model in equation 3). See Appendix C for further details.

Our empirical strategy identifies an intention-to-treat effect. Because these reduced-form estimates do not adjust for imperfect compliance and include firms regardless of whether or not they hire apprentices, our estimated effect will be smaller than the treatment effect on the treated (those who took the subsidy). Section 5.1 illustrates how our approach estimates the effects of the policy on take-up (the "first-stage") and apprentice hiring, respectively.

5 The Effect of the Policy Threshold

This section is organized as follows. First, it graphically illustrates our approach to the policy's effects on two key outcomes: take-up and apprentice hiring. Second, it provides evidence of model validity, showing that the difference in covariates of firms just above and below the discontinuity do not change over time. Finally, it documents the policy's null effects across other outcomes.

5.1 Policy Take-Up and Apprentice Hiring: An Illustration of the Design

Figure 2 and Figure 3 deconstruct the regression specification. The top panel of each figure is a binned scatter plot approximating the conditional expectation function in January 2007 of tax-break take-up and apprentice hiring, respectively. Overlaid in grey are best-fit lines excluding different windows of data, and overlaid in black exclude a window of firm-size 1, our preferred estimates. Because Equation 3 is fully interacted, the research design repeats this estimation in each period, shown in the second panel. The third panel plots the measured discontinuity over time, and our reduced-form difference-in-discontinuity estimates are obtained by subtracting the value at base period, January 2006.

In Figure 2, the likelihood of take-up increases by 2 p.p. per month. The plot shows the change is abrupt, and our design exploits variation driven over time. Naturally, because the policy did not exist prior to January 2007, the estimates for take-up are zero in the pre-period—the DD estimates normalized to January 2006 are mechanically identical. The policy's effect on take-up declines through the end of our analysis period.

Figure 3 is constructed analogously, and it examines the policy's effect on apprentice hiring. None of the binned scatterplots show any visual sign of discontinuity. The time-series of the discontinuity estimates shows no appreciable change in January 2007 or subsequently—the DD estimates normalized to January 2006 are virtually identical. The noisy appearance of the time-series belies the precision of the estimates owed to the large administrative sample.

While difficult to see, Figure 3 also shows that the conditional expectation function is rotating clockwise, coincident with a general slowdown in overall hiring and apprentice hiring through the end of 2010. A clockwise rotation would drive down the mean of the unsubsidized firms above the threshold. In a standard difference-in-differences specification, this would lead to conclusions that the subsidy supported hiring. See Appendix C.

5.2 Tests of Validity

In standard DD settings, treatment is defined ex-ante. Parallel pre-trend validity tests ensure that treatment and control differences are constant prior to the program start date. We also test for constant pre-policy differences in our augmented specification. For example, Figure 3 shows flat pre-trends in the discontinuous differences.

However, our setting is complicated by the fact that program eligibility is defined contemporaneously. The tax cost of hiring apprentices for firms depends on their policy-relevant firm size at the time of hiring, and comparing firms just above and just below the eligibility cut-off compares different firms in each period. In particular, the policy provides employers an incentive to manipulate their policy-relevant firm size, raising the possibility that our results are the consequence of different patterns of selection rather than unbiased causal estimates.

In this section, we provide two additional sets of evidence that our design consistently compares observationally similar firms. First, we show covariate balance by estimating our main specification with firm characteristics measured at baseline. The covariate differences between firms just above and just below the threshold are constant and do not depend on when policy-relevant firm size is measured. Second, we show that the marginal distribution of policy-relevant firm size is constant over time, exhibiting no bunching or manipulation.

5.2.1 Covariate Balance and Observable Differences

Covariate differences between firms just above and just below the threshold do not change over time. Table 2 and Table 3 show covariate stability over firm age and type; firm industry; and firm location, respectively. These tables report the effects of being above the threshold (b_t) from the main DD specification in Equation 3, where the outcome variables are general firm characteristics. For parsimony, we report a subset of the estimates.[12] The first two columns report the pre-reform DD estimates for $t_0 - 48$ (January 2003) and $t_0 - 24$ (January 2005). Columns 3–5 report the post-reform estimates for t_0 (January 2007), $t_0 + 12$ (January 2008), and $t_0 + 35$ (December 2009), respectively. The last three columns report Wald F-statistics testing the null that all the DD coefficients, the pre-reform coefficients, and the post-reform coefficients are zero, respectively. There is no imbalance along age or firm type in Panel A of Table 2.

While the vast majority of covariates show no signs of imbalance, in Panel B of Table 2, the balance tests detect statistically significant DD coefficients for Utilities and Transportation and Construction dummies. Appendix Figure A.5 and Figure A.6 show that these do not seem to be the consequence of systematic changes but rather some random variation plus precision from our large administrative data.

Similarly, most region dummies are strongly balanced (Table 3), but some DD coefficients for Lombardy, Liguria, Umbria, and Molise dummies are significant at the 10% level. Nevertheless, the coefficient plots show no evidence of systematic inconsistencies (Appendix Figures A.7, A.8, A.9, and A.10).

Altogether, there is no evidence that measuring policy-relevant firm size contemporaneously results in compositional shifts or comparisons between observationally different groups. To summarize the covariate balance validity checks, we assess the policy's effects on a covariate index, the predicted values from a regression of apprentice hiring on time-invariant firm characteristics.

Figure 4 shows that being above versus below the cutoff does not correspond to changes in covariates that systematically predict apprentice hiring. The estimates are extremely precise. The statistically insignificant point estimates fluctuate between -0.0005 and +0.0005, almost two orders of magnitude smaller than the statistically insignificant effects on apprentice hiring that fluctuate between -0.01 and +0.01 (Appendix Figure A.3).

5.2.2 Stability of Marginal Distributions and Unobservable Differences

Section 5.2.1 shows that the differences in observable characteristics are stable over time, evidence that the empirical specification's validity is not compromised by comparing unlike firms over time. Our rich administrative data shows stability across a wide collection of firm characteristics.

To show that comparisons are not contaminated by unobservable confounders, researchers complement RD analyses with McCrary tests to show evidence against manipulation of

[12] The results for the full subset of estimates are available upon request.

the running variable. The McCrary tests are not informative in our setting because the distribution of the running variable exhibits bunching and is not smooth—there is excess mass at whole numbers. However, in the same spirit of the analyses, we plot the CDFs of the running variable for each of the 84 periods in Figure 5. CDFs prior to January 2007 are plotted in orange; those starting from January 2007 are plotted in green.

This figure shows that the marginal distributions are highly stable. The 84 CDFs are virtually identical and exhibit almost no change in the periods before and after the reform. Though the DD specification does not require that the marginal distribution of the running variable is stable, the plots provide strong evidence against firms manipulating firm size to become eligible for the subsidy.

5.3 Reduced-Form Effects of Subsidizing Apprentice Hiring

Figure 6 and Figure 7 show the reduced form effects of being *above* the policy threshold on take-up and social security contributions.

Despite requiring firms to opt in, the policy has bite. In a given month, firms with a policy-relevant firm size just *above* 9 are 1–1.5 p.p. *less* likely to hire an apprentice *and* claim a discount on social security contributions. Correspondingly, *larger* firms pay 25 euros *more* per month in social security contributions than *smaller* firms. The fiscal impact of 25 euros per month per firm may seem small at first glance; however, we emphasize that these reduced-form estimates do not adjust for incomplete compliance and include firms regardless of whether or not they hire apprentices. We discuss the cost-effectiveness (or lack thereof) in further detail in Section 6.

The differences between smaller and larger firms are largest at the onset of the policy in January 2007 and decline through 2010. The pre-trends for take-up (Figure 6) are mechanically zero because there was no policy prior to January 2007. However, we measure firms' social security contributions throughout our analysis period (Figure 7). The pre-trends are flat, and the differences between eligible and ineligible firms do not emerge until January 2007.

Despite claiming fewer discounts, Figure 8 shows that firms just above the policy cutoff do not have relatively more or fewer apprentices than they did before the enactment of the policy. We also see no effects on the net apprentice wage bill, the total pecuniary compensation for the firm's apprentices net of taxes and SSCs (Figure 9). The changes subsequent to the policy's implementation are neither significant nor discernible from longer-run trends. This may be consistent with the downward wage rigidity imposed by CBAs.[13] While the null

[13]CBAs impose floors for workers' compensation that all firms in the industry have to abide by.

wage bill effects are precise in absolute terms, they are not precise enough to detect effects similar to those in Figure 7, owing to greater overall variation in apprentice tax bill across firms. Consequently, we cannot reliably determine whether the firms or the workers enjoy the welfare benefits of the tax break's incidence.

5.4 Reduced-Form Effects On Other Outcomes

Although the reform has no impact on the stock of apprentices, one concern is that the reform may change the composition of apprentices and/or induce firm strategic behavior, such as churning through more apprentices, "re-labeling" existing contracts, reducing transformations to open-ended contracts, lowering the quality of new hires, and substituting temporary workers with apprentices. Table A.4 reports the effects of being above the threshold (b_t) from the main DD specification in Equation 3 for $t_0 - 48$ (January 2003), $t_0 - 24$ (January 2005), t_0 (January 2007), $t_0 + 12$ (January 2008), and $t_0 + 35$ (December 2009), mirroring the first five columns of Table 2.

No increased churning If apprentices received little to no training and were perfectly substitutable with new untrained workers, in the absence of search costs, we would expect eligible firms to let their incumbent apprentices go and substitute them with newly hired apprentices at lower costs. However, Panel A of Table A.4 shows that the reform does not impact the number of new apprenticeship contracts or apprentice separations. In other words, the reform did not induce firms to substitute incumbent apprentices with cheaper, newly hired ones. This is consistent both with valuable match-specific training and search costs.

No re-labeling existing contracts The reform may induce firms to "re-label" existing contracts as apprenticeships to take advantage of the lower social contributions in the first two years. Panel A of Table A.4 reports the estimated impact of the reform on the number of new apprenticeship contracts and the number of apprentices hired from outside the firm. These results are virtually indistinguishable, showing that the vast majority of newly hired apprentices are hired from outside the firm.

No reduction in transformations The reform may induce firms to hire young workers as apprentices to take advantage of the discount in SSCs with no intention of hiring them permanently at the end of the apprenticeship. This would generate a reduction in transformations to open-ended contracts. Panel A of Table A.4 shows that this is not the case.

No changes in the quality of new hires If the reform induces firms to use apprenticeship contracts as a mere source of cheap labor, firms may choose to invest less in the search for talented apprentices and compromise on the quality of new hires. Panel B of Table A.4 shows that firms above the threshold do not hire apprentices that look different along proxies for ability—previous salary, previous experience, or starting salary—or along demographic dimensions.

No substituting (or complementing) hiring We also study whether the reform induces firms to substitute from other contract types to apprenticeship contracts. The closest substitute to apprentices are temporary workers. Table A.4 shows no effects on temporary worker hires and separations (panel A) and, consequently, no effects on the stock of temporary workers (panel C). Apprentices are typically younger; substitution patterns would manifest in firms becoming "younger." Panel A of Table A.4 shows no effects on the hiring and separations of young workers (or any workers).

Whether they are substitutes or complements, directly estimating the threshold's effects on permanent workers is complicated by the fact that they are used to compute policy-relevant firm size, the running variable pivotal to our design. However, the stability of the marginal distributions of policy-relevant firm size and absence of bunching over time (Figure 5) point away from the policy's incentives affecting the firm's permanent labor demand. Altogether, our evidence suggests that the policy subsidized inframarginal decisions with no corresponding increases in labor demand or substitution effects.

5.5 Why No Reduced Form Effects?

The reduction in SSC we study shares several similarities with policies implemented in other European countries (Cahuc et al., 2019; Saez et al., 2019). It targets contracts for young workers, and it is implemented around the same time. However, we do not find short- or medium-run employment effects.

Here, we consider several explanations for the diverging results: (1) measurement error; (2) the size of the subsidy; (3) a lack of saliency or awareness; (4) firm incentives to maintain eligibility; and (5) the temporary nature of the subsidy. None of these explanations offer a simple reason to explain the divergent results, leaving differences in demand elasticity and research design as the most plausible reasons for diverging estimates.

Measurement error Our monthly data is high quality and high-frequency. However, we do not measure the running variable, policy-relevant firm size, at the precise moment that firms hire apprentices. If firm size fluctuates dramatically within a given month, measurement error may attenuate the reduced-form results toward zero. (With respect to the IV estimates of cost-effectiveness, uncorrelated measurement error would not generate a violation of exclusion. IV estimates would be robust so long as the first stage remains strong.)

To avoid our null results being a consequence of measurement error, our preferred specifications exclude firms within a window of 1 of the threshold, relying on the pre-period discontinuity to remove the bias associated with extrapolation. Our null results on apprentice hiring are robust to the amount of excluded data (Appendix Figure A.4).

Is the subsidy too small? The size of the subsidy is substantial, worth roughly two months of earnings for the average apprenticeship contract or 8% of the earnings for a typical 20-month apprenticeship. The SSC subsidy is similar in size to the subsidy studied by Cahuc et al. (2019) who find large employment effects on targeted workers.

Are firms aware of the policy? One possibility is that firms do not respond to the SSC discount because they were unaware of it. Ultimately, however, 80% of eligible firms that hired apprentices received the discount. It is also worth noting that the SSC discount is not applied automatically: firms must claim it. This has two implications. First, this feature does not necessarily decrease (or increase) salience relative to other policies, all else equal. Second, this feature makes the SSC discount relatively *more* cost-effective because the government does not subsidize firms that are unaware of the subsidy. Conversely, in comparable settings where the wage subsidy is implemented universally, unaware firms impose a fiscal cost, yet do not respond to the lower marginal costs by hiring more employees.

Do firms restrict hiring to maintain eligibility? Importantly, hiring apprentices does not affect eligibility because apprentices are not included in policy-relevant firm size.

Does the temporary nature of the subsidy hinder its effectiveness? The subsidy covers the first two years of each apprenticeship contract at eligible firms. Because the typical apprenticeship contract lasts 20 months, contracts are effectively subsidies for their entire duration. Moreover, previous studies suggest that temporary subsidies should be, if anything, more effective than permanent ones (Cahuc et al., 2019).

We conclude that the size of the subsidy, the lack of saliency, firm incentives to maintain eligibility, and the temporary nature of the subsidy are unlikely to explain our results. Firms simply exhibit inelastic demand for apprentices. This contrasts with other studies that utilize different designs in different settings. Next, we study the cost-effectiveness of this policy.

6 Cost Effectiveness

Most cost-effectiveness estimates either rely on structural assumptions (Cahuc et al., 2019) or are obtained through back-of-the-envelope calculations (Neumark, 2013). In contrast to previous studies, our setting allows us to estimate cost-effectiveness simply and transparently using instrumental variables. Recalling Section 4.1, in this section, we estimate the average number of jobs supported by the tax break, $\gamma = -E[\frac{\partial L^*_{ij}}{\partial R_i}]$ and compares our estimates with those from previous studies.[14]

Specifically, we estimate the following system using 2SLS:

$$\begin{aligned} L^*_{ijt} &= -\gamma R_{it} + g_L(Z_{it}, t) + \varepsilon_{it} \\ R_{it} &= bT_{it} \times Post_t + g_R(Z_{it}, t) + \eta_{it}, \end{aligned} \tag{5}$$

where L^*_{ijt} measures employment of type j (the outcome), and R_{it} measures social security contributions (the endogenous regressor). The excluded instrument is $T_{it} \times Post_t$ and $g_Y(Z_{it}, t)$ are controls for time dummies and the running variable in each period.[15] Equation 5 differs from Equation 3 only because it averages the dynamic effects into a single parameter so that the system is just-identified. For example, the first-stage equation for R_{it} is identical to its reduced-form specification, except there is a single parameter b corresponding to a single $T_{it} \times Post_t$ indicator rather than the set b_t parameters corresponding to each of the time dummies $T_{it} \times \Delta_t$. Appendix Table A.5 reports the first stage coefficient estimate, which is highly statistically significant with an F-statistic of 230.

