The Cost of Non-Compliance: Firm Adjustments to Threshold-Based Labor Policies*

Alix Naugler†

January 7, 2026

Abstract

This paper examines the unintended consequences of size-dependent formalization policies that raise the cost of informality for firms, focusing on a provision in Vietnam's Labor Code 2012. The policy sharply increases the financial penalty for firms with at least 10 formally contracted, paid employees that fail to comply with pre-existing labor regulations at this threshold. I develop a profit maximization model illustrating how this policy incentivizes firms to avoid compliance by substituting toward unregulated labor arrangements or partially formalizing. Using Vietnamese micro-, small-, and medium-enterprise panel data, I employ a difference-in-discontinuities design to estimate the causal effect of the increased financial penalty at the threshold. McCrary density tests indicate no evidence of bunching below the 10-formal, paid employee threshold post-policy. Instead, firms adjusted along alternative margins: those just below the threshold increased their reliance on unpaid labor while those just above it registered with the government but continued using informal labor. Firms just above the threshold also realized profit and labor productivity gains. These findings show that threshold-based labor policies can lead to selective—rather than comprehensive—firm formalization, suggesting informality is restructured instead of reduced.

*I am grateful to Jenny Aker, Chris Barrett, Ben Leyden, Dan Maggio, Shubhalakshmi Nag, Elio Nimier-David, Anh Nguyen, Zhuan Pei, Heather Schofield, and Daniela Scur for their helpful feedback. All errors are my own.

†Cornell University. Email: ann34@cornell.edu

1 Introduction

Developing economies are characterized by a large number of small firms, the majority of which operate informally (La Porta & Shleifer, 2014; McKenzie, 2017). Informal firms play an important role in sustaining household incomes, generating employment, absorbing surplus labor, and fostering entrepreneurship; yet they face significant constraints concerning growth, productivity, and formal market integration (De Soto, 1989; La Porta & Shleifer, 2008). Following the conceptual framework of Ulyssea (2018), firm formality can be classified along three mutually exclusive dimensions—fully formal, partially formal, and fully informal—depending on whether firms comply with government registration and labor regulations. A firm's position along these dimensions is endogenous, reflecting its optimization of the costs and benefits associated with full regulatory compliance.

Compared to fully formal firms—registered businesses that adhere to formal labor standards—firms operating along at least one dimension of informality (i.e., partially formal or fully informal) tend to be smaller in size and less productive (La Porta & Shleifer, 2008, 2014; Benjamin & Mbaye, 2012; Amin & Okou, 2020). Because these firms typically lack financial records or legal recognition, they experience restricted access to credit and investment and are generally excluded from public procurement and support programs designed to help augment firm growth and productivity (Perry, 2007). Firm informality also poses challenges for fully formal firms and the broader economy since it allows firms to bypass taxes and regulations, lowering their operating costs. This cost advantage creates unfair competitive pressure and contributes to market fragmentation that weakens demand for fully formal firms' goods and services (Perry, 2007; La Porta & Shleifer, 2014). Informality also reduces tax revenues and productivity. This limits governments' capacity to provide infrastructure, enforce laws, and deliver public services, further reinforcing informality (Prado, 2011).

To address pervasive informality, governments have introduced policies to encourage or enforce firm formalization, aimed to improve regulatory compliance, increase tax revenues, and stimulate local economic growth. However, prior studies show that even when registration costs are removed or incentives (like subsidies or free services) are offered, firm formalization rates remain extremely low (Bruhn & McKenzie, 2014). This suggests that governments may overestimate the attractiveness of formalization to firms and underestimate its perceived costs. For firms, potential benefits of formalization include improved access to formal credit markets and greater opportunities to secure government contracts (Perry, 2007; La Porta & Shleifer, 2014); yet substantial costs include higher taxes, increased labor expenses, and administrative burdens (De Soto, 1989; Bruhn & McKenzie, 2014; Benhassine et al., 2018). If the expected benefits do not outweigh these costs, it pushes firms to partially formalize or be entirely informal. Such trade-offs can create size-dependent distortions, as firms strategically limit their size to avoid crossing over regulatory thresholds that impose higher compliance costs. This behavior reinforces the persistence of small firms in developing economies.

This paper analyzes how Vietnamese firms respond to an updated formalization policy anchored in a pre-existing firm-size threshold. Under Article 119 of Vietnam's Labor Code 2012, firms with 10 or more formal, paid (FP) employees are required to develop, submit, and update internal work regulations (IWRs) with the government. This policy substantially increased the financial penalty for non-compliance—raising fines five- to 10-fold—in an effort to create a more credible economic deterrent. Using panel data from the UNU-WIDER Vietnam Small- and Medium-Enterprise (SME) database, I explore how the increased financial penalty shapes firm behavior around the threshold, including adjustments in formal labor force size, labor composition, and formality status. Spillover effects on firm revenue, profitability, productivity, and costs are also assessed to determine whether regulatory avoidance imposes broader economic consequences. Moreover, these data support classifying firms as fully formal, partially formal, or fully informal based on their registration status and use of formal labor contracts. This allows for an examination of whether firms selectively comply with only *certain* dimensions of firm formality to minimize regulatory compliance costs.

Methodologically, this study employs a difference-in-discontinuities (DiDisc) design, which combines the cross-sectional precision of regression discontinuity design (RDD) with the temporal variation of difference-in-differences (DiD). This strategy identifies the causal effect of the increased financial penalty by comparing firms just above and just below the regulatory threshold before and after policy implementation. Conventional RDD methods are inappropriate in this setting because the running variable—the number of FP employees—is discrete, highly left-skewed, and characterized by a pronounced mass point. A standard DiD approach is also ill-suited since it averages across heterogeneous firms and cannot isolate the policy-relevant threshold margin. I therefore implement the DiDisc design using a local randomization procedure that selects optimal symmetric and asymmetric windows around the cutoff via covariate balance tests. These data-driven bandwidths define estimation samples in which pre-policy firm characteristics are statistically indistinguishable across the threshold, allowing for credible causal inference on firms operating at the margin.

To interpret firm responses to the higher threshold-induced cost, I develop a profit maximization model in which firms choose both their labor composition—across FP, informal paid, and unpaid workers—and formality status subject to compliance costs. The model underscores that the policy introduces a discontinuous jump in expected costs through a higher financial penalty, yielding three testable predictions. First, if the financial penalty for non-compliance is sufficiently large, firms will "bunch" below the 10-FP employee threshold to avoid these costs (Theorem 1). Second, firms just below the threshold will substitute toward unregulated workers as a margin of evasion (Theorem 2). Third, firms just above the threshold will adopt hybrid strategies: registering with the government while continuing to rely on informal labor, thereby reducing compliance costs without constraining firm size (Theorem 3). These predictions guide the empirical analysis that follows.

Did firms endogenously adjust their formal labor force to circumvent compliance requirements at the 10-FP employee threshold? To evaluate this prediction from Theorem 1, I apply McCrary (2008) density tests to assess whether the firm distribution around the threshold changed between the pre- and post-policy periods using the optimal windows. Results show no significant discontinuity post-policy despite the higher financial penalty. The absence of such bunching behavior provides evidence that most firms did not find it optimal to manipulate their FP employment to be just below the threshold. Firms instead adapted along alternative margins, consistent with Theorems 2 and 3.

Indeed, subsequent analysis employing the DiDisc methodology reveals distinct behavioral responses to the policy on either side of the threshold. Firms just *below* the 10-FP employee threshold increased their reliance on unpaid workers, likely reflecting an attempt to maintain operational capacity while remaining outside the regulatory boundary to avoid compliance costs. This is consistent with Theorem 2, which predicts that higher threshold costs induce substitution toward unregulated labor. By contrast, firms just *above* the threshold decreased their use of unpaid workers and experienced significant increases in gross profits and labor productivity. This suggests that, despite facing higher compliance costs, these firms achieved efficiency improvements—possibly through scale effects or labor force restructuring. Yet consistent with Theorem 3, these firms did not fully formalize. They instead adopted a partially formal status: registering with the government while continuing to leverage informal labor arrangements. This behavior reflects strategic non-compliance, in which firms selectively satisfy more visible dimensions of formality but evade the costlier or less enforceable ones. In doing so, they minimize regulatory exposure *and* preserve operational flexibility.

Several studies analyze the distortionary effects of size-dependent government policies on firms (Garibaldi et al., 2004; Guner et al., 2006; Schivardi & Torrini, 2008; Candela, 2013; Gourio & Roys, 2014; Garicano et al., 2016; Benedek et al., 2017; Dabla-Norris et al., 2018; Amirapu & Gechter, 2020; Mulligan, 2020; Bertrand et al., 2021; Padmakumar, 2022; Qian & Vereshchagina, 2022; Aghion et al., 2023). This literature consistently finds that such policies impact firm growth, productivity, resource allocation, innovation, and gender discrimination in developed and developing economies. Garibaldi et al. (2004) and Schivardi and Torrini (2008) show Italian firms close to a 15-employee threshold—which triggers stricter dismissal protections—reduce their growth propensity. Gourio and Roys (2014) and Garicano et al. (2016) find similar effects in France, where firms with more than 50 employees face additional labor requirements. In India, Amirapu and Gechter (2020) show a 10-employee threshold raises labor costs and induces bunching while Padmakumar (2022) finds a 100-employee threshold drives substitution toward capital and temporary labor inputs.

This research introduces three contributions to the literature on firm responses to size-dependent government regulations. First, it leverages a unique policy environment to provide evidence on the deterrence effect of an explicit financial penalty associated with a threshold-based labor policy. Prior studies demonstrate that firms tend to strategically adjust their size and labor composition to avoid crossing costly regulatory thresholds (Garibaldi et al., 2004; Schivardi & Torrini, 2008; Garicano et al., 2016; Aghion et al., 2023; Padmakumar, 2022). However, these studies focus on thresholds that generate *implicit* compliance costs—such as administrative burdens, tax obligations, or mandatory labor protections. None examine the behavioral response to an *explicit* economic consequence tied to non-compliance with a firm-size threshold. Comparatively, Vietnam's Labor Code 2012 includes a revised provision that increases the financial penalty for non-compliance five- to 10-fold for firms with 10 or more FP employees who do not submit IWRs. This setting provides a unique empirical opportunity to evaluate how firms respond to formalization incentives when non-compliance is not merely symbolic but also economically costly. It isolates the deterrence effect of financial penalties by contrasting responses to perceived economic costs with those to general compliance burdens.

Second, this paper advances empirical strategies for studying firm-level distortions by applying a DiDisc design. Although this method has recently gained traction (Takahashi, 2024), it remains underused relative to more established empirical approaches. Several studies (Gourio & Roys, 2014; Garicano et al., 2016; Padmakumar, 2022) employ structurally calibrated models to stimulate firm responses under counterfactual policies. While useful for capturing general equilibrium dynamics, such models depend on strong functional form assumptions and extensive data, often unavailable in developing countries. Existing studies also use reduced-form techniques but they do not integrate cross-sectional and temporal variation in a unified framework. By contrast, the DiDisc identification strategy used here enables direct estimation of the causal effects of policy changes at size-based thresholds without simulation or restrictive modeling assumptions. This is advantageous in settings like Vietnam, where localized policy variation and repeated survey rounds can be credibly exploited.

Lastly, this research enhances understanding of firm informality by adopting a multi-dimensional definition based on both government registration and labor contract compliance. Prior research uses binary classifications, typically tied to registration status alone (Perry, 2007; La Porta & Shleifer, 2008; Dabla-Norris et al., 2018; Padmakumar, 2022) which overlooks variation in *how* firms comply with labor regulations. Guided by the conceptual framework used in Ulyssea (2018), I distinguish among fully formal, partially formal, and fully informal firms to identify intermediate stages of formalization—such as formal firms that operate informally or vice versa. By capturing these hybrid strategies, the analysis uncovers forms of strategic non-compliance that dichotomous definitions obscure. Therefore, this approach provides a more nuanced and policy-relevant characterization of how firms adjust their formality status in response to threshold-based labor policies.

The remainder of this paper proceeds as follows: Section 2 introduces the country context and the threshold-based labor policy. Section 3 describes the data and variable construction. Section 4 presents the profit maximization model and its testable predictions. Section 5 outlines the empirical strategy, including the DiDisc design, the outcome variables, and the procedure for selecting optimal bandwidths. Section 6 reports the results and explores mechanisms. Section 7 concludes.

2 Country & Policy Setting

Since Vietnam's liberalization movement in 1986, the country has transitioned from a centrally planned economy to a market-oriented one characterized by rapid industrialization and sustained economic growth (Rand & Tarp, 2020). As part of the transition, the government prioritized small- and medium-enterprise (SME) development and business environment reform to catalyze structural change. In 2020, SMEs accounted for 95 percent of all firms, half of the labor force, and 40 percent of GDP in Vietnam (Rand & Tarp, 2020). Yet despite policy efforts, significant constraints—lack of a skilled labor force, limited credit access, and an excessive informal sector—continue to hinder SME entrepreneurial activity and growth (Angelino et al., 2021). Consequently, many SMEs still straddle the boundary between formality and informality, limiting their full economic potential.

A specific attempt of the Vietnamese government to directly encourage firm formalization among SMEs is Decree No. 10/2012/QH13, also known as Labor Code 2012. It is a detailed legal framework that governs labor relations including employment contracts, dispute resolution, wages, and working conditions. Passed in June 18, 2012 and then implemented in May 1, 2013, it is designed to protect workers' rights and promote fair labor practices across firms. Article 119 of Labor Code 2012—hereafter referred to as the 2012 employee threshold policy—specifically requires firms with at least 10 employees to prepare, issue, and then submit internal work regulations (IWRs) to a local Ministry of Labor, Invalids, and Social Affairs (i.e., MOLISA). According to Article 3 of Labor Code 2012, an "employee" is defined as someone "15 years or older, has the ability to work, works under a labor contract, is paid with wage and is managed and controlled by an employer." Thus based on the 2012 employee threshold policy, a firm's size is determined by the number of working-age individuals who are both formally contracted and paid. Unpaid workers and paid workers without a formal contract (i.e., informal workers) are not considered part of a firm's labor force according to the government. For simplicity, workers that satisfy the employee definition outlined in Labor Code 2012 (including the 2012 employee threshold policy) will be referred to as "formal, paid (FP) workers" in this paper. While this 10-FP employee threshold is an exogenous policy parameter, a firm's decision to adjust their size in terms of its number of FP workers is endogenous.

The 2012 employee threshold policy dictates that IWRs submitted by firms must be consistent with current labor laws and include employee protocols related to workplace order, occupational safety, working hours, and disciplinary measures specific to their firm. Any firm that fails to create IWRs when it employs at least 10 FP workers, to register such IWRs with the appropriate provincial MOLISA, or to update expired IWRs is liable to pay a financial penalty contingent on inspection.[1] The economic consequence associated with the 2012 employee threshold policy ranges from 5 to 10 million Vietnamese Dong (VND) (i.e., 240-475 USD) per infraction.[2]

The previous iteration of this legal framework was Labor Code 1994 which included the same 10-FP employee threshold requirement in its respective Articles 82 and 83; however, the financial penalty for non-compliance was 1 million VND (i.e., 47 USD).[3] Thus, the updated provision imposes a financial penalty that is five- to 10-times larger. Notably, this provision is the *only* one in Labor Codes 1994 and 2012 that explicitly links regulatory obligations to firm size in terms of the number of employees, regardless of how "employee" is defined. This enables a clearer attribution of observed changes in firm size to this provision itself, reducing the risk of confounding with other regulations.

3 Data

This research uses panel data from the UNU-WIDER Vietnam Small- and Medium-Enterprise (SME) database which includes private formal and informal manufacturing firms surveyed biennially

[1]The fine associated with a firm failing to comply with the 2012 employee threshold policy is specified in Article 15 of Decree No. 95/2013/ND-CP, which was implemented on October 10, 2013.

[2]Conversion to USD used the 10/10/2013 VND to USD exchange rate, coinciding with the day that Decree No. 95/2013/ND-CP was implemented.

[3]The fine associated with a firm failing to comply with Articles 82 and 83 of Labor Code 1994 is specified in Article 20 of Decree No. 38/CP, which was implemented on July 1, 1996. Conversion to USD used the 10/9/2013 VND to USD exchange rate, aligning with the day *before* Decree No. 95/2013/ND-CP was implemented.

from 2005 to 2015. These firms are sampled from 10 Vietnamese provinces and distributed across roughly 18 manufacturing sectors.[4] Data was collected from June to August for each survey year for approximately 2,500 firms. Each survey captures firm data across two distinct time frames. The first includes modules that document firm characteristics at the time of the survey, such as formality status, owner or manager attributes, sales structure, investment activity, credit history, production technology and practices, and perceived constraints. The second focuses on retrospective, year-end data, including firms' annual economic accounts (e.g., sales revenue, gross profit, labor costs, debt, etc.), employment levels, and labor force composition for the previous calendar year.

Additionally, the dataset supports classification of firms into fully formal, partially formal, and fully informal dimensions.[5] Firms that have an Enterprise Code Number (ECN) and provide all (i.e., 100 percent) of their full-time workers with a formal, written labor contract are considered fully formal.[6] These firms are both registration and labor contract compliant. By contrast, firms that do not have an ECN and fail to provide all (i.e., less than 100 percent) of their full-time workers with a formal, written labor contract are considered fully informal.[7] These firms are neither registration nor labor contract compliant. Firms that comply with only one of the two are considered partially formal. Specifically, firms that have an ECN but fail to provide a formal, written labor contract to all their full-time workers are referred to as formal firms that operate informally. Conversely, firms that do not have an ECN but do provide a formal, written contract to all their full-time workers are referred to as informal firms that operate formally. Hence, a firm's formality status is endogenous because it is determined by the firm's *own* registration and labor contract compliance decisions.

The survey samples are stratified by ownership type to capture the full range of legal structures among private SMEs (e.g., registered households, cooperatives, limited liability companies, etc.). The sampling scheme for this dataset is based on a representative sample of registered firms drawn from the General Statistics Office of Vietnam enterprise census information. However, this census only includes firms registered with the government. Unregistered firms that are a part of the sample were identified "on-site" as they operated along-side registered firms that were surveyed. A random selection of these unregistered firms were included in the sample in each survey year. Therefore, the sample of informal firms is not representative of the country's informal manufacturing sector. This limitation affects the external validity of the results. Nevertheless, it is not a major concern as the analysis only focuses on within-in firm changes over time, causal relationships, and mechanisms—not generalizations to Vietnam's informal manufacturing sector.

In reference to the 2012 employee threshold policy, the survey does not explicitly ask firms to report their exact number of formal, paid (FP) workers for the previous calendar year's year-end.

[4] Provinces were not selected randomly but included the country's main urban cities and specific rural areas.

[5] I extend the conceptual framework from Ulyssea (2018) by operationalizing firm formality along two dimensions: government registration and labor contract compliance.

[6] An Enterprise Code Number (ECN) is a unique identifier for registered firms in Vietnam, combining each firm's Business Registration Certificate and Tax Code numbers as mandated by Decree No. 43/2010/ND-CP.