Table 4 reports the IV estimates. In each month, the point estimates imply that €1M of lost social security contribution revenue supports the employment of 347 apprentices for one month (approximately 29 (= 347/12 jobs for one year) and €647 thousand in apprentice

[14]Note that our estimates do not take into account the potential savings stemming from (1) lower unemployment benefits and welfare transfers that would have been paid in the absence of the policy and (2) higher taxes that workers and employers pay on the jobs generated by this policy. Because the employment effects are null, these savings are likely to be minimal in our setting implying little to no difference between the net and the gross cost of the policy.

[15]Specifically, the controls in the IV system mirror those in the full reduced-form specification in Equation 4. Omitting indices for outcome equation for parsimony,

$$\begin{aligned} g\left(Z_{it}, t\right) = & a_{1,\text{Jan 2006}} && + \sum_{t \neq \text{Jan 2006}} a_{1t}\Delta_t \\ & + a_{2,\text{Jan 2006}} T_{it} \\ & + g_{1,\text{Jan 2006}} Z_{it} && + \sum_{t \neq \text{Jan 2006}} g_{1t}\left(Z_{it} \times \Delta_t\right) \\ & + g_{2,\text{Jan 2006}} Z_{it} \times T_{it} && + \sum_{t \neq \text{Jan 2006}} g_{2t}\left(Z_{it} \times T_{it} \times \Delta_t\right). \end{aligned}$$

earnings. The effects are not statistically different from zero. The 95% confidence intervals rule out effects larger than 1,700 apprentices and €2.5 million per 1€ million of tax revenue, respectively.

To give a sense of cost-effectiveness, note that the maximum tax break is 8.5%. This implies that marginal apprentice hires generate €11.8 million (1/0.085) in earnings per €1 million of foregone government revenue. An inframarginal firm that does not change behavior but accepts the subsidy would not generate any additional wage income by definition. We cannot reject a null that the policy supported no additional employment and wage income for apprentices, and the 95% confidence interval implies that the government would have to subsidize four inframarginal apprentice hires for each marginal apprentice hire (2.5/11.8). These estimates imply that if the government wanted to subsidize apprenticeship contracts at small firms, offering reduced SSCs would be 2.6 times more expensive than hiring apprentices at the prevailing wage.

Increased apprenticeships are only an intermediate goal; the ultimate goal of subsidizing apprenticeships is increasing permanent employment. Only a subset of subsidized apprentices become permanent employees. Thus, one can alternatively evaluate the subsidy against the ultimate goal, using as the endogenous variable the number of apprentices that transformed into permanent contracts. In line with our point estimates, €1M of lost social security contribution revenue does not support any transformed contracts (the point estimate is negative). Altogether, these estimates suggest that wage subsidies are a cost-ineffective method of supporting both the temporary and permanent employment of marginalized workers.

Considering reduced-form employment effects and ignoring costs, our study adds a precise zero to the collection of mixed results (Benzarti and Harju, 2021b,a; Bohm and Lind, 1993; Bennmarker et al., 2009; Korkeamaki and Uusitalo, 2006; Saez et al., 2019). However, the programs are difficult to compare because they vary in fiscal costs. Only a small subset of studies have evaluated the cost-effectiveness of these reforms (Cahuc et al., 2019; Egebark and Kaunitz, 2013; Neumark, 2013; Saez et al., 2021). Examining differing policies across different countries is inherently difficult, but normalizing employment effects against fiscal costs offers a unified way of comparing results across studies. We compute the implied number of jobs supported by € 1 million of foregone revenue implied by structural or back-of-the-envelope estimates and compare the literature to our IV estimates.[16]

Figure 10 reports the results. We find that most programs are not cost-effective. With two notable exceptions (Bartik, 2001; Cahuc et al., 2019), Figure 10 suggests that the cost

[16] The specific studies are Bartik (2001); Bartik and Erickcek (2010); Dupor and Mehkari (2016); Dupor and McCrory (2018); Egebark and Kaunitz (2013); Feyrer and Sacerdote (2011); Neumark (2013); Saez et al. (2021); Wilson (2012).

of generating employment effects is extremely high, even for programs that generate positive employment effects (Saez et al., 2019, 2021).

7 Conclusion

This paper studies the cost-effectiveness of a targeted payroll tax cut in stimulating labor demand. Using a difference-in-discontinuity framework, we find that the reduction in SSCs did not have employment effects for either apprentices or their substitutes. The program also did not have discernible wage effects. Its only effects were on tax revenue.

To evaluate the cost-effectiveness of the policy, we use the policy variation in an instrumental variables strategy to estimate the number of jobs sustained by each euro of foregone revenues. Each €1 million euro of lost social security contribution supports the employment of 347 apprentices for one month and no open-ended positions (and the estimates are not statistically significant).

Our precise null employment effects contrast with the literature, which lacks consensus on the responsiveness of labor demand to policy. However, when benchmarking other studies against their fiscal cost basis, our disciplined IV approach yields estimates that generally accord with other studies: wage subsidies to increase employment are generally fiscally ineffective. These results suggest caution in the use of payroll tax credits to stimulate employment.

References

Albanese, A., Cappellari, L., and Leonardi, M. (2017). The effects of youth labour market reforms: evidence from italian apprenticeships. *Oxford Economic Papers*.

Bartik, T. and Erickcek, G. A. (2010). The employment and fiscal effects of michigan's mega tax credit program.

Bartik, T. J. (2001). Jobs for the poor: Can labor demand policies help?

Bennmarker, H., Mellander, E., and Öckert, B. (2009). Do regional payroll tax reductions boost employment? *Labour Economics*, 16(5):480–489.

Benzarti, Y. and Harju, J. (2021a). Can payroll tax cuts help firms during recessions? *Journal of Public Economics*, 200:104472.

Benzarti, Y. and Harju, J. (2021b). Using Payroll Tax Variation to Unpack the Black Box of Firm-Level Production. *Journal of the European Economic Association*, 19(5):2737–2764.

Benzarti, Y., Harju, J., and Matikka, T. (2020). Does mandating social insurance affect entrepreneurial activity? *American Economic Review: Insights*, 2(2):255–268.

Bohm, P. and Lind, H. (1993). Policy evaluation quality: A quasi-experimental study of regional employment subsidies in sweden. *Regional Science and Urban Economics*, 23(1):51–65.

Borusyak, K., Jaravel, X., and Spiess, J. (2021). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:2108.12419*.

Bozio, A., Breda, T., and Grenet, J. (2017). Incidence of social security contributions: evidence from france. *Paris School of Economics Working Paper*.

Bozio, A., Breda, T., and Grenet, J. (2020). Does Tax-Benefit Linkage Matter for the Incidence of Social Security Contributions? Working paper.

Cahuc, P., Carcillo, S., and Le Barbanchon, T. (2019). The effectiveness of hiring credits. *The Review of Economic Studies*, 86(2):593–626.

Caicedo, S., Espinosa, M., and Seibold, A. (2020). Unwilling to train? firm responses to the colombian apprenticeship regulation. Technical report, CESifo Working Paper.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cappellari, L., Dell'Aringa, C., and Leonardi, M. (2012). Temporary employment, job flows and productivity: A tale of two reforms. *The Economic Journal*, 122(562):F188–F215.

Card, D., Kluve, J., and Weber, A. (2018). What works? a meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 16(3):894–931.

Citino, L. (2020). What are the returns to apprenticeships? evidence from italy. Technical report, WorkINPS paper n.21.

D'Arcangelo, A., Romito, A., et al. (2019). L'apprendistato tra continuità e innovazione: 18. rapporto di monitoraggio. Technical report.

De Chaisemartin, C. and D'haultfœuille, X. (2023). Two-way fixed effects and differences-in-differences estimators with several treatments. *Journal of Econometrics*, 236(2):105480.

Dupor, B. and McCrory, P. B. (2018). A cup runneth over: Fiscal policy spillovers from the 2009 recovery act. *The Economic Journal*, 128(611):1476–1508.

Dupor, B. and Mehkari, M. S. (2016). The 2009 recovery act: Stimulus at the extensive and intensive labor margins. *European Economic Review*, 85:208–228.

Egebark, J. and Kaunitz, N. (2013). Do payroll tax cuts raise youth employment? Working Paper Series 2013:27, IFAU - Institute for Evaluation of Labour Market and Education Policy.

Feyrer, J. and Sacerdote, B. (2011). Did the stimulus stimulate? real time estimates of the effects of the american recovery and reinvestment act. Technical report, National Bureau of Economic Research.

G20-OECD-EC Conference (2014). G20-OECD-EC Conference on Quality Apprenticeship: country information on apprenticeships: country responses. Technical report.

Garicano, L., Lelarge, C., and Van Reenen, J. (2016). Firm size distortions and the productivity distribution: Evidence from france. *American Economic Review*, 106(11):3439–79.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Goos, M. and Konings, J. (2007). The impact of payroll tax reductions on employment and wages: A natural experiment using firm level data. Technical report, LICOS Discussion Paper.

Grembi, V., Nannicini, T., and Troiano, U. (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics*, pages 1–30.

Huttunen, K., Pirttilä, J., and Uusitalo, R. (2013). The employment effects of low-wage subsidies. *Journal of Public Economics*, 97:49–60.

Katz, L. F. (1998). *Wage Subsidies for the Disadvantaged*, pages 21–53. Russell Sage Foundation, New York. NBER WP No. 5679, 1996.

Korkeamaki, O. and Uusitalo, R. (2006). Employment Effects of a Payroll-Tax Cut: Evidence from a Regional Tax Exemption Experiment. Discussion Papers 407, VATT Institute for Economic Research.

Levy Yeyati, E., Montané, M., and Sartorio, L. (2019). What works for active labor market policies? *CID Working Paper Series*.

Marx, K. (1910). *Value, price, and profit*, volume 5. CH Kerr & Company.

Neumark, D. (2013). Spurring job creation in response to severe recessions: Reconsidering hiring credits. *Journal of Policy Analysis and Management*, 32(1):142–171.

OECD (2003). *Employment Outlook, Towards More and Better Jobs.*

OECD (2011). *Taxation and Employment.*

Rambachan, A. and Roth, J. (2023). A More Credible Approach to Parallel Trends. *The Review of Economic Studies*, 90(5):2555–2591.

Roth, J. and Sant'Anna, P. H. (2023). When is parallel trends sensitive to functional form? *Econometrica*, 91(2):737–747.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics.*

Rubolino, E. (2021). Taxing the Gender Gap: Labor Market Effects of a Payroll Tax Cut for Women in Italy. Working paper.

Ryan, P. (2012). Apprenticeship: between theory and practice, school and workplace. In *The future of vocational education and training in a changing world*, pages 402–432. Springer.

Saez, E., Matsaganis, M., and Tsakloglou, P. (2012). Earnings determination and taxes: Evidence from a cohort-based payroll tax reform in greece. *The Quarterly Journal of Economics*, 127(1):493–533.

Saez, E., Schoefer, B., and Seim, D. (2019). Payroll taxes, firm behavior, and rent sharing: Evidence from a young workers' tax cut in sweden. *American Economic Review*, 109(5):1717–63.

Saez, E., Schoefer, B., and Seim, D. (2021). Hysteresis from employer subsidies. *Journal of Public Economics*, 200:104459.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. Number smith1776 in History of Economic Thought Books. McMaster University Archive for the History of Economic Thought.

Snell, K. D. (1996). The apprenticeship system in british history: the fragmentation of a cultural institution. *History of Education*, 25(4):303–321.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

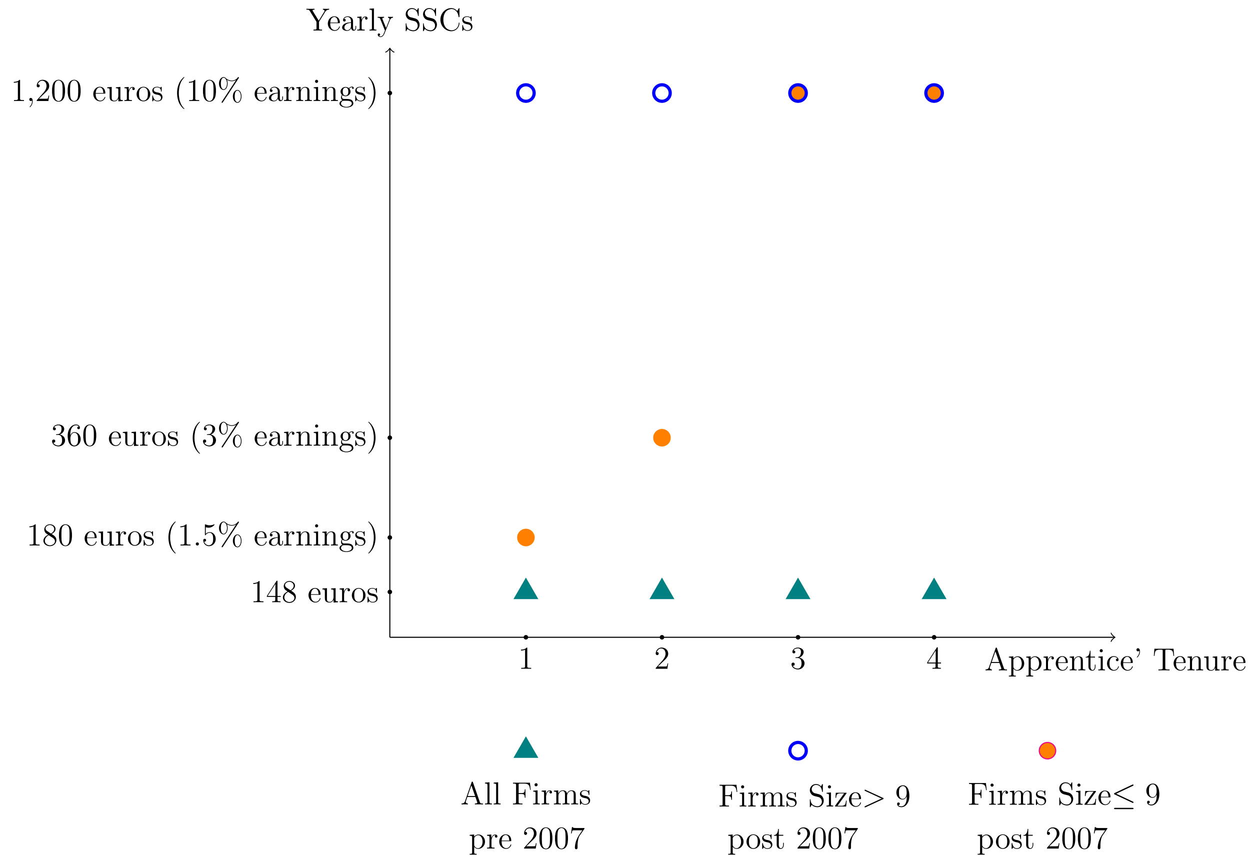
Tiraboschi, M. (2014). Young workers in recessionary times: A caveat (to continental europe) to reconstruct its labour law? *GundersonM. FazioF.(Eds.), Tackling youth unemployment*, pages 3–26.

Wilson, D. J. (2012). Fiscal spending jobs multipliers: Evidence from the 2009 american recovery and reinvestment act. *American Economic Journal: Economic Policy*, 4(3):251–282.

Zurla, V. (2021). Firm Responses to Earned Income Tax Credits: Evidence from Italy. Working paper.