[7] Appendix Table C.1 presents the year-end distribution of firms providing formal, written labor contracts to their full-time workers using relevant survey years from the UNU-WIDER Vietnam SME database. Across time, 78 to 81 percent are consistently *not* labor contract compliant, with the remainder split between full and partial compliance.

Instead, I estimate this metric independently using the survey's question structure and sequencing. Enumerators instructed firms to separately report their number of full-time, part-time, and casual workers, along with the percentage of their total workers who were unpaid and the percentage of their full-time workers with formal, written labor contracts. Using this information, I estimate the number of FP *full-time* workers for each firm i at time t as defined in Equation 1:[8]

$$FP_{it} = C_{it} \times FT_{it} \times (1 - U_{it}) \tag{1}$$

where $C_{it} \in [0, 1]$ is the percentage of full-time workers with a formal, written labor contract; FT_{it} is the number of full-time workers; and, $U_{it} \in [0, 1]$ is the percentage of total workers that are unpaid. Note that total workers is the summation of full-time, part-time, and casual workers.[9]

Article 3 of Labor Code 2012 defines "employee" without any reference to the *amount* of time worked, making no distinction between full-time, part-time, or casual workers. Because the data can only determine the percentage of *full-time* workers that have a formal, written labor contract, only the number of FP *full-time* workers can be estimated for the sample (see Equation 1). The dataset does not include survey questions that can be used to estimate the number of formally contracted part-time or casual workers. Although excluding part-time FP workers may slightly underestimate a firm's size in terms of its total FP employment—which defines the firm-size threshold in the 2012 employee threshold policy—this does not undermine the analysis.[10] This is because FP full-time workers have a higher probability of driving firm compliance decisions and performing compliance tasks compared to their part-time counterparts given their more stable and consistent roles in the firm as well as their greater responsibility and involvement in operational and regulatory processes (Mulligan, 2020; Morikawa, 2023). Hence, using the number of FP full-time workers still strongly aligns with the policy's intent to capture a firm's substantive employment obligations.

Figure 1 illustrates the timing of Labor Code 2012's passage and implementation relative to the UNU-WIDER Vietnam SME survey years. The timeline raises a potential concern about using year-end 2012 and 2014 firm data to represent the pre- and post-policy periods, respectively, due to the possible anticipatory affects during the interim period. Given the year-long gap shown in Figure 1, firms may have preemptively adjusted their behavior *before* the 2012 employee threshold policy's formal enforcement to mitigate the risk of future financial penalties associated with non-compliance.

To identify the appropriate pre-policy period (i.e., year-end 2010 or 2012), I calculate the number of (1) FP full-time workers, (2) informal, paid full-time workers, and (3) unpaid full-time workers

[8]Equation 1 presents a simplified form. Its mathematical derivation is detailed in Appendix A.

[9]Labor force classification guidelines referenced by enumerators during UNU-WIDER Vietnam SME data collection: "unpaid labour refers to those who do not receive wages or other remuneration directly related to the work they perform...Full-time is considered as a person working more than than 183 days per year, more than 20 days a month and more than 20 hours per week. Part-time is a person working under 20 hours per week and/or between 5 and 20 days a month. Casual labour [is] the residual that is working on average less than 5 days a month and/or few hours some weeks a month" (Institute of Labour Science and Social Affairs & the University of Copenhagen, 2009, pg. 21).

[10]As further reassurance, the McCrary (2008) density tests performed in Section 6.1 find no evidence of post-policy manipulation in firms' number of FP full-time workers at the 10-FP employee threshold, suggesting that firms report their employment levels and related data accurately.

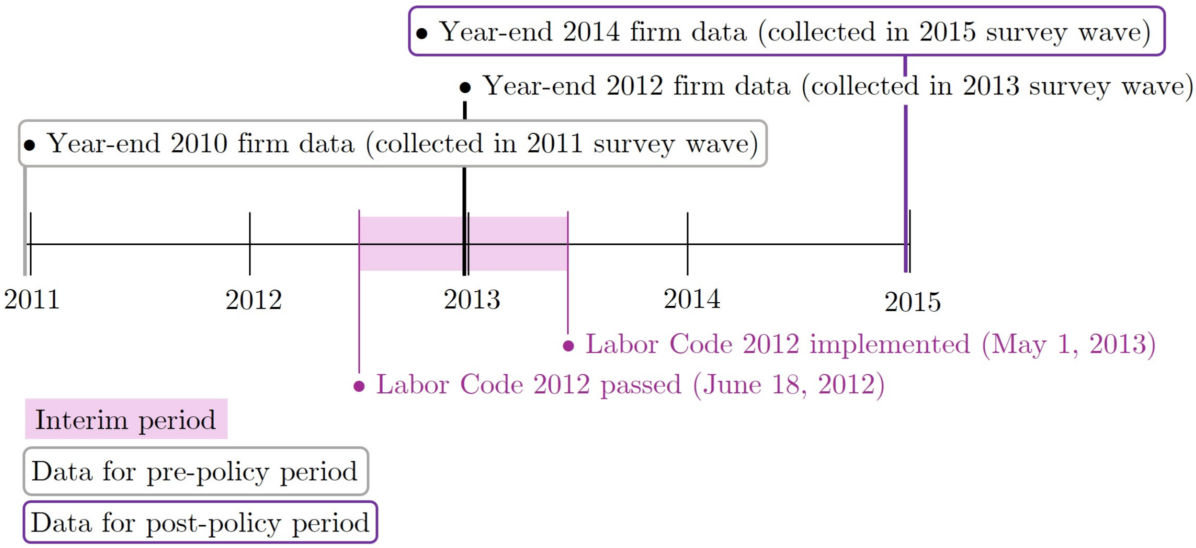


Figure 1: Policy & Data Collection Timelines

for each firm in the analytical sample.[11] I further restrict the analytical sample to firms with fewer than 50 total workers, consistent with the World Bank's definition of micro- and small-enterprises.[12] Year-end averages for each of the three worker types are plotted in Figure 2.

Figure 2 displays the average number of full-time workers by formal contract and pay status from year-end 2008 to 2014 for this restricted sample. The average number of FP full-time workers remains stable from year-end 2008 (i.e., 3.0) to 2010 (i.e., 3.1), declines modestly from 3.1 in year-end 2010 to 2.7 in year-end 2012, and then levels off again through year-end 2014 (i.e., 2.6). There is no statistical difference in means between year-end 2008 and 2010 (i.e., $p = 0.9578$) or between year-end 2012 and 2014 (i.e., $p = 0.5442$). However, the decline from year-end 2010 to 2012 *approaches* statistical significance (i.e., $p = 0.1301$) whereas that from year-end 2010 to 2014 is statistically significant (i.e., $p = 0.0322$). Comparatively, the average number of informal, paid full-time workers shows a more pronounced and steady decrease over time. Unpaid labor remains relatively constant, averaging around 1.3 to 1.4 workers throughout. These trends suggest that selecting year-end 2010 as the pre-policy period is prudent, as it precedes any potential behavioral adjustments firms may have made in anticipation of the 2012 employee threshold policy's implementation.

To help contextualize the change in magnitude of the financial penalty tied to the 2012 employee threshold policy, Table 1 presents the fine amount as a percentage of average year-end sales revenue, gross profit, and labor costs. The table is only representative of firms in the restricted sample across the pre-policy, interim, and post-policy periods. For the pre-policy (i.e., year-end 2004 to 2010) and interim (i.e., year-end 2012) periods, the calculation uses the original 1 million VND fine specified in Labor Code 1994, as the revised penalty denoted in Labor Code 2012 had not yet taken effect. For

[11] Appendix E details the rounding procedures used to derive the analytical sample from the Vietnam SME dataset.
[12] See https://ieg.worldbankgroup.org/sites/default/files/Data/Evaluation/files/SME_Synthesis.pdf for details.

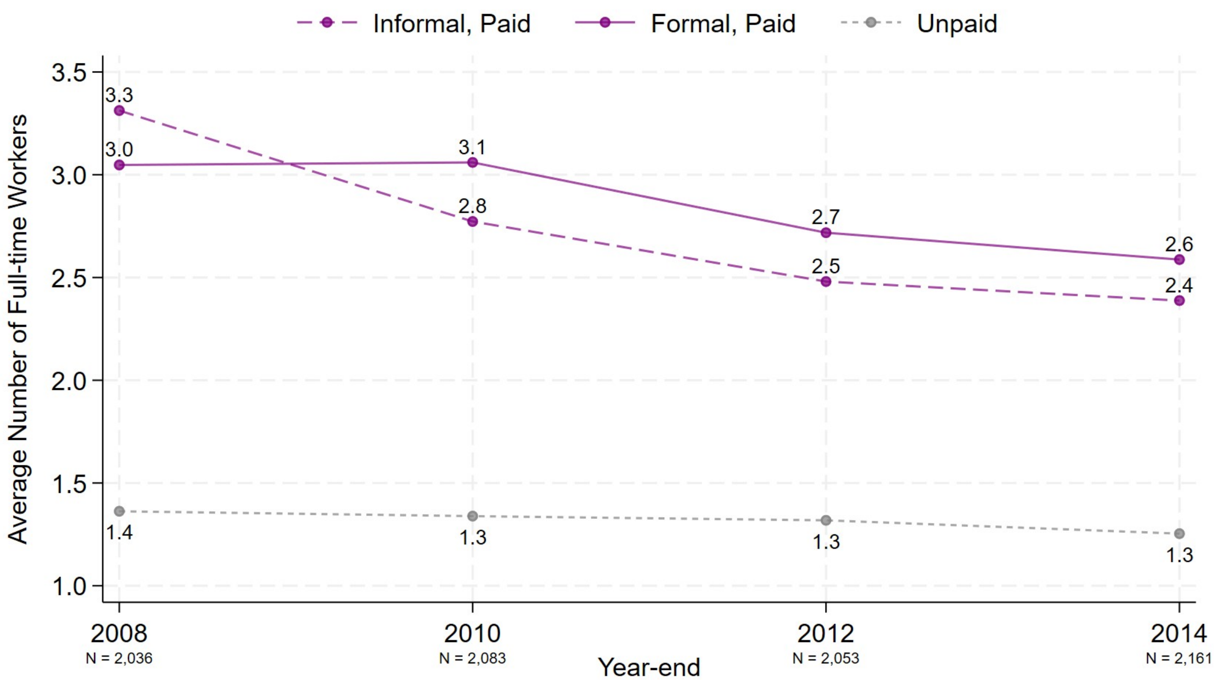


Figure 2: Mean Number of Worker Types for Restricted Sample by Year-End Period

Notes: This figure shows the average number of (1) informal, paid, (2) formal, paid, and (3) unpaid full-time workers across multiple year-end periods for the analytical sample. The sample used in this figure is also restricted to firms with fewer than 50 total workers, consistent with the World Bank's definition of micro- and small-enterprises. Each year-end period's sample size is indicated below the corresponding year.

the post-policy period (i.e., year-end 2014), the percentages instead reflect the increased penalty of 5 to 10 million VND established under Labor Code 2012, thus the range of values shown in Table 1. The findings underscore the significant rise in the economic burden of non-compliance, illustrating how the 2012 employee threshold policy raised the costs of informality for firms.

Table 1: Financial Penalty as a Percent of Firm Economic Account Metrics Over Time

	Pre-Policy Period			Interim Period	Post-Policy Period
Percent of Average Annual	2006	2008	2010	2012	2014
Sales Revenue	0.07%	0.03%	0.04%	0.04%	0.13–0.27%
Gross Profit	0.43%	0.22%	0.28%	0.27%	1.02–2.05%
Labor Costs	0.90%	0.31%	0.43%	0.33%	1.53–3.06%
Observations	2,420	2,434	2,318	2,371	2,452

Notes: Data reflect year-end values (i.e., December 31st). Since Labor Code 2012 was not implemented until May 2013, the financial penalty associated with Labor Code 1994 (i.e., 1 million VND) was applied to compute percentages from year-end 2004 to 2012. Year-end 2012 serves as the interim period, marking the gap between the passage and implementation of Labor Code 2012 (see Figure 1). Samples only include firms with less than 50 total workers, which is consistent with the World Bank's definition of micro- and small-enterprises. Gross profit is the difference between sales revenue and total costs—including raw materials, indirect expenses, and labor costs (e.g., wages, allowances, recruitment, training, as well as health, social, and unemployment insurance).

Figure 3 complements these findings by documenting a significant rise ($p < 0.001$) in the average number of total and policy inspections per firm between year-end 2010 and 2014 for the restricted sample. Total inspections rose from 0.15 to 0.73 per firm while policy inspections—those targeting firm compliance with labor and tax laws—increased from 0.07 to 0.33.[13] Table 1 indicates that the higher financial penalty greatly increased the cost of non-compliance while Figure 3 shows that this cost was actively enforced rather than merely theoretical. The concurrent escalation in policy inspections make the threat of detection and sanction both salient and credible. Altogether, these shifts created a dual mechanism of deterrence: firms faced higher expected costs of informality from the increased magnitude of the financial penalty *and* from the heightened enforcement risk. This institutional environment made the 2012 employee threshold policy more binding for firms.

4 Firm Profit Maximization Problem

The following firm profit maximization model builds on those with heterogeneous productivity and size-dependent regulatory costs—as developed by Garicano et al. (2016), Dabla-Norris et al. (2018), and Ulyssea (2018)—to capture firm responses to Vietnam's 2012 employee threshold policy. Following Garicano et al. (2016) and Dabla-Norris et al. (2018), the model abstracts from endogenous firm entry decisions and focuses on firm-level choices conditional on market participation.

[13] Policy inspections include announced and unannounced government visits to verify labor and tax law compliance. Non-policy inspections include technical compliance (i.e., workplace standards) or investigations following accidents.

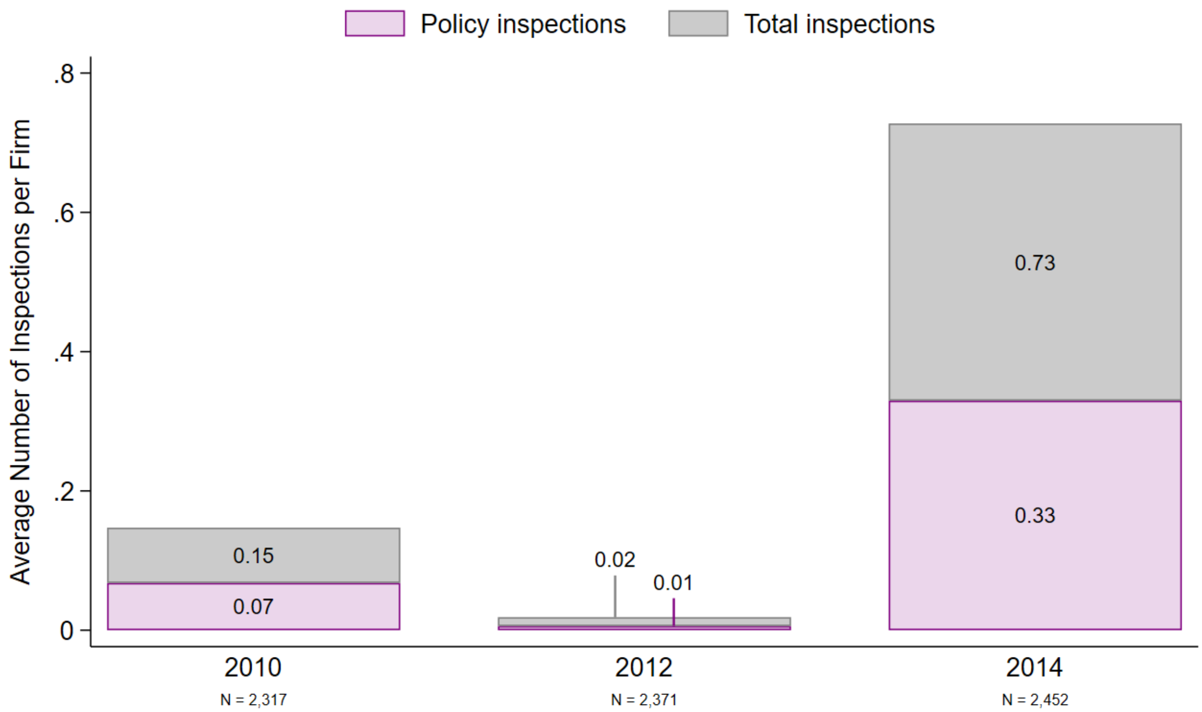


Figure 3: Mean Number of Inspections per Firm by Year-End Period

Notes: The figure shows the average number of total (gray) and policy (purple) inspections per firm across multiple year-end periods. Samples only include firms with fewer than 50 total workers, which is consistent with the World Bank's definition of micro- and small-enterprises. Each period's sample size is indicated below the corresponding bar.

After the implementation of the 2012 employee threshold policy in period t, a representative firm i operating in a homogeneous goods sector maximizes profits in period $t+1$ by choosing both its (1) labor composition and (2) formality status. The simultaneity of these decisions reflects how firm-level adjustments to (1) and (2) are often jointly constrained by compliance costs and market considerations introduced by the policy. Specifically, firm i in period $t+1$ chooses:

- $L_{FP,i,t+1}$: number of formal, paid (FP) workers[14]
- $L_{IP,i,t+1}$: number of informal, paid workers
- $L_{U,i,t+1}$: number of unpaid workers
- $F_{i,t+1} \in \{0, 1, 2\}$: categorical variable denoting firm i's formality status as either fully informal (0), partially formal (1), or fully formal (2)

Firm i's total labor size in the post-policy period is defined as:

$$L_{i,t+1} = L_{FP,i,t+1} + L_{IP,i,t+1} + L_{U,i,t+1}$$

[14]Consistent with the 2012 employee threshold policy's legal definition of "employee," L_{FP} only includes **formal, paid** workers, excluding unpaid and non-contract (i.e., informal) labor.

which is a mechanical outcome of firm i's chosen labor composition.

Firm i's formality status $F_{i,t+1}$ is determined by the summation of two binary, post-policy choices:[15]

- $R_{i,t+1} \in \{0,1\}$: whether firm i is registered with the government (1) or not (0)
- $K_{i,t+1} \in \{0,1\}$: whether firm i is fully compliant with labor regulations in terms of providing formal, written labor contracts to all workers (1) or not (0)

Table 2 summarizes how these firm choices map to each formality status. For instance, $F_{i,t+1} = 0$ if firm i neither registers with the government nor provides formal, written labor contracts to all of its workers, $F_{i,t+1} = 2$ if firm i does both, and $F_{i,t+1} = 1$ if firm i complies with only one of the two (i.e., government registration or full labor contract compliance but not both).