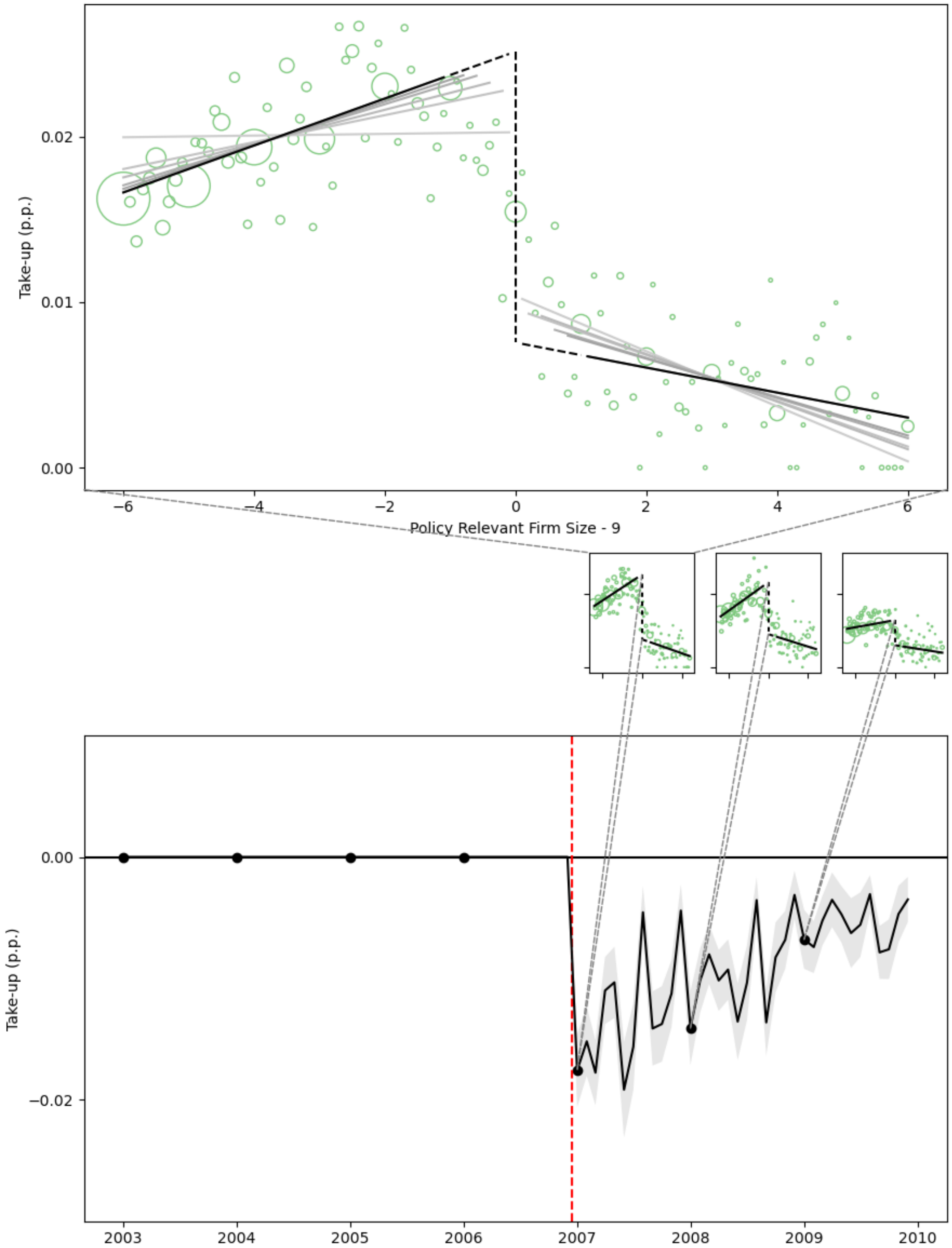
8 Figures

Figure 1: Social Security Contributions for Apprenticeship Contracts



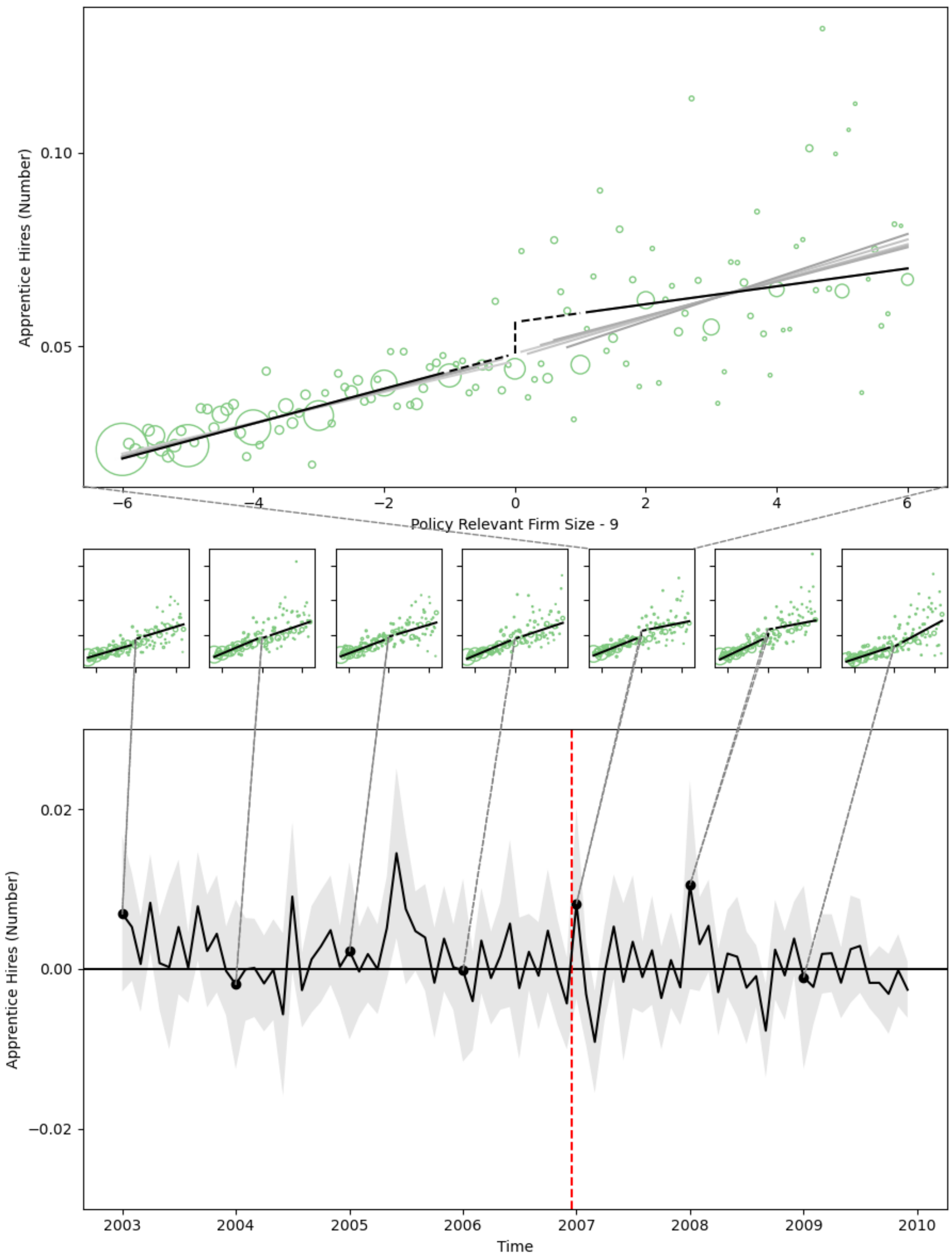
Notes: This figure illustrates how yearly social security contributions (SSCs) for apprenticeship contracts changed in response to the 2007 Budget Bill. Before 2007, employers paid a fixed weekly fee of 2.85 euros per apprenticeship contract. The yearly social contributions are computed as $2.85 \times 52 = 148.2$ euros (green triangles). After January 1, 2007, yearly social contributions are computed as a percentage of the apprentice's yearly earnings; their schedule differs between firms below or above the 9-employee threshold. Social contributions amount to 10% of the apprentice's earnings for firms with more than 9 employees (blue hollow circles). Firms with 9 employees or less pay 1.5% of the apprentice's earnings in the first year of the contract, 3% in the second year, and 10% in the third year and all the following ones (orange circles). To compute the change in social contributions implied by this policy, we use the yearly average 2006 apprentice earnings, which are equal to 12,000 euros.

Figure 2: Reduced-form effects on take-up



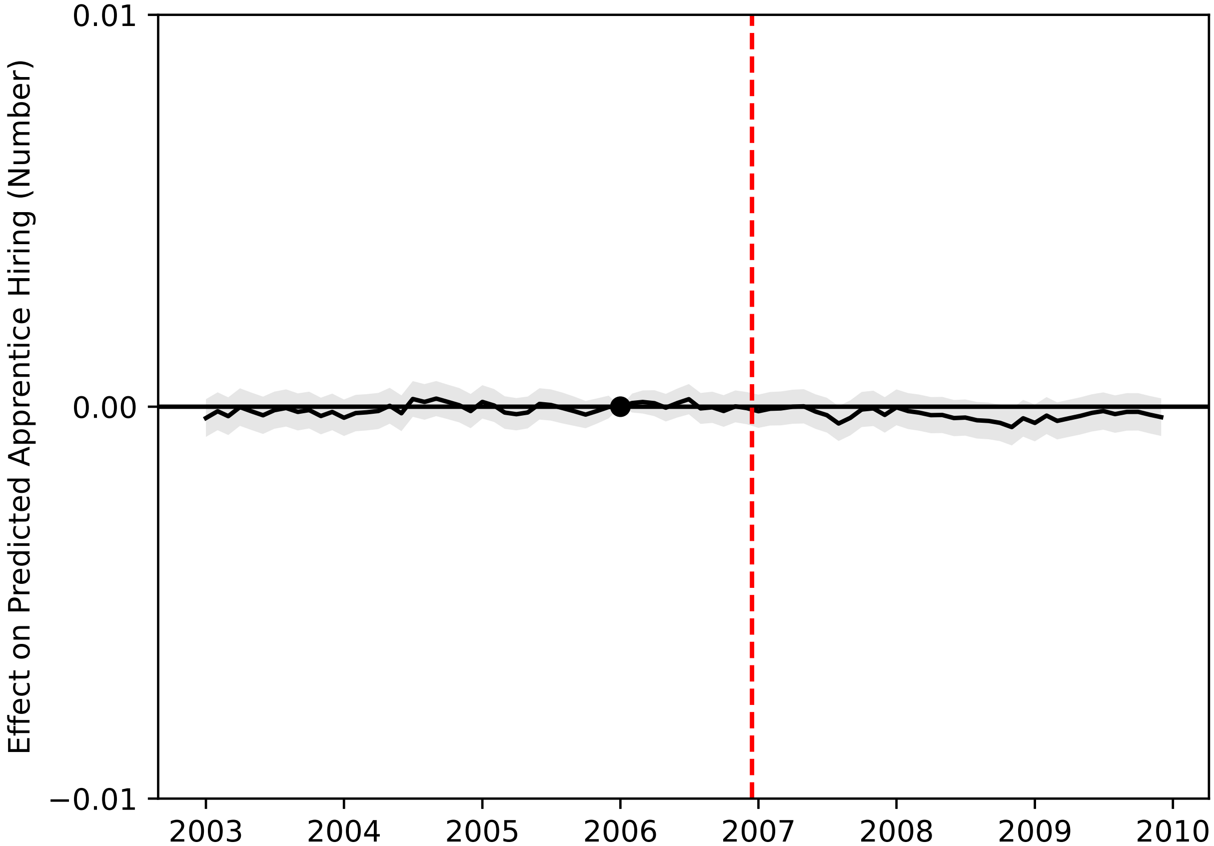
Notes: Social Security Administration data (January 2003–December 2009). This figure shows the effect of the policy on take-up. The top panel shows a binned scatterplot of take-up against policy-relevant firm size in January 2007, the first month of the policy. The size of the green dots indicates the number of firms within the bin. Fitted values from piece-wise linear regressions are overlayed. The black line indicates regressions estimated, excluding a window of 1 around the discontinuity. (Grey lines are fit using windows of 0, 0.2, 0.4, and 0.8.) The first panel is a zoomed example of the conditional expectation function in each period, shown in the second panel. The third panel plots a time-series of the discontinuity estimates. 95% confidence intervals are shaded in grey. Note that take-up is mechanically zero before January 2007.

Figure 3: Reduced-form effects on apprentice hiring



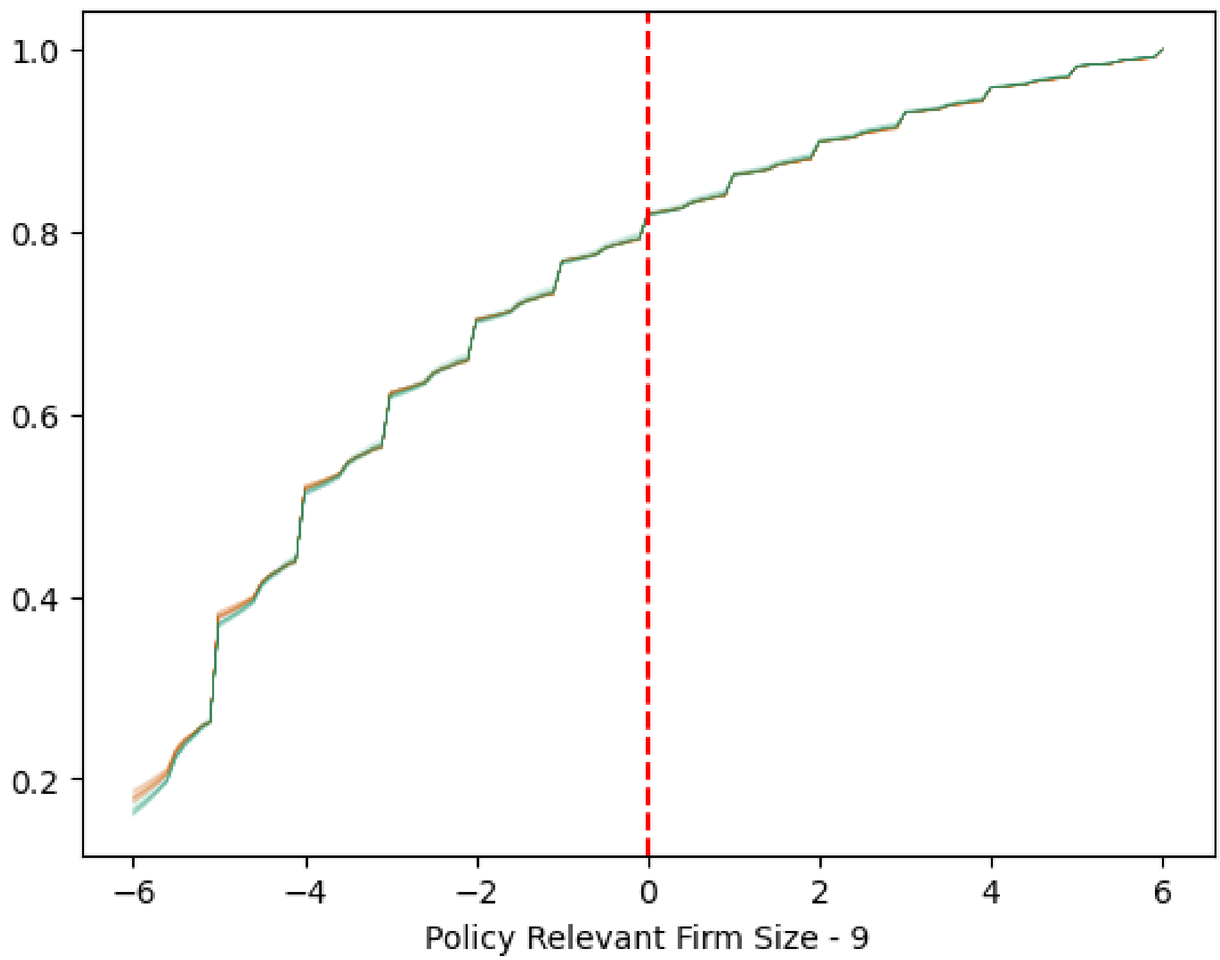
Notes: Social Security Administration data (January 2003–December 2009). This figure shows the effect of the policy on apprentice hiring, mirroring Figure 2. See notes for Figure 2 for details.

Figure 4: Reduced Form Estimates of Threshold on Covariate Index



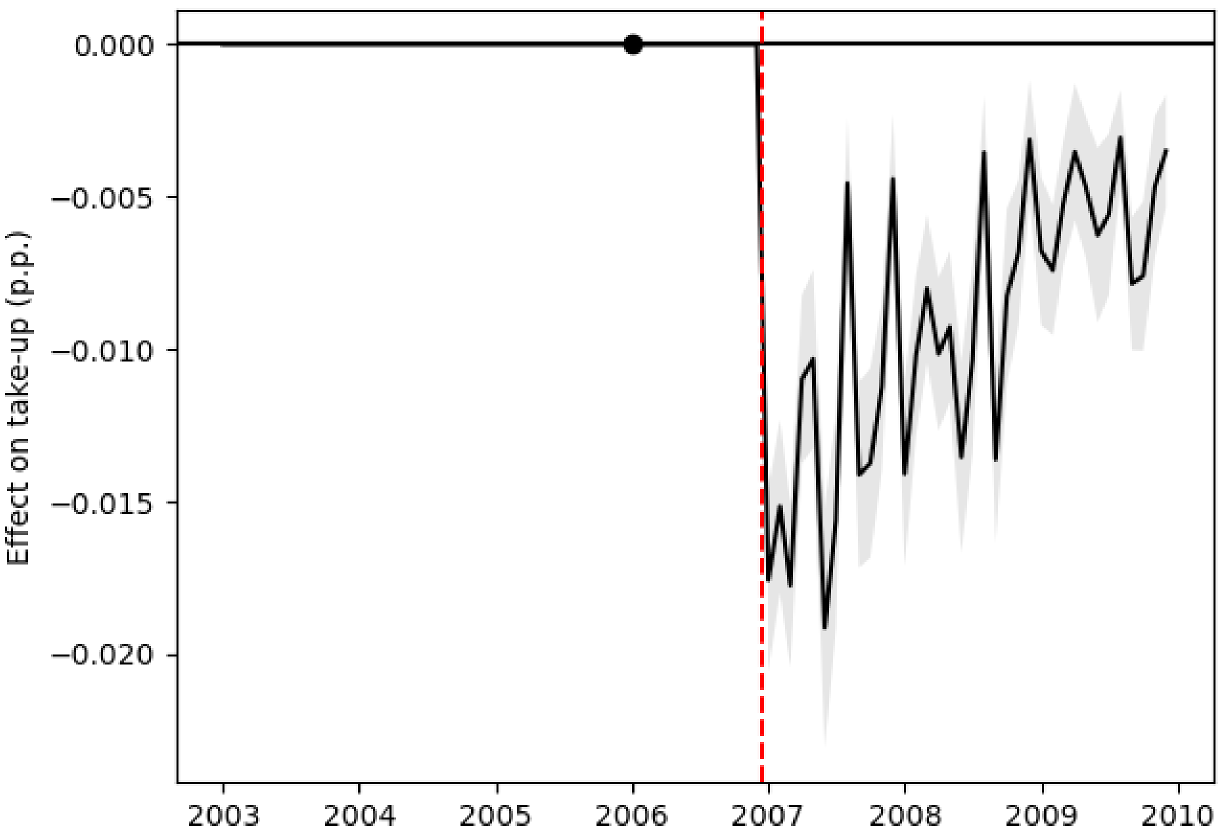
Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in equation Equation 3 where the outcome variable is a covariate index, the predicted values from a regression of apprentice hiring on time-invariant firm characteristics. Estimates are relative to January 2006, the omitted category. 95% confidence intervals are shaded in grey.

Figure 5: Empirical CDFs of Policy Relevant Firm Size, Jan 2003–Dec 2009



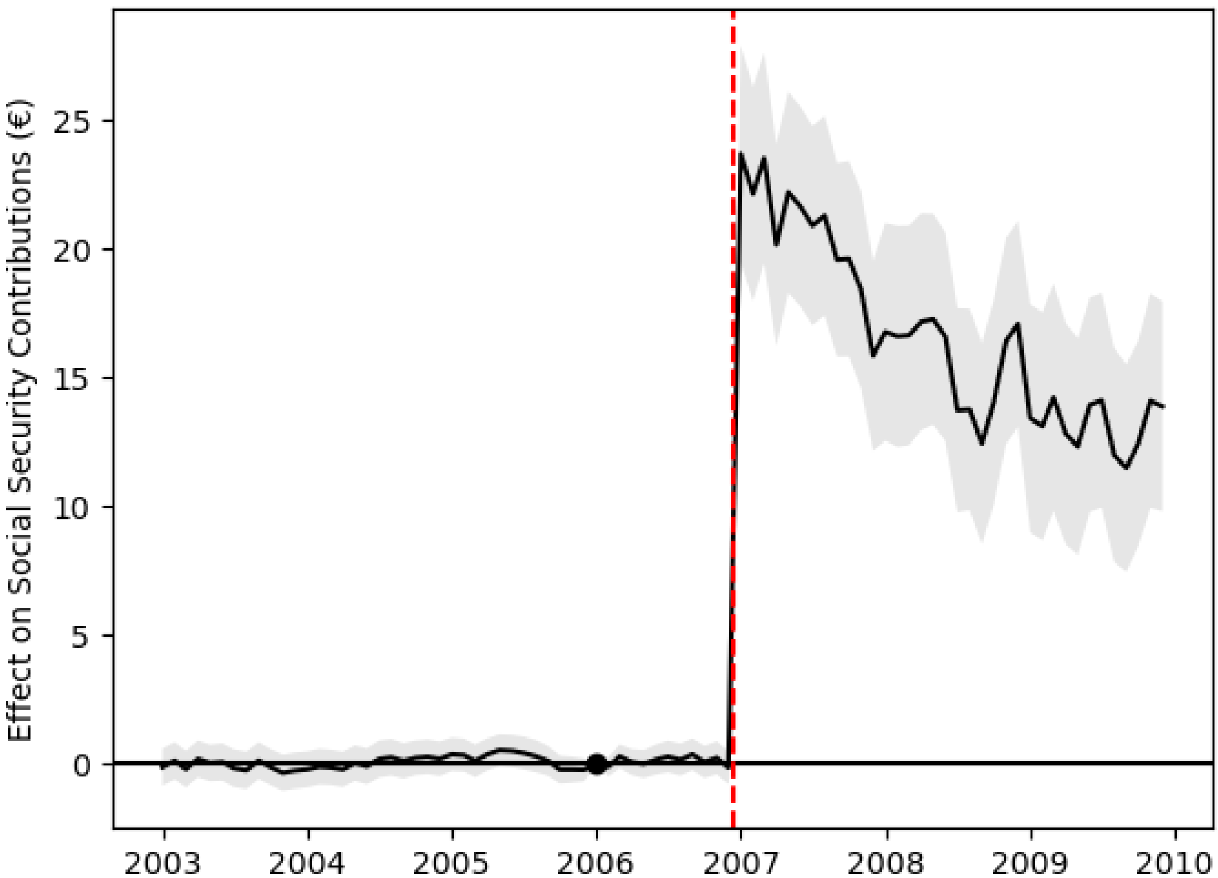
Notes: Social Security Administration data (January 2003–December 2009). This figure overlays all 84 CDFs of policy-relevant firm size from Jan 2003 to Dec 2009 for firms with a policy-relevant firm size between 3 and 15. CDFs prior to Jan 2007 are plotted in orange. Those subsequent to Jan 2007 are plotted in green. Because they overlap, most CDFs are not visible.

Figure 6: Reduced Form Estimates of Threshold on Take-Up



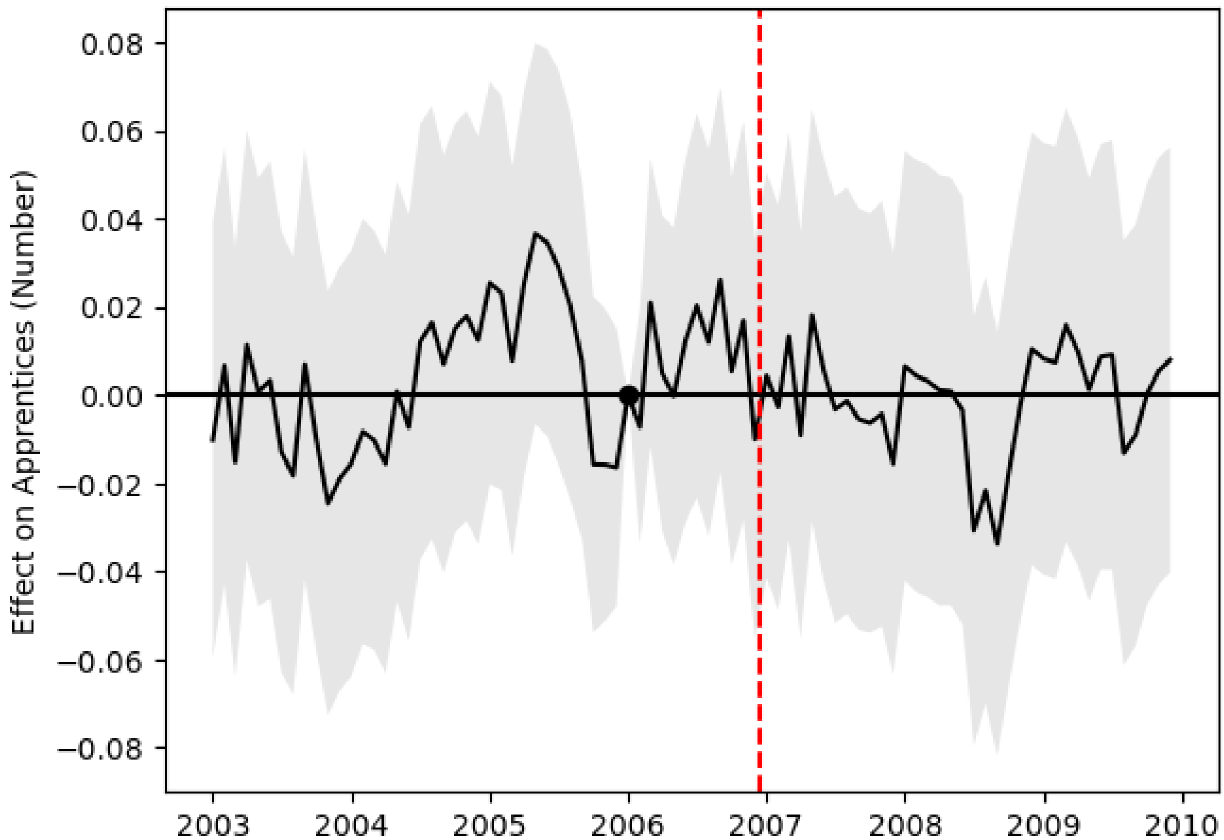
Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in Equation 3 where the outcome variable is an indicator variable for taking the tax break. See Figure 4 notes for details.

Figure 7: Reduced Form Estimates of Threshold on Social Security Contributions



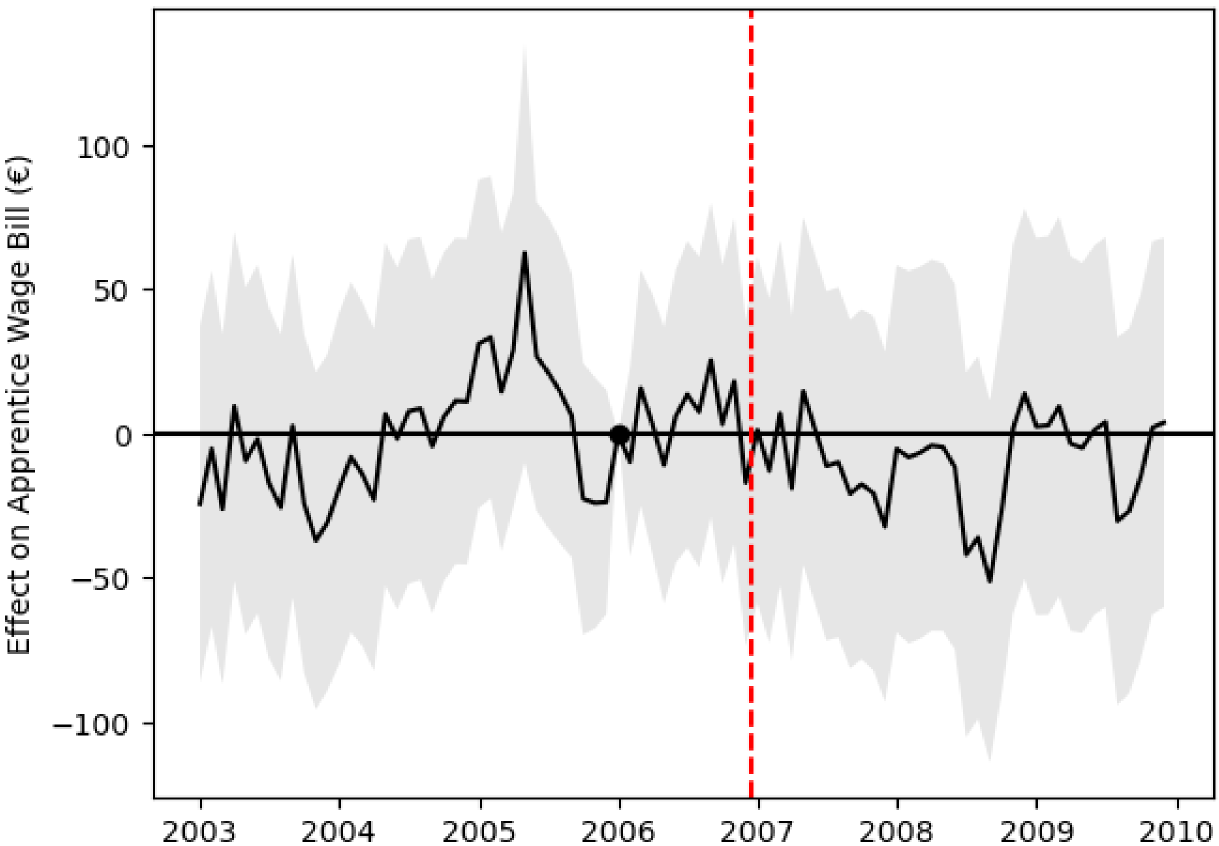
Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in equation Equation 3 where the outcome variable is firm's total social security contribution. See Figure 4 notes for details.

Figure 8: Reduced Form Estimates of Threshold on Number of Apprentices



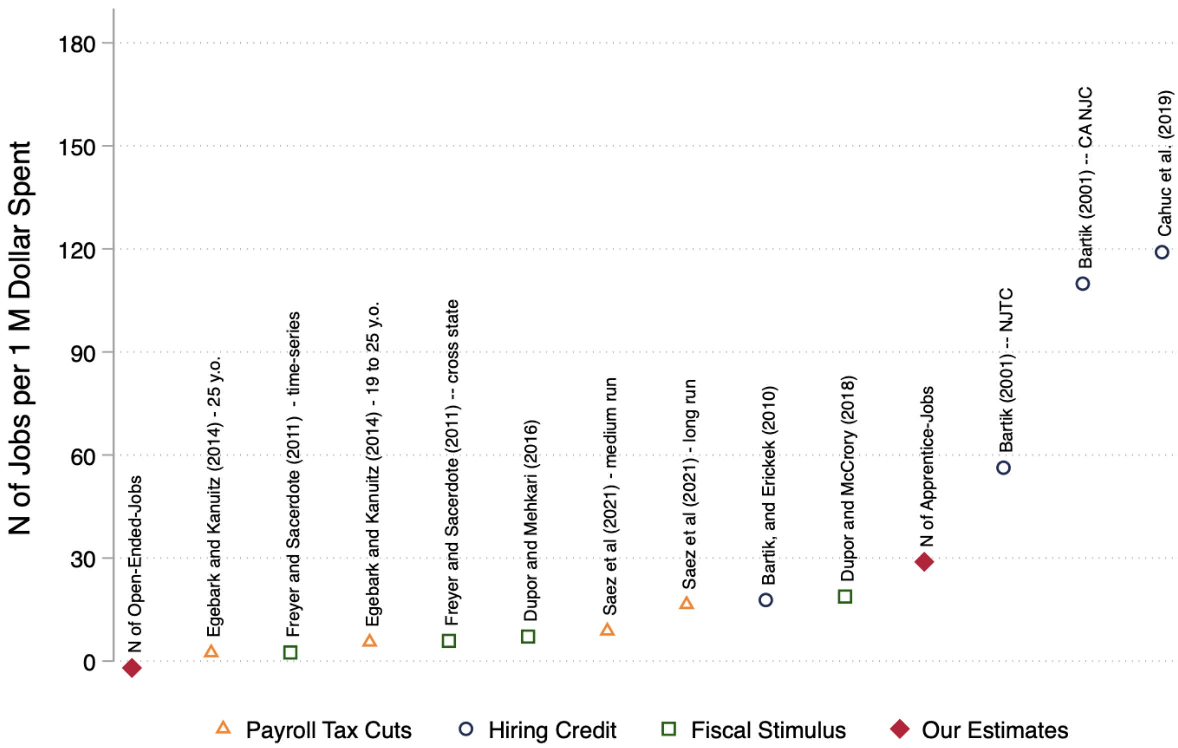
Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in Equation 3 where the outcome variable is the firm's number of apprentices. See Figure 4 notes for details.