Table 2: Firm Choice Set for Each Formality Status Post-Policy

	Registered with the Government $R_{i,t+1}$	Formal Contracts for All Workers $K_{i,t+1}$	Formality Status $R_{i,t+1} + K_{i,t+1} = F_{i,t+1}$
Fully Informal	0	0	0
Partially Formal			
Formal but Operate Informally	1	0	1
Informal but Operate Formally	0	1	
Fully Formal	1	1	2

Firms differ in their productivity, denoted by $A_i > 0$, which captures firm-specific efficiency or technology. Following Garicano et al. (2016) and Dabla-Norris et al. (2018), firm output is modeled via a concave Cobb-Douglas production function with diminishing returns to effective labor input:

$$y_{i,t+1} = A_i (L^{eff}_{i,t+1})^{\alpha} \tag{2}$$

where $\alpha \in (0,1)$ is the output elasticity of effective labor—capturing returns to scale with respect to labor input—and effective labor input is a weighted sum of different worker types' productivity:

$$L^{eff}_{i,t+1} = \gamma_{FP} L_{FP,i,t+1} + \gamma_{IP} L_{IP,i,t+1} + \gamma_U L_{U,i,t+1} \tag{3}$$

where $\gamma_{FP} > \gamma_{IP} > \gamma_U > 0$.[16] Without loss of generality, I normalize γ_{FP} at 1 and define $\gamma_{IP} = \theta$ and $\gamma_U = \delta$, where $0 < \delta < \theta < 1$.

Total cost for firm i in period $t+1$ consists of fixed and variable costs associated with government registration, full labor contract compliance, and threshold-triggered regulatory burdens:[17]

[15] The distinction between government registration (i.e., $R_{i,t+1}$) and full labor contract compliance (i.e., $K_{i,t+1}$) mirrors the dual margins of formality discussed in Ulyssea (2018).

[16] This labor productivity hierarchy is supported by theoretical models (Lewis et al., 1954; Amaral & Quintin, 2006; Ulyssea, 2018) and empirical evidence from developing countries (La Porta & Shleifer, 2008; McCaig & Pavcnik, 2013).

[17] Productivity weights and cost parameters are assumed to be constant across firms and time, except through A_i.

$$
\begin{aligned}
C_{i,t+1} = {} & C_R R_{i,t+1} + C_K K_{i,t+1} + C_{FP} L_{FP,i,t+1} + C_{IP} L_{IP,i,t+1} + \psi(L_{U,i,t+1}) \\
& + C_T \cdot \mathbb{1}\{L_{FP,i,t+1} \geq 10 \text{ and } S_{i,t+1} = 1\} \\
& + \phi(R_{i,t+1}) P_{t+1} \cdot \mathbb{1}\{L_{FP,i,t+1} \geq 10 \text{ and } S_{i,t+1} = 0\}
\end{aligned}
\tag{4}
$$

The components of Equation 4 are as follows:

- C_R: fixed cost of government registration (e.g., administrative fees and taxes) incurred when $R_{i,t+1} = 1$.
- C_K: fixed cost of full labor contract compliance incurred when formal, written contracts are provided to all workers (i.e., $K_{i,t+1} = 1$). It includes costs related to contract administration, documentation, and preparation for labor inspections.
- $C_{FP} = w_{FP} + b_{FP}$: per-worker cost of FP labor, consisting of wages w_{FP} and mandatory benefits b_{FP} (e.g., unemployment, social, or health insurance contributions).
- $C_{IP} = w_{IP}$: per-worker cost of informal, paid labor that consists only of wages.
- $\psi(L_{U,i,t+1})$: an increasing and convex function (i.e., $\psi'(L_{U,i,t+1}) > 0$ and $\psi''(L_{U,i,t+1}) \geq 0$) representing the disutility, supervision burden, or opportunity cost of unpaid labor. L_U incurs no monetary cost (i.e., $w_U = 0$) but its use imposes non-monetary costs on firm i such as managerial effort, reduced operational efficiency, or limited availability of household labor.
- C_T: fixed cost of complying with the 2012 employee threshold policy (e.g., drafting, registering, and updating internal work regulations (IWRs)) incurred when firm i has 10 or more FP workers (i.e., $L_{FP,i,t+1} \geq 10$) and submits IWRs to the government (i.e., $S_{i,t+1} = 1$).
- $\phi(R_{i,t+1}) = \phi_L + (\phi_H - \phi_L)(R_{i,t+1})$: probability that firm i is inspected for non-compliance with the 2012 employee threshold policy, where $0 \leq \phi_L < \phi_H \leq 1$. Registered firms (i.e., $R_{i,t+1} = 1$) face a greater enforcement probability than unregistered firms (i.e., $R_{i,t+1} = 0$).
- P_{t+1}: expected financial penalty for failing to comply with the 2012 employee threshold policy incurred when firm i has 10 or more FP workers but does *not* submit IWRs to the government (i.e., $S_{i,t+1} = 0$). This post-policy monetary fine is five- to 10-times greater than its pre-policy counterpart (i.e., $P_t < P_{t+1}$) and it is scaled by the enforcement probability $\phi(R_{i,t+1}) \in [0, 1]$.

Consistent with the 2012 employee threshold policy, this model imposes two feasibility constraints. First, firms choosing $K_{i,t+1} = 1$ must not hire any informal, paid workers (i.e., $L_{IP,i,t+1} = 0$). This restriction captures that full labor contract compliance requires formal, written contracts for *all* workers and the exclusion of any informal labor. Second, when firms are at or above the 10-FP employee threshold and choose to submit IWRs, they must be registered with the government (i.e., $L_{FP,i,t+1} \geq 10 \wedge S_{i,t+1} = 1 \Rightarrow R_{i,t+1} = 1$). This operationalizes the 2012 employee threshold policy's requirement that IWRs must be submitted to provincial MOLISAs, a government process that presumes a legally recognized (i.e., registered) firm. However, this model does not impose a feasibility constraint that requires government registration (i.e., $R_{i,t+1} = 1$) to hire FP labor (i.e., $L_{FP,i,t+1} > 0$), thereby allowing unregistered firms (i.e., $R_{i,t+1} = 0$) to hire such labor. This reflects Vietnam's regulatory environment in which *all* firms—regardless of government registration—must comply with the 2012 employee threshold policy once they employ 10 or more FP workers.[18]

These feasibility constraints apply to two distinct forms of labor compliance in the model: $K_{i,t+1}$ and $S_{i,t+1}$. $K_{i,t+1}$ indicates full labor contract compliance (i.e., providing formal, written contracts to all workers) while $S_{i,t+1}$ indicates compliance with the 2012 employee threshold policy, triggered when a firm has 10 or more FP workers and submits IWRs. Though both are associated with labor regulation, each form imposes different costs and is subject to different enforcement mechanisms.

Firm i chooses its labor composition and formality status in period $t+1$ by selecting choice sets $\{\ell_{i,t+1}, f_{i,t+1}\}$ where $\ell_{i,t+1} = \{L_{FP,i,t+1}, L_{IP,i,t+1}, L_{U,i,t+1}\}$ and $f_{i,t+1} = \{R_{i,t+1}, K_{i,t+1}\}$, respectively, to maximize profits:[19]

$$\max_{\ell_{i,t+1}, f_{i,t+1}} A_i \cdot (\gamma_{FP} L_{FP,i,t+1} + \gamma_{IP} L_{IP,i,t+1} + \gamma_U L_{U,i,t+1})^\alpha - C_{i,t+1}(\ell_{i,t+1}, f_{i,t+1}) \tag{5}$$

subject to:

$$\gamma_{FP} > \gamma_{IP} > \gamma_U > 0, \quad C_{FP} > C_{IP} > 0, \quad w_U = 0, \quad \alpha \in (0,1),$$
$$A_i > 0, \quad \psi'(L_{U,i,t+1}) > 0, \quad \psi''(L_{U,i,t+1}) \geq 0$$

This formulation captures the post-policy trade-offs firms face between productivity gains from FP employment and the costs induced by government registration, full labor contract compliance, and crossing the regulatory threshold. The model yields three key predictions.

Theorem 1. *Let firm i choose $L_{FP,i,t+1}$ to maximize profits in the post-policy period, holding all other labor inputs and formality decisions fixed. If $L_{FP,i,t+1} \geq 10$, firm i incurs a discrete increase in fixed costs: a compliance cost $C_T > 0$ when $S_{i,t+1} = 1$ or an expected penalty for non-compliance $\phi(R_{i,t+1})P_{t+1} \geq 0$ when $S_{i,t+1} = 0$. Firms whose marginal benefit from increasing FP labor at the threshold is insufficient to cover this threshold-induced cost optimally choose $L_{FP,i,t+1} < 10$. If this cost is binding, this generates bunching at $L_{FP,i,t+1} = 9$ in the firm-size distribution.*

Proof. See Appendix B.1.

Theorem 2. *Suppose firm i employs fewer than 10 FP workers pre-policy (i.e., $L_{FP,i,t} < 10$). To reiterate, when $L_{FP,i,t+1} \geq 10$, firm i incurs a fixed compliance cost (i.e., C_T if $S_{i,t+1} = 1$) or an expected penalty for non-compliance (i.e., $\phi(R_{i,t+1})P_{t+1}$ if $S_{i,t+1} = 0$). To avoid these costs and*

[18] See also Ulyssea (2018) for comparable modeling of firm partial formality without registration constraints.

[19] $S_{i,t+1}$ indicates compliance with the 2012 employee threshold policy and is only defined for firms with $L_{FP,i,t+1} \geq 10$. Only these firms face a regulatory choice of whether to submit IWRs (i.e., $S_{i,t+1} = 1$) or not (i.e., $S_{i,t+1} = 0$). As such, $S_{i,t+1}$ is a *conditional* choice variable and not included in firm i's choice sets $\{\ell_{i,t+1}, f_{i,t+1}\}$ to maximize profits as defined by Equation 5. Note that if $L_{FP,i,t+1} < 10$, then $S_{i,t+1}$ is undefined and does not enter Equation 5.

maintain their optimal effective labor input (and thus output), firms will choose to stay below the threshold by setting $L_{FP,i,t+1} < 10$ and substituting toward informal, paid labor (i.e., $L_{IP,i,t+1}$), unpaid labor (i.e., $L_{U,i,t+1}$), or both. Formally, let τ denote the marginal threshold cost (e.g., C_T or $\phi(R_{i,t+1})P_{t+1}$) induced by the policy. Then, in equilibrium:

$$\frac{\partial L_{FP,i,t+1}}{\partial \tau} < 0, \quad \frac{\partial L_{IP,i,t+1}}{\partial \tau} \geq 0, \quad \frac{\partial L_{U,i,t+1}}{\partial \tau} \geq 0$$

and firm i's labor composition shifts away from regulated employment toward one or both unregulated labor arrangements as τ increases.