Figure 9: Reduced Form Estimates of Threshold on Apprentice Wage Bill



Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in Equation 3, where the outcome variable is the firm's wage bill for their apprentices. See Figure 4 notes for details.

Figure 10: Cost Effectiveness



Notes: This figure reports the number of jobs/job-years supported by 1 million dollars spent. We compare our estimates (red diamonds) with those from other studies on payroll tax cuts (orange triangles), hiring credits (blue circles), and fiscal stimulus (green squares). The estimates of cost-effectiveness for Bartik (2001) and Bartik and Erickcek (2010) are taken from Neumark (2013).

9 Tables

Table 1: Characteristics of Firms in January 2006

	(1) All firms	(2) Firms with apprentices	(3) Firms that ever take-up
Employees	7.088	8.796	7.056
	[3.634]	[3.822]	[3.015]
Full-time Equivalents	6.875	8.606	6.893
	[3.475]	[3.719]	[2.922]
Permanent Workers	6.629	8.322	6.611
	[3.552]	[3.727]	[2.979]
Apprentices	0.427	1.712	0.963
	[0.954]	[1.205]	[1.325]
Temporary Workers	0.429	0.458	0.414
	[1.129]	[1.067]	[1.006]
Seasonal Workers	0.033	0.028	0.039
	[0.478]	[0.449]	[0.515]
All Hires	0.332	0.430	0.312
	[1.354]	[1.640]	[1.122]
New Hires (Under Age 30)	0.109	0.206	0.137
	[0.538]	[0.817]	[0.579]
Apprentice Hires	0.030	0.122	0.066
	[0.261]	[0.512]	[0.352]
Temp Hires	0.072	0.078	0.068
	[0.441]	[0.431]	[0.369]
All Separations	0.182	0.221	0.153
	[0.690]	[0.741]	[0.464]
Separations (Under Age 30)	0.061	0.107	0.070
	[0.309]	[0.414]	[0.293]
Apprentice Separations	0.015	0.060	0.032
	[0.140]	[0.276]	[0.191]
Temporary Separations	0.034	0.036	0.032
	[0.242]	[0.228]	[0.213]
Transformations to Open-Ended	0.009	0.023	0.017
	[0.100]	[.160]	[0.138]
Year Established	1992.400	1993.100	1993.200
	[10.630]	[9.796]	[9.681]
Share Sole Proprietorship	0.217	0.209	0.213
	[0.412]	[0.406]	[0.409]
Share LLC	0.783	0.791	0.787
	[0.412]	[0.407]	[0.409]
N	398,412	99,311	59,670

Notes: Social Security Administration data (January 2006). This table reports the summary statistics for the firms in our sample at baseline (January 2006). The standard deviation is reported in brackets. All statistics are calculated across firm observations.

Table 2: Covariate Balance: Firm Characteristics and Industry Shares

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	t_0-48	t_0-24	t_0	t_0+12	t_0+35	Full	Post	Pre
Panel A: Firm Characteristics								
Age	-0.431	0.064	-0.155	-0.154	-0.297	0.855	0.729	0.958
	(0.223)	(0.198)	(0.199)	(0.216)	(0.231)	<0.824>	<0.884>	<0.554>
General	-0.002	-0.009	-0.003	0.012	-0.003	0.924	1.271	0.640
Partnership	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	<0.673>	<0.128>	<0.974>
LLC	-0.002	0.009	0.001	-0.008	-0.002	0.811	0.799	0.784
	(0.011)	(0.009)	(0.009)	(0.010)	(0.010)	<0.894>	<0.798>	<0.857>
Panel B: Industry Shares								
Agriculture	-0.001	0.001	0.001	0.001	0.001	0.911	1.217	0.697
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	<0.704>	<0.174>	<0.943>
Manufacturing	0.006	-0.005	-0.003	0.006	0.000	1.460	1.711	1.260
	(0.010)	(0.009)	(0.009)	(0.009)	(0.010)	<0.004>	<0.005>	<0.109>
Utilities	0.002	-0.000	0.004	0.001	0.001	1.056	1.077	1.153
	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	<0.343>	<0.346>	<0.219>
Transportation and	-0.013	-0.003	0.002	-0.004	-0.003	1.310	1.659	1.070
Construction	(0.008)	(0.007)	(0.007)	(0.008)	(0.008)	<0.031>	<0.008>	<0.345>
Trading	0.016	-0.001	0.014	0.003	0.012	0.834	0.863	0.852
	(0.008)	(0.007)	(0.007)	(0.008)	(0.008)	<0.861>	<0.702>	<0.754>
Services	-0.008	0.010	-0.010	-0.003	-0.009	1.110	1.007	1.231
	(0.008)	(0.007)	(0.007)	(0.008)	(0.008)	<0.230>	<0.457>	<0.133>
Public Admin, Health,	-0.005	-0.005	-0.006	-0.004	-0.005	0.842	0.673	0.986
and Education	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	<0.847>	<0.932>	<0.499>

Notes: Social Security Administration data (January 2003–December 2009). N=24,532,943. This table reports the effects of being above the threshold (b_t) from the main DD specification in Equation 3 where the outcome variables are general firm characteristics. Each row reports the estimates for a different outcome variable. Estimates are relative to $t_0 - 12$ (January 2006). The first two columns report the pre-reform DD estimates for $t_0 - 48$ (January 2003) and $t_0 - 24$ (January 2004). Columns 3-5 report the post-reform estimates for t_0 (January 2007), t_0+12 (January 2008), and t_0+35 (December 2009), respectively. The last three columns report Wald F-statistics testing the null that all the DD coefficients, the pre-reform coefficients, and the post-reform coefficients are zero, respectively. The dependent variables are firm characteristics and industry dummies in Panels A and B, respectively. Robust standard errors clustering by firms reported in parenthesis. p-values from Wald tests are reported in triangular brackets.

Table 3: Covariate Balance: Regional Shares

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	t_0-48	t_0-24	t_0	t_0+12	t_0+35	Full	Post	Pre
Valle d'Aosta	-0.000	-0.000	0.000	0.001	-0.000	0.798	0.717	0.835
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	<0.911>	<0.895>	<0.781>
Lombardy	0.011	-0.002	-0.004	0.002	-0.006	1.370	1.326	1.377
	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	<0.014>	<0.091>	<0.044>
Piedmont	0.006	0.001	-0.006	-0.001	0.003	0.830	0.805	0.867
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	<0.868>	<0.790>	<0.728>
Liguria	-0.010	-0.006	-0.002	-0.005	-0.001	1.228	1.104	1.332
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	<0.077>	<0.307>	<0.063>
Veneto	-0.001	-0.004	-0.006	-0.006	-0.008	1.018	0.692	1.317
	(0.007)	(0.006)	(0.006)	(0.006)	(0.007)	<0.434>	<0.918>	<0.071>
Trentino-Alto Adige	0.005	0.001	0.004	0.000	0.001	1.012	0.708	1.248
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	<0.449>	<0.904>	<0.119>
Friuli-Venezia Giulia	-0.001	0.006	0.001	0.003	0.003	0.918	0.693	1.124
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	<0.688>	<0.917>	<0.259>
Emilia-Romagna	-0.002	0.004	0.002	0.008	-0.003	0.888	0.772	0.977
	(0.006)	(0.005)	(0.005)	(0.006)	(0.006)	<0.758>	<0.834>	<0.517>
Tuscany	-0.005	-0.002	-0.002	-0.006	-0.011	0.995	1.522	0.547
	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)	<0.493>	<0.023>	<0.995>
Abruzzo	-0.005	-0.002	-0.004	-0.005	-0.005	1.062	1.474	0.841
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	<0.330>	<0.033>	<0.771>
Marche	-0.000	-0.002	0.006	0.003	0.002	1.016	1.182	0.972
	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)	<0.438>	<0.210>	<0.527>
Umbria	-0.000	0.001	0.001	-0.006	-0.001	1.390	1.291	1.423
	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	<0.011>	<0.113>	<0.030>
Molise	0.000	-0.001	-0.001	0.001	-0.001	1.353	1.409	1.362
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	<0.018>	<0.053>	<0.050>
Basilicata	-0.000	-0.001	0.003	0.000	0.000	1.134	1.195	1.127
	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	<0.190>	<0.196>	<0.254>
Lazio	0.005	0.003	0.003	0.005	0.003	0.701	0.838	0.619
	(0.006)	(0.005)	(0.005)	(0.005)	(0.006)	<0.983>	<0.742>	<0.981>
Campania	-0.002	-0.001	0.003	-0.000	0.005	1.087	1.135	0.998
	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	<0.275>	<0.266>	<0.477>
Calabria	0.002	0.003	0.001	0.003	0.007	1.019	1.065	0.969
	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	<0.431>	<0.363>	<0.533>
Sicily	-0.004	-0.001	-0.001	0.000	0.001	0.904	0.681	1.029
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	<0.721>	<0.927>	<0.418>
Sardinia	-0.003	-0.001	-0.004	-0.004	-0.001	1.116	1.073	1.156
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	<0.220>	<0.351>	<0.216>
Apulia	0.005	0.004	0.007	0.007	0.013	0.954	1.027	0.995
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	<0.599>	<0.424>	<0.482>

Notes: Social Security Administration data (January 2003-–December 2009). N=24,532,943. This table reports the effects of being above the threshold (b_t) from the main DD specification in Equation 3 where the outcome variables are region dummies. See notes to Table 2 for details.

Table 4: IV Estimates of Cost-Effectiveness.

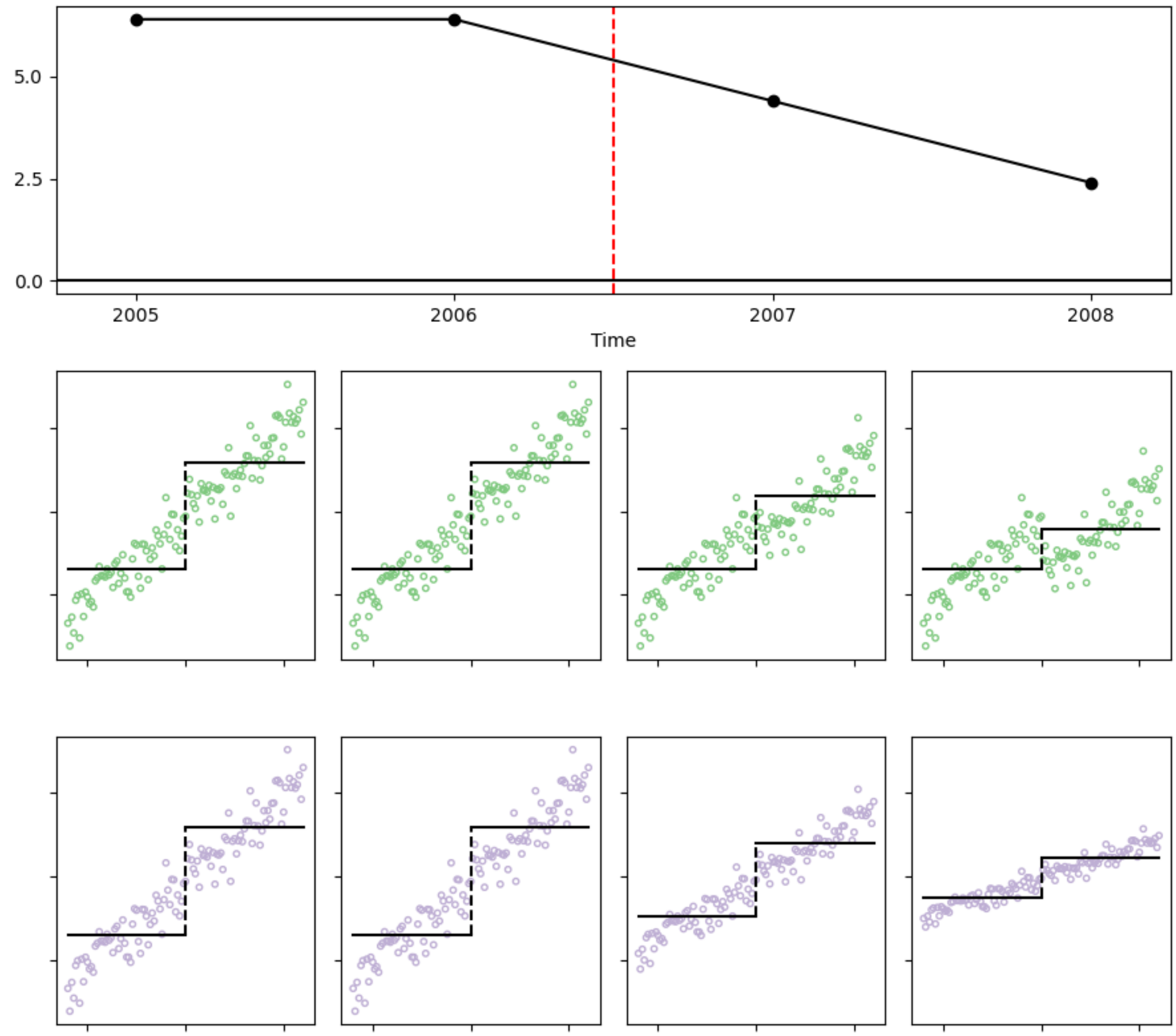
Apprentices per €1M	Total Apprentice Compensation per €1M	Transformations per €1M
347	647,237	-2
(698)	(921,320)	(21)

Notes: Social Security Administration data (January 2003–December 2009). N=24,523,943. This table reports IV coefficient estimates of apprentice jobs supported and apprentice compensation supported (β) per €1M of lost social security contributions from Equation 5. The excluded instrument is a dummy variable for being above the policy cut-off in a month after January 2007. Each IV regression controls for policy-relevant firm size and policy-relevant firm size interacted with being above the threshold in each month, mirroring the reduced-form estimates. The first-stage F-statistic is 230 (see Appendix Table A.5). Robust standard errors clustering by firms reported in parenthesis.