Proof. See Appendix B.2.

Theorem 3. *Suppose firm i employs at least 10 FP workers pre-policy (i.e., $L_{FP,i,t} \geq 10$). In the post-policly period, the expected penalty for non-compliance at the threshold increases from P_t to P_{t+1} where $P_t < P_{t+1}$. If firm i does not submit IWRs (i.e., $S_{i,t+1} = 0$), and thus not registered with the government (i.e., $R_{i,t+1} = 0$), it incurs an expected penalty for non-compliance of $\phi_L P_{t+1}$. If instead it submits IWRs (i.e., $S_{i,t+1} = 1$), which requires government registration (i.e., $R_{i,t+1} = 1$), it incurs the fixed cost $C_R + C_T$. Therefore, whenever $\phi_L P_{t+1} > C_R + C_T$, firms with $L_{FP,i,t+1} \geq 10$ minimize costs by submitting IWRs and thus registering with the government.*

To characterize firm i's post-policy labor composition, the effective labor input and cost functions in Equations 3 and 4, respectively, imply the following (marginal) per-effective-labor unit costs for paid (unpaid) labor:

$$c_{FP} \equiv \frac{C_{FP}}{\gamma_{FP}} = C_{FP}, \qquad c_{IP} \equiv \frac{C_{IP}}{\gamma_{IP}} = \frac{w_{IP}}{\theta}, \quad c_U(L_{U,i,t+1}) \equiv \frac{\psi'(L_{U,i,t+1})}{\delta}$$

Since full labor contract compliance entails fixed cost $C_K > 0$, if $c_{IP} \leq c_{FP}$, firm i strictly prefers not to provide formal, written contracts to all its workers in period t+1 (i.e., $K_{i,t+1} = 0$). Consequently, firms just above the 10-FP employee threshold pre-policy will be registered with the government (and thus IWR-compliant) but not labor contract compliant due to using informal labor arrangements post-policy. These firms are "formal but operate informally" with $(R_{i,t+1}, K_{i,t+1}) = (1, 0)$ (see Table 2). This outcome does not require employing informal, paid labor as informality may also arise from hiring unpaid workers without formal, written contracts. However, if additionally $c_{IP} \leq c_U(L^*_{U,i,t+1})$ and either $c_{IP} < c_{FP}$ or $c_{IP} < c_U(L^*_{U,i,t+1})$, then $L_{IP,i,t+1} > 0$.

Proof. See Appendix B.3.

5 Empirical Strategy

This section presents the empirical strategy used to estimate the causal effects of the increased financial penalty introduced by the 2012 employee threshold policy on firm outcomes. Section 5.1 describes the difference-in-discontinuities (DiDisc) identification framework, Section 5.2 defines the outcome variables, and Section 5.3 outlines the procedure for selecting optimal bandwidths around the 10-formal, paid (FP) employee threshold for DiDisc estimation.

5.1 Difference-in-Discontinuities Identification Framework

The empirical strategy employs a DiDisc approach which combines the localized identification of regression discontinuity design (RDD) with the intertemporal variation exploited by difference-in-differences (DiD) (Lalive, 2008; Grembi et al., 2016; Bennedsen et al., 2022; Ferguson & Kim, 2023). While increasingly used in applied research, DiDisc remains only partially formalized in the economic literature, with relatively few empirical applications to date (Takahashi, 2024). The 2012 employee threshold policy increased the financial penalty for non-compliance at a sharply defined firm-size threshold, creating a cross-sectional discontinuity and a time-based shift. These conditions are ideally suited for a DiDisc approach because it identifies the causal effect of the policy change on firm behavior at the margin before *relative to* after its implementation. Given this, a conventional RDD is insufficient because the policy generates temporal—not purely cross-sectional—identifying variation at the threshold. Appendix D provides diagnostic evidence that a standard DiD approach is also ill-suited and motivates the use of the DiDisc specification defined in Equation 6.

$$
\begin{aligned}
Y_{it} = \alpha &+ \beta_G \, Above10_i + \beta_T \, Post_t + \delta_0 \, (Above10_i \times Post_t) \\
&+ \theta_X \, FPEmp_i + \theta_G \, (Above10_i \times FPEmp_i) + \theta_T \, (Post_t \times FPEmp_i) \\
&+ \delta_1 \, (Above10_i \times Post_t \times FPEmp_i) + \mu_i + \epsilon_{it}
\end{aligned} \tag{6}
$$

In Equation 6, Y_{it} is an outcome for firm i in time t. The running variable $FPEmp_i$ measures the number of FP full-time workers employed by firm i, centered at the 2012 employee threshold policy cutoff of 10 using pre-policy values (i.e., at time $t-1$). $Above10_i$ equals 1 if firm i has 10 or more FP full-time workers pre-policy, defining the treatment group. Firms with less than 10 FP full-time workers form the control group. $Post_t$ equals 1 for the post-policy period (i.e., year-end 2014) and 0 for the pre-policy period (i.e., year-end 2010), capturing temporal variation. Firm fixed effects (μ_i) control for unobserved, time-invariant firm characteristics and ϵ_{it} is the error term.

Equation 6's structure parallels a standard DiD design in its first line: $Above10_i$ captures the pre-policy group difference, $Post_t$ captures the time difference across firms, and their interaction—represented by coefficient δ_0—measures the DiDisc treatment effect. This is the primary coefficient of interest, representing the change in discontinuity at the threshold following the policy. It isolates the causal effect of the 2012 employee threshold policy on outcomes for firms *just* above the threshold (i.e., the treatment group) compared to firms *just* below it (i.e., the control group) post-policy. By comparing the size of the discontinuity at the cutoff before and after the policy, the DiDisc method nets out any pre-existing jump and recovers the policy-induced shift captured by δ_0.

The second line introduces the RDD components by including the running variable $FPEmp_i$ and its interactions with both group and time. It permits different slopes on either side of the cutoff and allows those slopes to change over time. The final line integrates the RDD and DiD components by including the triple interaction ($Above10_i \times Post_t \times FPEmp_i$) with coefficient δ_1. This captures post-policy changes in the slope of the running variable at the cutoff. Though informative, δ_1 is secondary to δ_0 for interpreting the causal effect of the 2012 employee threshold policy.

5.2 Outcome Variables

The primary outcomes analyzed using Equation 6 evaluate how the increased financial penalty associated with the 2012 employee threshold policy affects firm size and labor composition. Though the threshold-based labor policy targets firms with 10 or more FP workers, data limitations (see Section 3) restrict identification to FP *full-time* workers. Since this variable is the running variable in the causal identification strategy, it cannot be an outcome. Instead, Section 6.1 formally tests whether firms endogenously adjusted their FP full-time employment post-policy. Firm size is thus measured more broadly as the total number of workers—defined as the sum of full-time, part-time, and casual workers—to capture whether firms expanded or contracted their overall labor force post-policy, regardless of worker type. Firms that are close to the threshold may face strong incentives to manage their labor force size proactively to avoid triggering formalization requirements. This can include strategically scaling back hiring or reducing headcount across any worker type to maintain distance from the cutoff. Hence, the number of total workers is an observable indicator of behavioral responses to the policy, even if not all worker types contribute directly to the regulatory threshold.

Primary outcomes also include measures of labor composition to evaluate how firms may strategically restructure their labor force to minimize regulatory exposure while maintaining operational capacity and productivity. I assess four worker types: part-time workers, casual workers, informal, paid full-time workers, and unpaid full-time workers. The latter two worker types fall outside the legal definition of "employee" according to the 2012 employee threshold policy. Therefore they do not count toward the compliance threshold. Firms that anticipate higher regulatory costs may be incentivized to substitute away from *regulated* worker types toward these more flexible, lower-cost alternatives. For example, relying more heavily on informal or unpaid labor allows firms to expand capacity without formally increasing their size as defined by the government, thereby avoiding the administrative burdens and risk of financial penalty associated with formalization. Consistent with the prediction in Theorem 2, higher effective costs of regulated labor should reduce FP employment and prompt firms to substitute toward informal, paid or unpaid workers instead.

Secondary outcomes capture potential spillover effects of the policy on firm performance, measured through sales revenue, profitability, productivity, and labor costs. They include sales revenue per full-time worker, gross profit per full-time worker, labor costs per full-time worker, labor productivity, and capital productivity.[20] While the primary hypothesis is that policy effects are con-

[20] All size-normalized firm economic account metrics represent year-end values. Gross profit is calculated as sales revenue minus total costs, which include raw materials, indirect expenses, and labor costs. Labor costs include wages, allowances, recruitment and training expenses, as well as health, social, and unemployment insurance contributions. Labor productivity is measured as value added per full-time worker, where value added equals sales revenue minus production costs (including raw materials and indirect expenses). Capital productivity is measured as valued added per asset (including both physical and financial assets).

centrated on firm size and labor force structure, binding growth constraints may engender broader consequences. Downsizing or restructuring to avoid formalizing could limit a firm's ability to scale operations, invest, innovate, or remain competitive. Declines in sales revenue or profit may indicate that regulatory avoidance comes at a cost to firm performance while reductions in labor or capital productivity could signal efficiency losses from substituting away from more productive formal labor. Lower labor costs may reflect cost-minimization or reduced labor force stability and quality.

If there are no significant changes in these secondary outcomes, it would indicate that the policy's distortions are largely confined to a firm's size and structure, with firms successfully adapting to the threshold while preserving their operational and financial performance. Conversely, significant adverse effects would suggest broader unintended consequences—such as constrained growth, efficiency losses, or reduced profitability—that extend beyond the policy's intended compliance objectives. Identifying whether these spillovers occur is important for understanding the full scope of how threshold-based labor policies affect firm behavior and outcomes in developing economies.

Lastly, outcomes include indicators for firm formality status: fully formal, formal but operating informally, informal but operating formally, and fully informal (see Table 2). Estimating Equation 6 on these outcomes helps identify the mechanisms through which firms near the threshold respond to the policy change. The 2012 employee threshold policy raises the cost of informality for firms above the threshold, but it does not necessarily lead to full formalization. This is because formality status is endogenously determined by government registration and labor contract compliance decisions. If the expected benefits of full formalization outweigh its costs, firms above the threshold will more likely fully formalize. Yet in other cases, consistent with the prediction in Theorem 3, firms above the threshold may choose partial formality. For example, firms may register with the government but avoid providing formal, written labor contracts to all workers to minimize costs while retaining operational legitimacy. This hybrid strategy would allow these firms to satisfy the more enforceable requirements while avoiding the costlier ones. Analyzing changes in firm formality status directly tests this mechanism, revealing if the policy encouraged meaningful formalization or incentivized compliance avoidance strategies that undermine its intent and effectiveness.

5.3 Optimal Bandwidth Selection

To reiterate, the running variable is a firm's number of FP full-time workers. Appendix Figure C.1 presents its distribution for the analytical sample pre- and post-policy.[21] Three features violate the continuity assumptions that underpin conventional RDD and, by extension, DiDisc estimation: (1) the distribution is highly left-skewed, (2) the running variable is discrete and takes on relatively few values, and (3) there is a pronounced mass point at zero where approximately 78 percent of firms cluster in both periods. This mass at zero suggests that many firms strategically avoid regulatory exposure *altogether* by not employing any workers who satisfy the government's legal definition of "employee." This persistent pattern indicates widespread evasion of compliance requirements and underscores the extent to which firms operate beyond the reach of size-dependent formalization

[21] Appendix E details the rounding procedures used to derive the analytical sample from the Vietnam SME dataset.

policies in settings characterized by high levels of informality. These features violate the smoothness and density continuity assumptions required for local polynomial estimation, the most commonly used implementation of RDD. They therefore motivate the use of the local randomization approach, which remains valid when the running variable is discrete and contains pronounced mass points.

The local randomization approach (Cattaneo et al., 2024) treats units within a selected window around the cutoff (i.e., 10 FP full-time workers) as if they were randomly assigned to treatment or control, circumventing the need for continuity or smoothness assumptions that are violated in this context.[22] Implementation involves conducting a series of covariate balance tests to determine the largest symmetric and asymmetric windows within which the "as-if" random assignment assumption is empirically plausible. This procedure tests for differences in the distribution of selected pre-policy covariates between treatment and control groups across possible candidate windows. The goal is to identify the widest window in which no statistically significant differences arise by using a p-value threshold of at least 0.15, following the recommendation of Cattaneo et al. (2024). Optimal window selection specifically follows a greedy expansion rule: starting from the smallest feasible window, each side is widened step by step until balance fails for at least one covariate. The optimal window is therefore defined as the largest passing window immediately before failure.

Figure 4 shows the running variable's distribution for the pre-policy period considering the optimal symmetric (h, h) and asymmetric (h_-, h_+) windows selected using the procedure in Cattaneo et al. (2024). In the local randomization approach, the bandwidths within these optimal windows define the treatment and control groups. The covariate balance tests are applied to the pre-policy covariates listed in Table 3 and support an optimal symmetric window of $[7, 12]$ with a bandwidth (h) of 3. The right-side bandwidth always includes the cutoff value since the 2012 employee threshold policy applies to firms with *at least* 10 FP workers. This optimal symmetric window yields a binomial test p-value of 0.314 and a minimum covariate balance test p-value of 0.322, indicating no significant difference in treatment assignment and well-balanced pre-policy characteristics.[23] Figure 5 plots the minimum covariate balance test p-values for all symmetric candidate windows around the cutoff, with the dotted horizontal line denoting the 0.15 threshold. The figure shows the widest symmetric window that satisfies covariate balance is $[7, 12]$ while the next widest, $[6, 13]$, yields a minimum p-value below 0.15 and thus fails the covariate balance test. Building on this, I adapt the greedy expansion procedure from Cattaneo et al. (2024) to obtain the optimal *asymmetric* window. Starting from the optimal symmetric window of $[7, 12]$, I iteratively expand one bandwidth at a time—first to the right, then to the left—testing covariate balance after each increment. The procedure alternates between sides until neither bandwidth can be further increased without violating covariate balance. Table 3 reports the corresponding results, which support an optimal asymmetric window of $[1, 25]$ with a left-side bandwidth (h_-) of 9 and a right-side bandwidth (h_+) of 15.

Using these optimal windows, the analytical sample narrows to firms whose pre-policy FP full-time employment falls within those bandwidths. The optimal symmetric window of $[7, 12]$ defines

[22] A "window" refers to the full range of values around the cutoff spanning both sides while a "bandwidth" refers to the distance from the cutoff on either side.

[23] Results were obtained using the `rdwinselect` Stata command as recommended by Cattaneo et al. (2024).

a subsample of 63 firms—27 in the treatment group and 36 in the control—whereas the optimal asymmetric window of $[1, 25]$ results in 187 firms—107 treatment and 80 control. Both bandwidth-defined subsamples constitute the updated estimation samples used in the subsequent analyses.

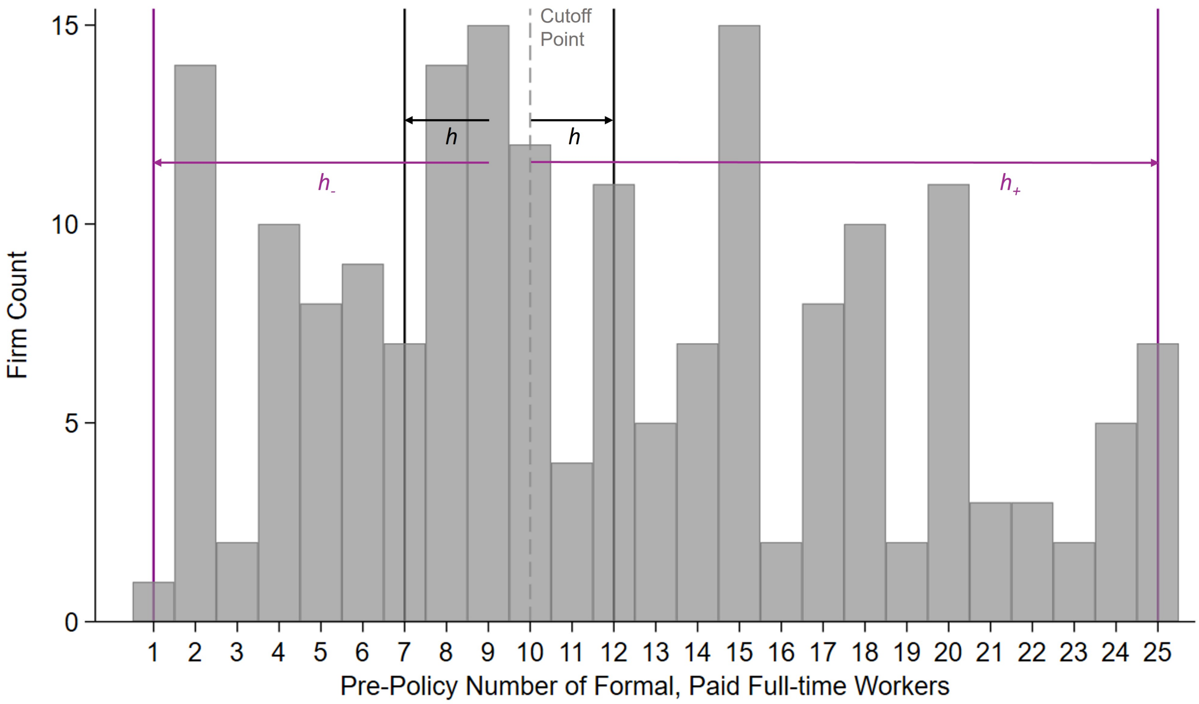


Figure 4: Distribution of Running Variable Pre-Policy within the Optimal Windows

Notes: This figure shows the distribution of the running variable for the pre-policy period within the optimal windows. The optimal symmetric window is [7,12] and has a symmetric bandwidth (h) of 3. The optimal asymmetric window is [1,25] and has an asymmetric bandwidth to the left of the cutoff (h_{-}) of 9 and to the right of the cutoff (h_{+}) of 15. These optimal windows were identified using the local randomization approach for a discrete running variable, following the greedy expansion rule described in Cattaneo et al. (2024) whereby each side of the window is widened until covariate balance fails and the selected window is the last that passes.