Online Appendix

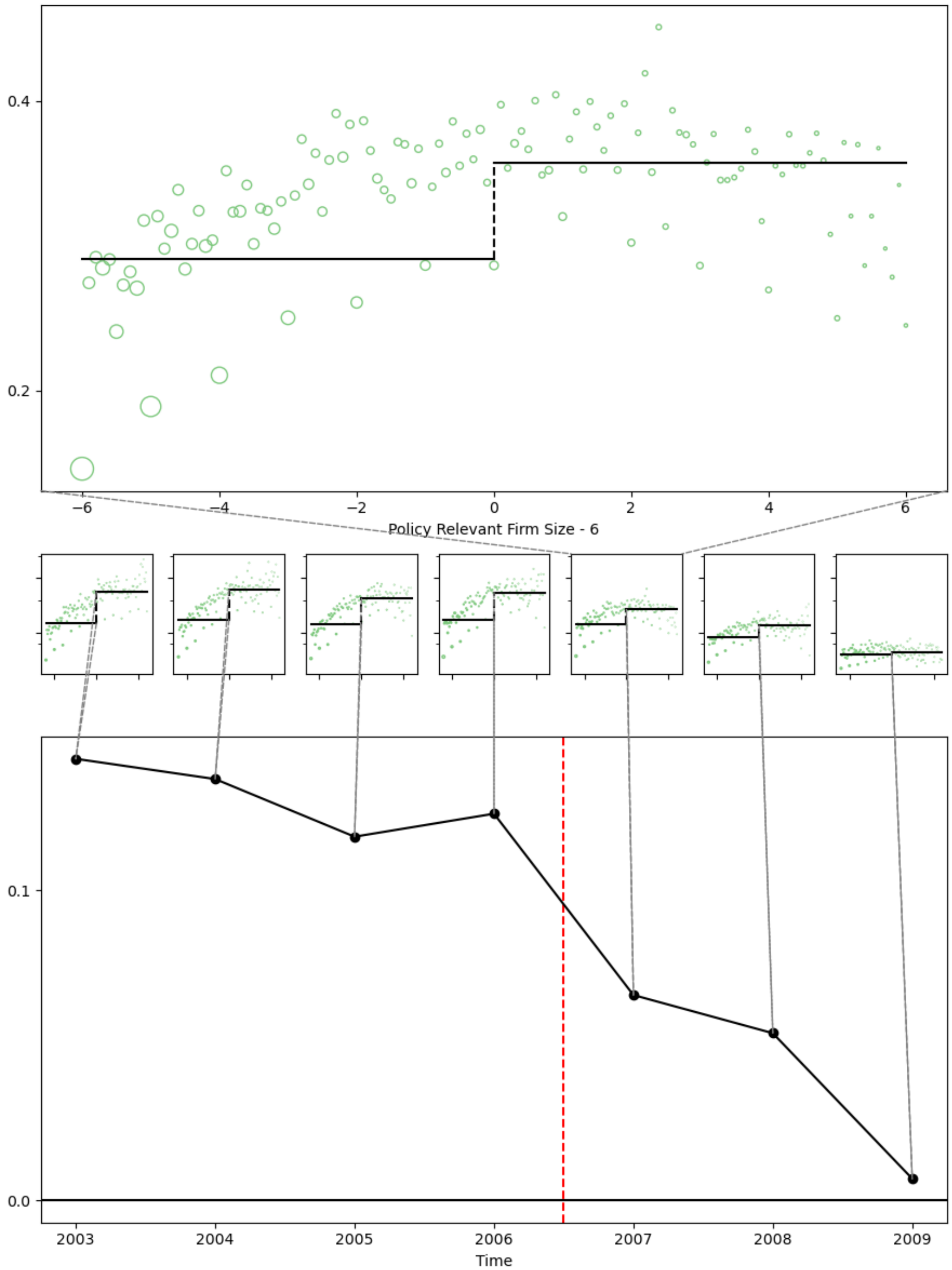
A Appendix Figures and Tables

Figure A.1: Two examples of joint distributions that generate observationally equivalent difference-in-differences estimates.



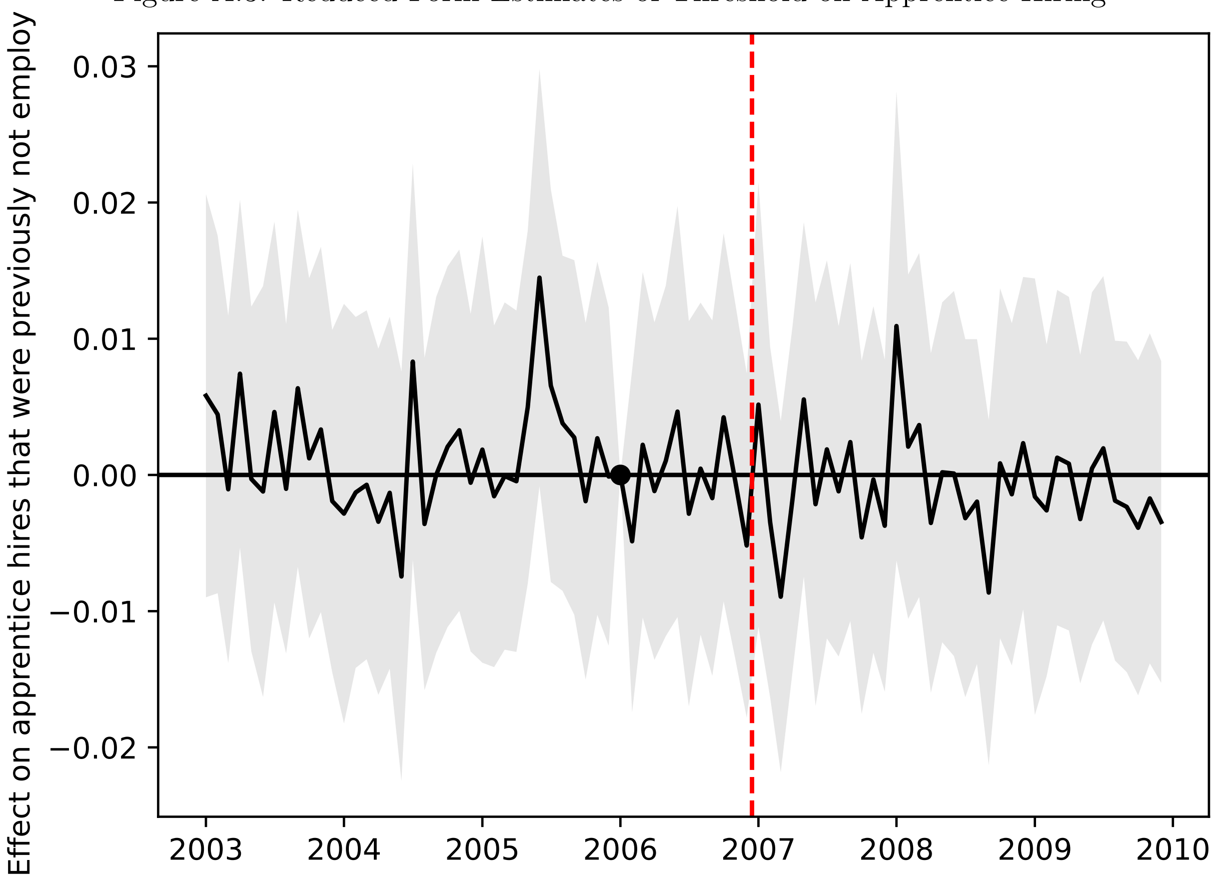
Notes: This figure uses simulated data to show that the same DD estimates can come from two different relationships of the outcome and the targeted characteristic. The top row plots a time series of first difference estimates. A standard DD specification would subtract the difference at a baseline period (e.g. 2006). The second row of figures plots the underlying relationship between the outcome Y and the targeted characteristic Z in green. A discontinuity emerges in 2007 and grows in 2008. The third row plots an alternative relationship between Y and Z in purple that generates the same estimates. The conditional expectation function is stable in the pre-period and only rotates in the post-period. There is little evidence that the outcome changes discontinuously at the targeted threshold.

Figure A.2: Spurious Effects on Apprentice Hiring under Standard Diff-in-Diff Specification



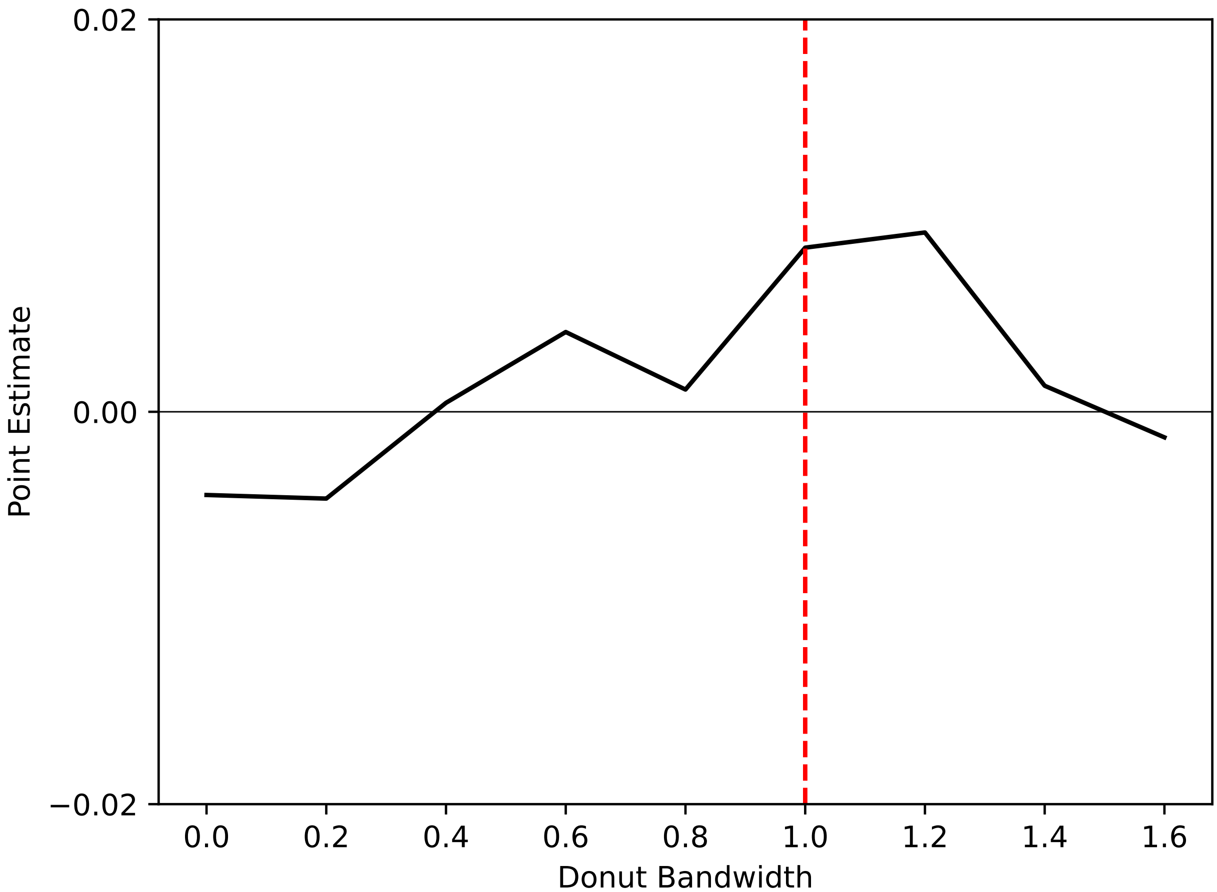
Notes: Social Security Administration data (January 2003–December 2009). This figure decomposes the comparisons made by the standard difference-in-differences specification. Treated firms are those whose average policy-relevant firm size over 2006 is at least 9. The top-panel shows a binned scatterplot of annual apprentice hiring against *average baseline* policy-relevant firm size in 2007, the first year of the policy. The size of the green dots indicates the number of firms within the bin. Means conditional on being in treatment on control—a piecewise zeroth order polynomial fit—are overlayed as black lines. The first panel is zoomed example of the fitted means in each period, shown in the second panel. The third panel, plots a time-series of the mean difference estimates.

Figure A.3: Reduced Form Estimates of Threshold on Apprentice Hiring



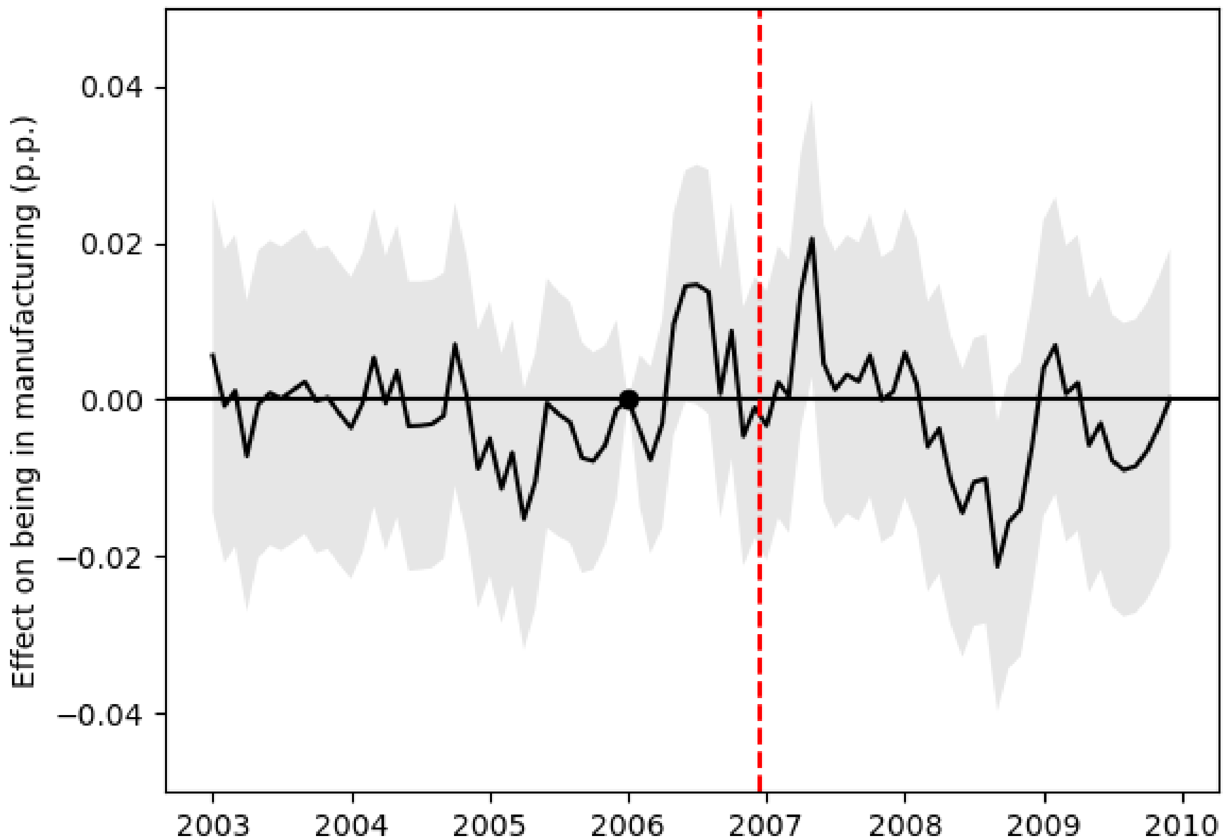
Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in Equation 3 where the outcome variable is new apprentice hires. See Figure 4 notes for details.

Figure A.4: Sensitivity of Apprentice Hiring Effects to Amount of Excluded Data



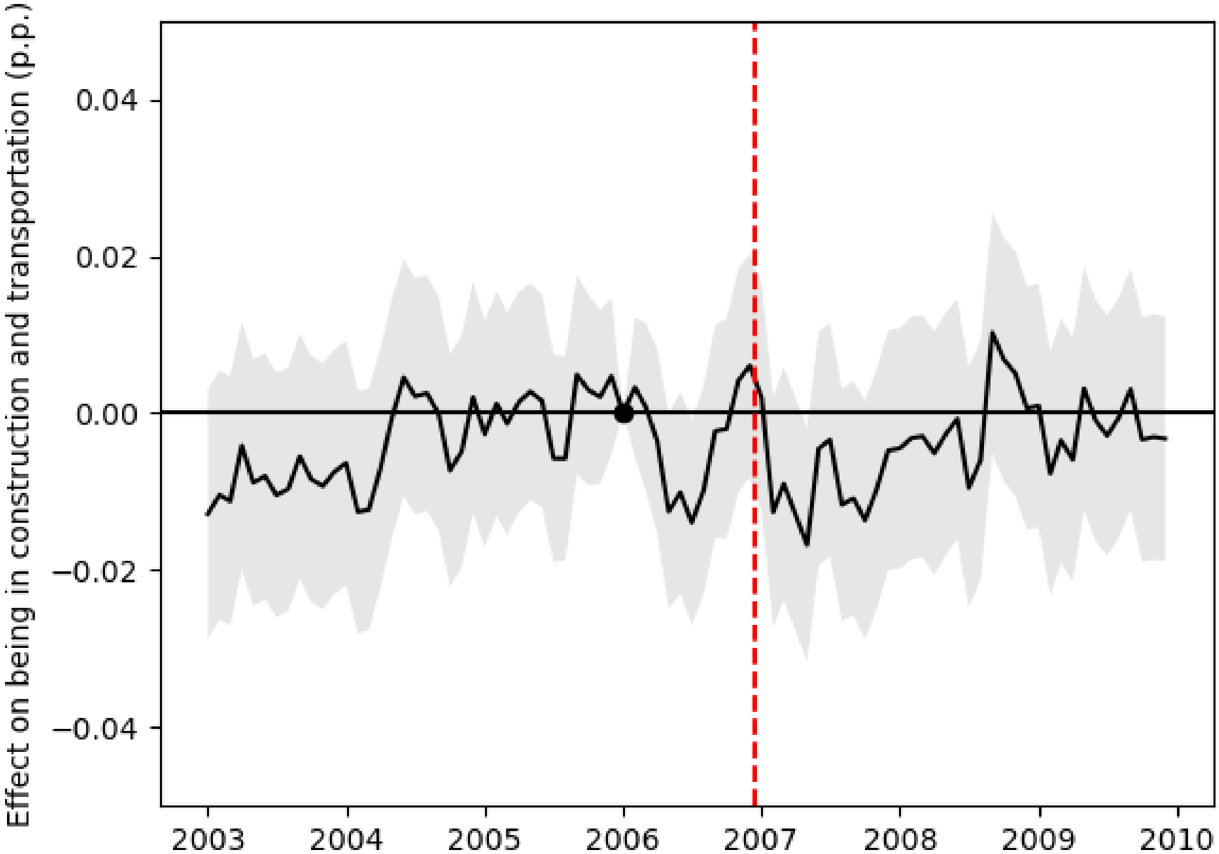
Notes: Social Security Administration data (January 2003–December 2009). This figure shows how the amount of excluded data—the donut bandwidth—affects the coefficient for being above the subsidy threshold in January 2007, ($b_{\text{Jan. 2007}}$) in Equation 3. The outcome variable in this figure is new apprentice hires.

Figure A.5: Reduced Form Estimates of Threshold on Manufacturing Indicator



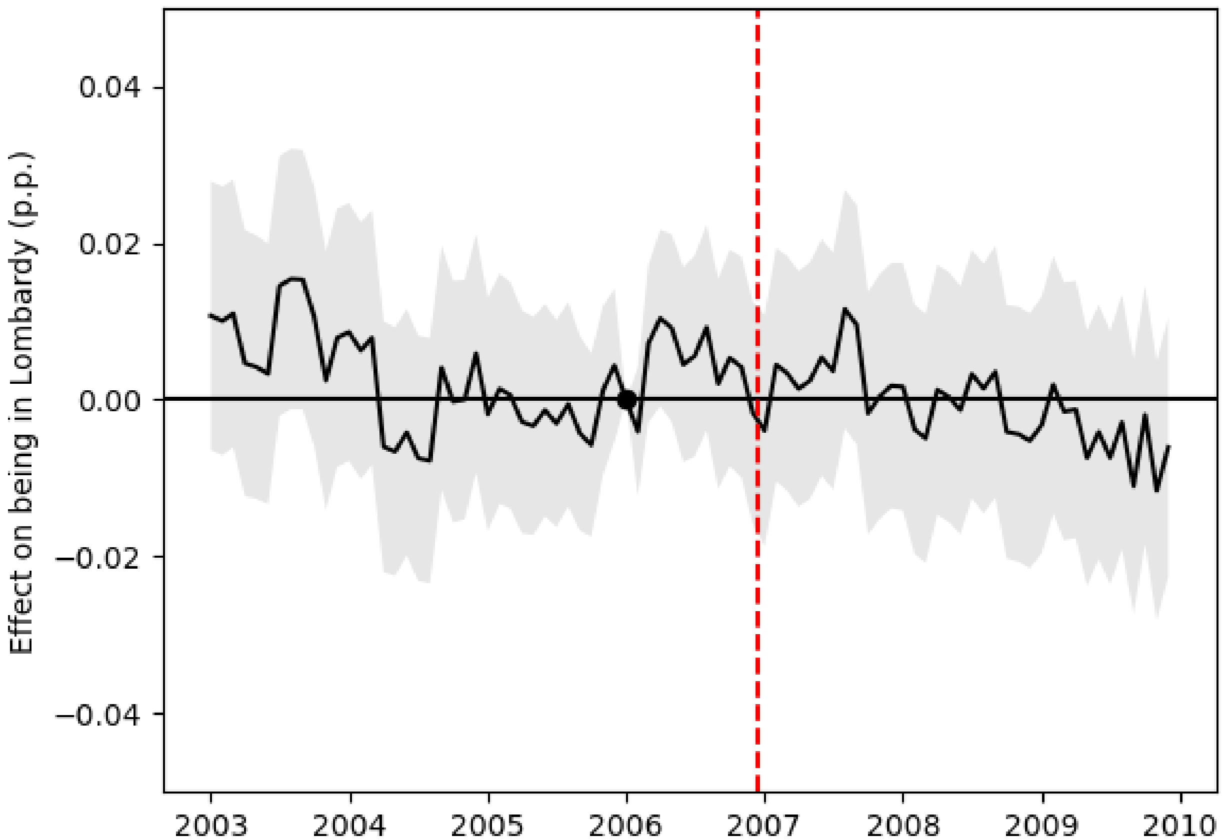
Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in Equation 3, where the outcome variable is an indicator variable for the firm being in manufacturing. See Figure 4 notes for details.

Figure A.6: Reduced Form Estimates of Threshold on Construction and Transportation Indicator



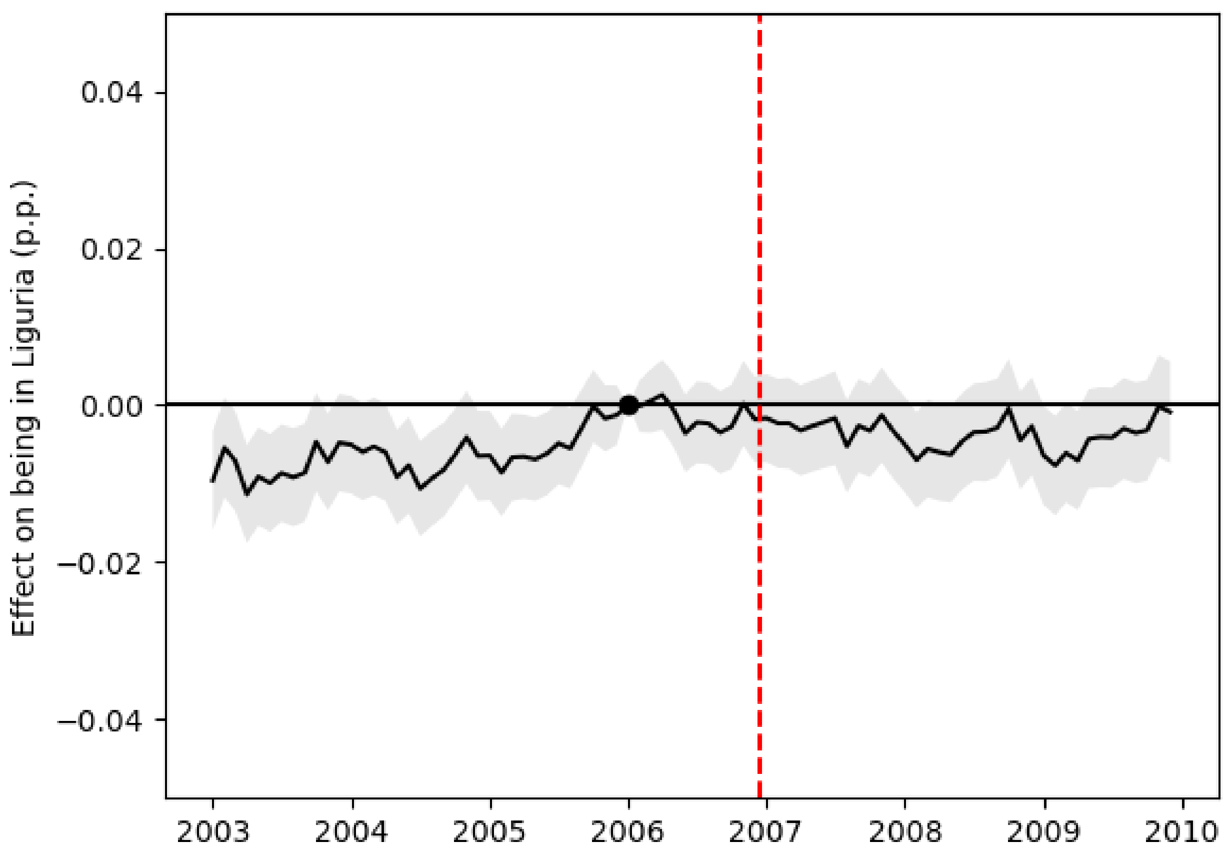
Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in equation Equation 3, where the outcome variable is an indicator variable for the firm being in construction and transportation. See Figure 4 notes for details.

Figure A.7: Reduced Form Estimates of Threshold on Lombardy Indicator



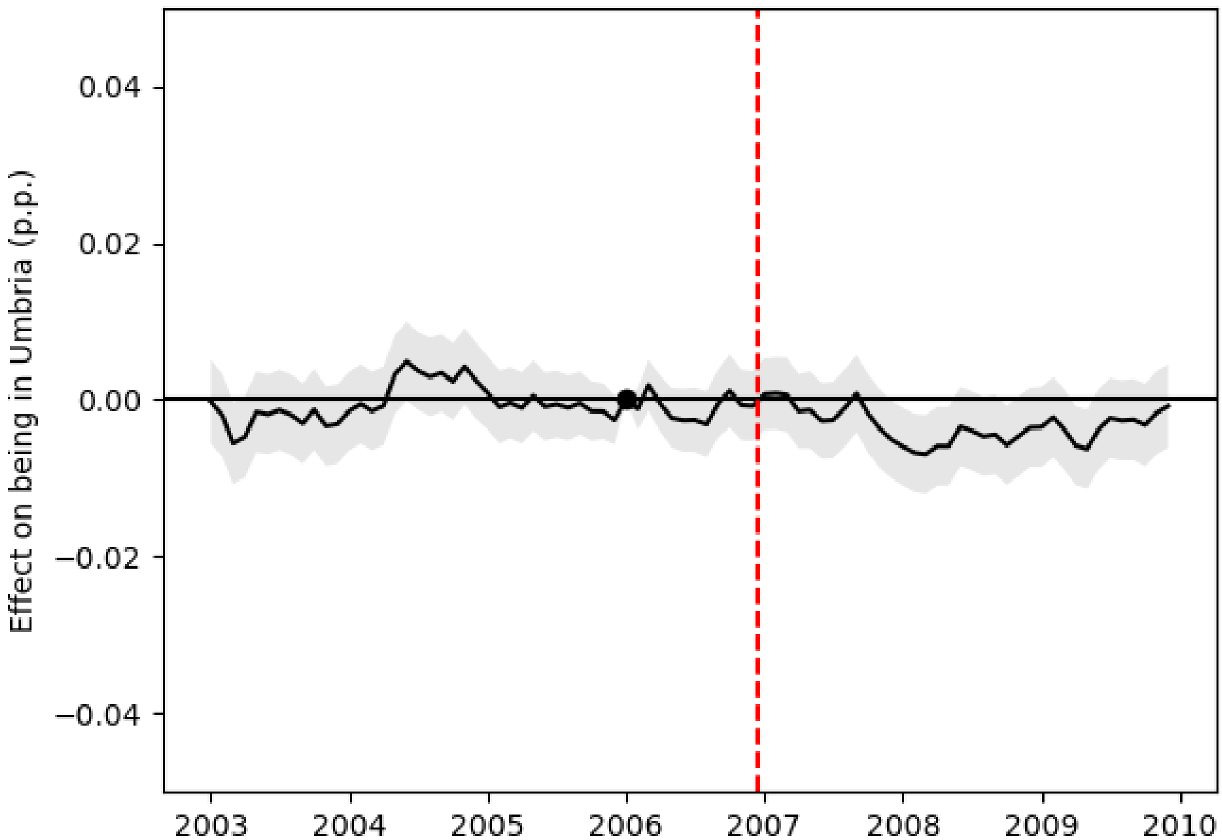
Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in Equation 3, where the outcome variable is an indicator variable for the firm being located in Lombardy. See Figure 4 notes for details.

Figure A.8: Reduced Form Estimates of Threshold on Liguria Indicator



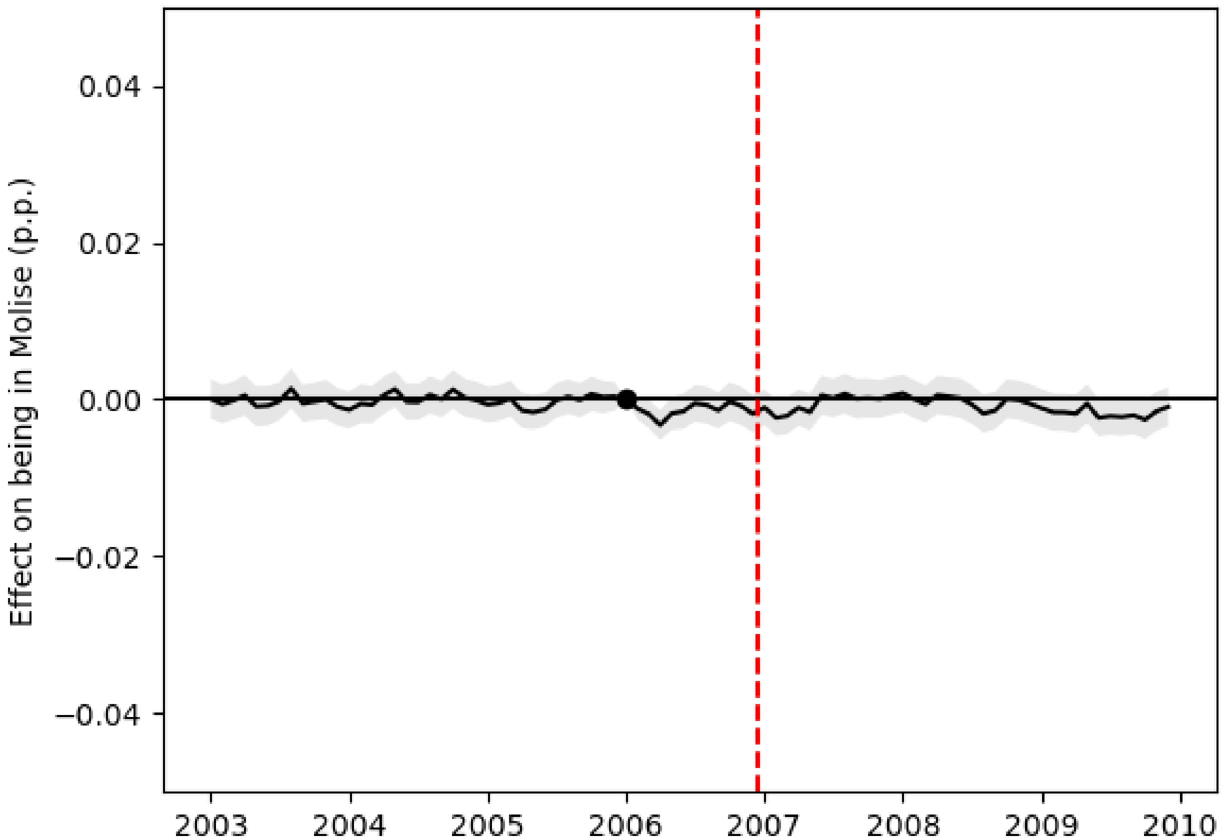
Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in Equation 3, where the outcome variable is an indicator variable for the firm being located in Liguria. See Figure 4 notes for details.

Figure A.9: Reduced Form Estimates of Threshold on Umbria Indicator



Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in Equation 3, where the outcome variable is an indicator variable for the firm being located in Umbria. See Figure 4 notes for details.

Figure A.10: Reduced Form Estimates of Threshold on Molise Indicator



Notes: Social Security Administration data (January 2003–December 2009). This figure reports the effects of being above the subsidy threshold (b_t) from the main DD specification in Equation 3, where the outcome variable is an indicator variable for the firm being located in Molise. See Figure 4 notes for details.

Table A.1: Yearly Social Contributions for the Average Apprenticeship Contract

Apprentice' Tenure	(1) Before Jan 1, 2007 Size> 9	(2) Size≤ 9	(3) After Jan 1, 2007 Size> 9	(4) Size≤ 9	(5) $\Delta_{After-Before}$ Size> 9	(6) Size≤ 9
1	148	148	1200	180	1052	32
2	148	148	1200	360	1052	212
3	148	148	1200	1200	1052	1052

Notes: This table illustrates how yearly social security contributions for the average apprenticeship contract changed in response to the 2007 Budget Bill. Before 2007 all employers paid a fixed weekly fee of 2.85 euros per apprenticeship contract. The yearly social contributions are computed as $2.85 \times 52 = 148.2$ euros. Yearly social contributions for the period after January 1, 2007 are computed as a percentage of yearly earnings and the schedule differs between firms above and below the 9 employee threshold. Social contributions amount to 10% of the apprentice's earnings for firms with more than 9 employees. Firms with 9 employees or less pay 1.5% of the apprentice's earnings in the first year of the contract, 3% in the second year, and 10% in the third year and all the following ones. To compute the change in social contributions implied by this policy, we use the yearly average 2006 apprentice wage, which is equal to 12,000 euros.

Table A.2: Characteristics of Apprentices in January 2006

	(1)
Male	0.585
	[0.493]
Native	0.889
	[0.314]
Age	22.573
	[2.834]
Previously employed	0.983
	[0.128]
Experience	3.736
	[5.035]
Monthly (net) earnings	994.330
	[337.500]
N	430,445

Notes: Social Security Administration data (January 2006). This table reports the summary statistics for the apprentices in our sample at baseline (January 2006). The standard deviation is reported in brackets. All statistics are calculated across apprentice observations.

Table A.3: Industry Composition of Firms in January 2006

	(1) All firms	(2) Firms with apprentices	(3) Firms that ever take-up
Ag., silviculture, fishing, and extraction	0.015	0.004	0.003
	[0.120]	[0.060]	[0.056]
Manufacturing	0.297	0.358	0.327
	[0.457]	[0.479]	[0.469]
Utilities	0.005	0.003	0.003
	[0.072]	[0.053]	[0.051]
Transportation, warehouse, and construction	0.226	0.225	0.219
	[0.418]	[0.418]	[0.414]
Trading	0.205	0.205	0.208
	[0.404]	[0.404]	[0.406]
Services	0.187	0.163	0.195
	[0.390]	[0.369]	[0.396]
Public admin, education, and health	0.032	0.010	0.011
	[0.175]	[0.097]	[0.106]
Other	0.032	0.032	0.033
N	398,412	99,311	59,670

Notes: Social Security Administration data (January 2006). This table reports the summary statistics for the firms in our sample at baseline (January 2006). The standard deviation is reported in brackets. All statistics are calculated across firm observations.

Table A.4: Reduced Form Estimates of Threshold on Other Outcomes

	(1) t_0-48	(2) t_0-24	(3) t_0	(4) t_0+12	(5) t_0+35
Panel A: Flows					
New Apprentice Contracts	0.007	0.002	0.008	0.011	-0.002
	(0.008)	(0.008)	(0.008)	(0.009)	(0.006)
New Apprentice Hires	0.006	0.002	0.005	0.011	-0.003
	(0.008)	(0.008)	(0.008)	(0.009)	(0.006)
Temporary Hires	-0.002	-0.008	0.003	0.007	-0.022
	(0.015)	(0.016)	(0.016)	(0.016)	(0.014)
All Hires	-0.015	0.025	-0.032	0.025	-0.016
	(0.046)	(0.047)	(0.045)	(0.047)	(0.035)
New Hires (Under Age 30)	0.011	0.003	0.004	0.003	-0.004
	(0.019)	(0.018)	(0.018)	(0.018)	(0.014)
Apprentice Separations	0.001	0.004	-0.001	0.002	0.012
	(0.003)	(0.003)	(0.003)	(0.003)	(0.006)
Temporary Separations	0.002	0.002	0.002	0.011	0.018
	(0.006)	(0.007)	(0.007)	(0.008)	(0.016)
All Separations	-0.008	-0.001	-0.037	-0.016	0.050
	(0.019)	(0.019)	(0.019)	(0.020)	(0.039)
Separations (Under Age 30)	0.001	0.004	-0.004	0.002	0.034
	(0.009)	(0.008)	(0.008)	(0.008)	(0.016)
Apprentice Transformations	-0.002	-0.001	-0.003	0.000	-0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
Panel B: Apprentice Characteristics					
Apprentice Avg. Age	-0.524	-0.279	0.080	-0.448	0.395
	(0.331)	(0.338)	(0.350)	(0.353)	(0.546)
Apprentice Avg. Experience	-0.365	-0.656	-0.301	-0.433	-0.253
	(0.311)	(0.316)	(0.333)	(0.331)	(0.473)
Apprentice Male Share	-0.002	0.016	-0.021	0.051	0.048
	(0.058)	(0.057)	(0.055)	(0.056)	(0.083)
Apprentice Native Share	-0.064	-0.011	-0.005	-0.004	0.004
	(0.038)	(0.039)	(0.039)	(0.041)	(0.061)
Apprentice Prev. Employed share	-0.011	-0.015	-0.047	-0.021	0.034
	(0.052)	(0.049)	(0.049)	(0.047)	(0.076)
Apprentice Wage Bill (New Hires)	-11.663	30.117	6.689	-21.862	-119.946
	(46.411)	(46.150)	(48.261)	(49.788)	(84.921)
Panel C: Stocks					
Number of Temporary Workers	-0.034	-0.011	-0.028	0.027	0.007
	(0.032)	(0.032)	(0.034)	(0.036)	(0.042)

Notes: Social Security Administration data (January 2003–December 2009). N=24,532,943. This table reports the effects of being above the threshold (b_t) from the main DD specification in Equation 3, where the outcome variables are general firm characteristics. Each row reports the estimates for a different outcome variable. Estimates are relative to $t_0 - 12$ (January 2006). The first two columns report the pre-reform DD estimates for $t_0 - 48$ (January 2003) and $t_0 - 24$ (January 2004). Columns 3-5 report the post-reform estimates for t_0 (January 2007), $t_0 + 12$ (January 2008), and $t_0 + 35$ (December 2009), respectively. Robust standard errors clustered by firms are reported in parenthesis.

Table A.5: First Stage

	(1) Social Security Contributions
Above × Post	16.411
	(1.082)
N firms	857,587
N obs	24,532,943
F-stat	230

Notes: Social Security Administration data (January 2003–December 2009). This table reports the first stage estimates from the main IV specification in Equation 5. Robust standard errors clustering by firms reported in parenthesis.

B The Policy-Relevant Firm Size

The 2007 Budget Bill does not define how to compute the policy-relevant firm size and delegates this task to the Italian Social Security Agency (INPS). INPS details how to compute the policy-relevant firm size in a provision issued in January 2007 (*circolare n. 22, 2007*). We follow this definition closely.

The firm size that determines the eligibility for the SSC discount is full-time equivalent employment excluding apprentices, temporary agency workers, workers who are on leave (unless the firm hires a substitute), and workers who have been hired with an on-the-job training contract. The types of job training contracts that are excluded from the computation of firm size are those created under the following provisions: exD.lgs.251/2004, D.lgs.n.276/2003, law n.223/1991.

Our rich administrative data contains detailed information on workers' contracts and allows us to construct an accurate proxy for the policy-relevant firm size. In this context, there are two sources of potential measurement error. First, INPS data does not contain a flag for the on-the-job training contracts created under the exD.lgs.251/2004. Anecdotally, this contractual arrangement is very rare and it is unlikely to generate substantial measurement error. Second, our proxy does not account for workers who are on temporary leave (e.g., sick leave or maternity leave).

C Pitfalls of Standard Difference-in-Differences when Program Eligibility is Defined Using a Continuous Variable

This section formalizes the argument that discretizing a continuous treatment in a standard difference-in-differences (DD) approach can inadvertently use variation unrelated to policy changes, leading to erroneous conclusions about the effect of the policy. First, we show that rotations of the conditional expectation function are a form of omitted variable bias in standard DD models. Second, we illustrate that a difference-in-discontinuity approach is robust to rotations of the conditional expectation function over time because it controls flexibly for the running variable in each period. Finally, we illustrate our findings using a concrete example.

C.1 RD *or* Diff-in-diff

We begin by stating the standard fuzzy RD assumptions.

Assumption 1 (Potential Outcomes and Exclusion)**.** *In each period t, each firm draws a pair of potential outcomes, potential choices under treatment, and the running variable $(Y_{it}(0), Y_{it}(1), D_{it}(0), D_{it}(1), Z_{it})$, and the observed outcome is $Y_{it}(D_{it}) = Y_{it}(0) \cdot (1 - D_{it}) + Y_{it}(1) \cdot D_{it}$.*

Assumption 2 (Regression Discontinuity)**.** *Assume:*

1. *Continuity in potential outcomes: $\mathbf{E}[Y_{it}(D_{it}) | Z_{it} = z]$ is continuous in z for each D_{it}*
2. *Continuity in take-up rate: $\mathbf{E}[D_{it}(T_{it}) | Z_{it} = z]$ is continuous in z for each T_{it}*

Local linear regression estimators of regression discontinuity also typically requires that the density of the running variable is continuous. In our setting, firm size bunches at round numbers (Figure 5), rendering infeasible standard RD estimators that compare observed outcomes in a small neighborhood around the discontinuity.

An alternative especially common in the literature on wage subsidies is to apply a difference-in-differences approach, comparing mean differences between large and small firms and subtracting selection bias by measuring pre-existing differences prior to the intervention (see e.g., Cahuc et al., 2019). This approach unwittingly imposes additional assumptions on firms' potential outcomes away from the threshold. To see this formally, consider the standard parallel trends assumption:

Assumption (Strong Parallel Trends)**.** *Assume that potential outcomes can represented by*

$$Y_{it}(0) = a_i + c_t + u_{it}$$
$$Y_{it}(1) = a_i + c_t + b_{it} + u_{it}$$

with u_{it} independent.

The difference-in-differences regression specification masks heterogeneity away from the threshold because it recodes a continuous variable, effectively approximating the conditional expectation function with horizontal lines (Figure A.1, Panel B). Difference-in-differences specifications are often operationalized by estimators derived from saturating indicator variables for time and their interactions with treatment,

$$Y_{it} = a_1 + a_2 T_{it} + \sum_{\tau \neq -12} a_3^\tau \Delta_t^\tau + b_4^\tau (T_{it} \times \Delta_t^\tau) + u_{it}, \tag{6}$$

where Δ_t^τ are dummies for each time period. Parallel pre-trends that check that $b_\tau = 0 \quad \forall \tau < 0$ are testing that u_{it} is mean independent of T_i, $\mathbf{E}\left[u_{it}|T_i\right] = 0$. [17] However, the strong parallel trends assumption also requires that u_{it} is fully independent of Z_{it}. Testing the significance of b_τ does not exhaust the available validity tests of the assumption.

Concretely, let $\mathbf{E}[u_{it}|Z_{it} = z] = g_t(z)$. By assuming that $u_{it} \perp Z_{it}$, a strict parallel trends assumption not only implies parallel trends in intercepts ($\mathbf{E}[g_t(Z)|T] = 0$) but also parallel trends in the slopes of the conditional expectation function of Y given Z ($\mathbf{E}[g_t'(Z)|T] = 0$).[18] If the conditional expectation function rotates over time—a problem that manifested in a previous version of this paper and resulted in substantively different conclusions—then Z_{it} is correlated with u_{it} and is an omitted variable. This can lead one to find no effect with regression discontinuity but find a spurious effect with difference-in-differences.

Figure A.1 simulates two scenarios that produce identical DD estimates. The DD specification cannot distinguish between a treatment effect generated by the discontinuity (green scatter plots) and rotations of the conditional expectation function (purple scatterplots), i.e., the conditional expectation function becomeing more/less flat over time. Failing to isolate variation close to the discontinuity means that RD estimates and DD estimates can diverge, even assuming constant treatment effects.

Notably, many empirical analyses often measure Z_{it} in some base year because it is not subject to manipulation and therefore less "endogenous." However, the conditional expectation function will often regress to the mean, generating a rotation.[19]

C.2 RD *and* Diff-in-diff: Difference in Discontinuities

Even without treatment effect heterogeneity ($b_{it} = b$ in the strong parallel trends assumption), the previous discussion shows how RD and difference-in-differences can yield different estimates. Differences-in-discontinuities rectifies this problem. If changing slopes are an omitted variable, a simple fix is to allow flexibility in the slope of the conditional expectation, isolating variation adjacent to the discontinuity to infer the causal effects of the policy. (One way to view differences-in-discontinuities is as an alternative to local linear regression methods to debiasing RD estimates.)

Formally, we make a weaker parallel trends assumption:

[17] When T_{it} is time-invariant, one can include unit fixed effects to obtain equivalent estimates with greater statistical power.

[18] For the identifying assumption to hold, $Cov[g(Z_{it}) \times \Delta_{it}^\tau, \varepsilon_{it}]$ for any function $g(\cdot)$.

[19] In a simple error-in-variables (white noise) model, $|Cov[Y_{it}, Z_{it}]| < |Cov[Y_{it}, Z_{i0}]|$ for $t \neq 0$.

Assumption 3 (Weak Parallel Trends)**.** *Assume that potential outcomes can represented by*

$$Y_{it}(0) = a_i + c_t + u_{it}$$
$$Y_{it}(1) = a_i + c_t + b_{it} + u_{it}$$

with $\mathbf{E}[u_{it}|Z_{it}] - \mathbf{E}^*[u_{it}|Z_{it}] = d \ \forall t$*, where* $\mathbf{E}^*[\cdot]$ *is a linear projection and* d *is a constant.*

Under this assumption, the curvature in the conditional expectation function of untreated potential outcomes is time-invariant.[20] Whereas the literature on RD has focused on minimizing d by estimating local quadratic regressions and restricting estimation to a narrow bandwidth, we subtract the bias generated by non-linearities using the pre-period.[21]

Combining difference-in-differences with regression discontinuity to exploit variation around the threshold yields model (3) in Section 4.2. Through the lens of this model, the main and interacted terms of Z_{it} can be viewed as omitted variables. The standard DD short regression specification constrains $g_{1t} = 0$ and $g^{\tau}_{2t} = 0$.

C.3 A Cautionary Tale

A previous version of this paper had applied a standard difference-in-differences specification to recover the causal effect of the policy. The paper had also employed the common strategy of defining treatment at baseline to avoid simultaneity bias arising from the "endogenous" choice of firm size. The analysis showed both null pre-trends but large policy effects, suggesting that wage subsidies for apprentices were very cost-effective.

As noted previously, the strategy of defining treatment at baseline to avoid simultaneity bias arising from the "endogenous" choice of firm size can itself induce a rotation from the regression coefficient exhibiting mean reversion. Whereas our difference-in-discontinuities specification is robust to rotations because it isolates variation near the discontinuity, the previous estimates were based upon variation derived from rotations of the conditional expectation function. We believe our original estimates were both too large and spurious.

In Figure A.2, we document a rotation defining Z_{it} in the year prior to the policy and show that a naive analysis of the subsidy policy generates estimates driven by such a rotation.

[20] Unlike other applications of diff-in-discontinuity designs (see e.g. Grembi et al., 2016), we are not trying to subtract the effect of other policies that share the same discontinuity.

[21] A technical literature has emerged to select a bandwidth that balances bias and precision while debiasing the estimates using controls for higher-order polynomials (Calonico et al., 2014). Calonico et al. (2014) Remark 7 notes that conventional point estimates from a quadratic regression specification coincide with their procedure that allows the point estimate and bias correction specifications to be fit on samples with differing bandwidths.

Can a rotating conditional expectation function be causal? A discontinuity at the threshold is generally considered to be "good variation" and strong evidence of policy effects. Nevertheless, it is worth asking whether variation away from the threshold is actually "bad variation." Specifically, if our design focuses on DD estimates just above versus just below the policy threshold, could a design that measures time variation in the slope of the conditional expectation function be consistent with causal effects?

We argue no. Estimating Equation 6 on a rotation would spuriously detect treatment effects in regions without policy variation. Consider the bottom panel of Figure A.1 and conditioning the analysis sample on firms entirely above or entirely below the policy discontinuity. In such a sample, there is no cross-sectional policy variation. However, the differences between large and small firms *within the subsample* are changing over time.

Robustness to over-identifying placebo tests (i.e., estimating the placebo effects moving the policy threshold to the left or to the right of the actual policy threshold) may ameliorate concerns, especially in the case of Figure A.1 when the conditional expectation function is linear. But, if the conditional expectation function exhibits concavity or convexity, a relatively flat portion of the conditional expectation function may rotate less, and the placebo test would fail to find spurious effects.