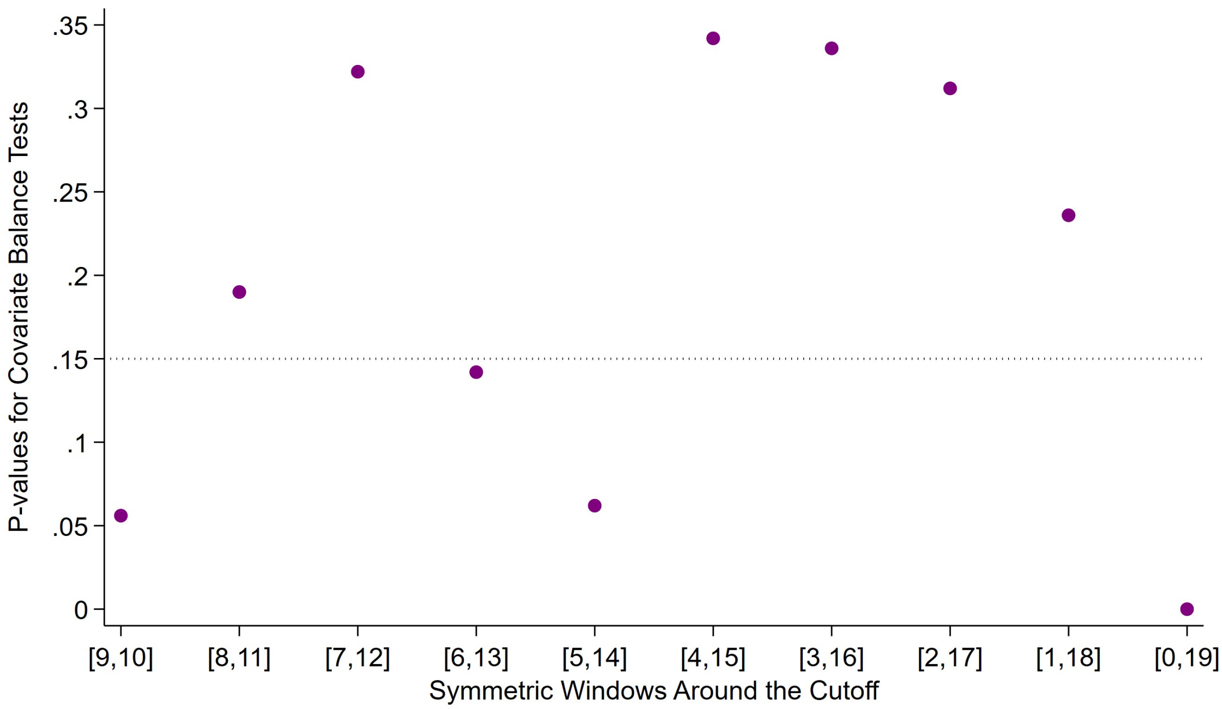


Figure 5: Minimum P-values for Covariate Balance Tests across Symmetric Candidate Windows

Notes: This figure shows the minimum p-value for covariate balance tests for each candidate symmetric window around the cutoff, expressed in terms of the number of pre-policy formal, paid full-time workers. Each dot corresponds to the lowest p-value across the covariates listed in Table 3. The dotted horizontal line marks the 0.15 threshold used to evaluate covariate balance. Windows whose dots are above this line pass the covariate balance test and are admissible; those with dots below the line fail. Following the procedure of Cattaneo et al. (2024), the optimal symmetric window is $[7, 12]$ as it is the last window to pass the covariate balance test before the next one fails.

Table 3: Pre-Policy Covariate Balance Within the Optimal Asymmetric Window of [1,25]

Covariate	Means		Fisherian P-value	Description	Rationale
	Control	Treatment			
Owner/Manager Characteristics					
Male	0.58	0.51	0.546	Gender	Gender-linked differences
Upper secondary school completion	0.80	0.89	0.160	Highest level of education completed is upper secondary school	Managerial ability
Firm Characteristics					
Province	43.68	47.24	0.438	Firm's operating province in Vietnam	Geographic location
Road access	0.85	0.85	1.000	A main, paved road leads to the firm	Infrastructure access
Age	10.23	9.57	0.690	Age of firm in years	Firm operational longevity
Value added per full-time worker	86,681	104,830	0.360	Year-end 2010 value added divided by the number of full-time workers	Firm productivity
Assets per full-time worker	732,109	571,051	0.316	Year-end 2010 physical and financial assets divided by the number of full-time workers	Firm wealth and resources
Sales revenue per full-time worker	298,647	403,645	0.262	Year-end 2010 sales revenue divided by the number of full-time workers	Firm performance
Gross profit per full-time worker	55,148	61,613	0.790	Year-end 2010 gross profit divided by the number of full-time workers	Firm profitability
Observations	80	107			

Notes: Firms with less than 10 formal, paid (FP) full-time workers pre-policy form the control group; those with 10 or more constitute the treatment group. In Vietnam, completion of upper secondary school is equivalent to a U.S. high school diploma and is *optional* after lower secondary school. Value added is calculated as sales revenue minus production costs (including raw materials and indirect expenses) while gross profit equals value added minus labor costs (e.g., wages, allowances, recruitment, training, as well as health, social, and unemployment insurance). All size-normalized firm economic account metrics are reported in thousands of Vietnamese Dong (VND). Covariate balance tests were conducted using the `rdrandinf` Stata command as recommended by Cattaneo et al. (2024). The finite sample two-sided Fisherian p-values test if the distribution of each covariate differs statistically across the cutoff within the optimal asymmetric window of 1 to 25 FP full-time workers, based on bandwidths of 9 units to the left and 15 units to the right of the cutoff. The right-side bandwidth includes the cutoff value because the 2012 employee threshold policy applies to firms with **10 or more** FP workers.

6 Results

The model presented in Section 4 predicts how threshold-based regulations distort firms' labor and formality decisions. Building on this framework, I empirically evaluate how the 2012 employee threshold policy shapes firm behavior around the 10-formal, paid (FP) employee threshold. Section 6.1 tests for endogenous adjustments in firms' employment of FP full-time workers at this threshold while Section 6.2 estimates average effects on labor composition, performance, and formality status.

6.1 McCrary Density Tests

To evaluate if firms strategically manipulated their size distribution around the 10-FP employee threshold, I conduct the McCrary (2008) density test separately for the pre- and post-policy periods using the optimal symmetric and asymmetric bandwidths from Section 5.3. This approach tests for discontinuities in the distribution of the running variable (i.e., the number of FP full-time workers) at the cutoff in each period to determine whether manipulation emerged after the 2012 employee threshold policy's implementation. A statistically significant jump in density just below the 10-FP employee threshold post-policy would suggest that firms adjusted their FP employment downward to avoid policy compliance, consistent with the bunching behavior predicted by Theorem 1.

Figure 6 shows the distribution of firms by the number of FP full-time workers for both periods and optimal windows. Panel A presents results using the optimal symmetric window of $[7, 12]$ while Panel B shows those using the optimal asymmetric window of $[1, 25]$. Under the optimal symmetric window, the McCrary (2008) density test finds no evidence of a discontinuity at the cutoff in the pre-policy period: the estimated log-difference in density across the threshold is positive ($T = 1.60$) but not statistically significant ($p = 0.110$). The estimate remains insignificant ($T = -1.13$, $p = 0.257$) for the post-policy period. In contrast, using the wider asymmetric window, the pre-policy density test produces a negative and statistically significant discontinuity ($T = -2.05$, $p = 0.041$). This indicates that relatively fewer firms were at or just above the 10-FP employee threshold compared to just below it. The pattern is consistent with firms strategically clustering just below 10 FP full-time workers to circumvent the penalty under the policy's earlier version, when the non-compliance fine was only 1 million VND. However, post-policy results under the asymmetric window are again insignificant ($T = -1.04$, $p = 0.298$). Visual inspection of Panel B reinforces this interpretation: the pre-policy period exhibits a clear "hollowing" at 10 FP full-time workers with excess mass just below, whereas the post-policy period distribution appears smooth across the cutoff. This suggests that the sharper discontinuity evident pre-policy does not persist after the introduction of the 2012 employee threshold policy. The difference between panels is unsurprising: the optimal asymmetric window includes more observations (i.e., 187 vs. 63), increasing power to detect pre-policy bunching but *not* altering the null post-policy result seen under the optimal symmetric window. The pattern under the optimal asymmetric window—a significant negative density discontinuity pre-policy and no discontinuity post-policy—also holds when using the analytical sample[24].

[24]The analytical sample (see Appendix E for details) includes 228 treatment firms and 1,211 control firms. Applying

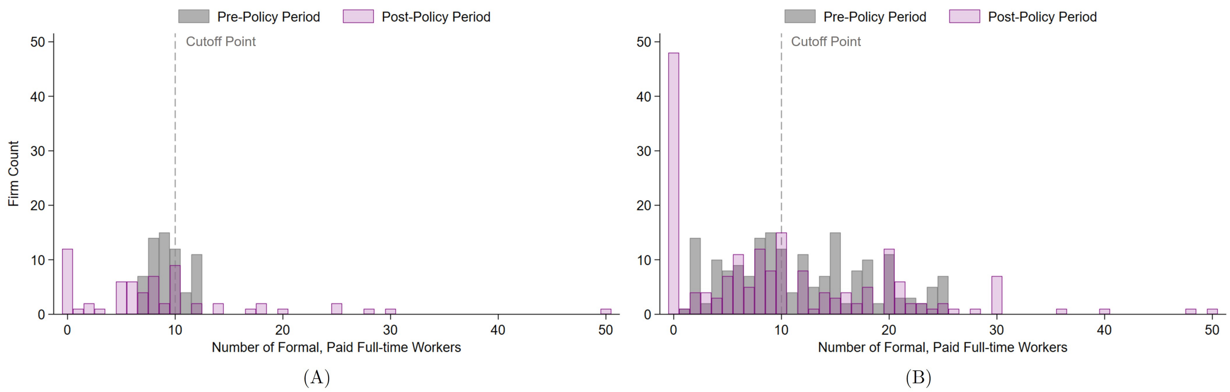


Figure 6: Distribution of Running Variable Pre- and Post-Policy within the Optimal Windows

Notes: This figure shows the distribution of firms around the 10-formal, paid (FP) employee threshold, with Panels A and B using the optimal symmetric and asymmetric windows, respectively. The gray bars depict the distribution of firms within these optimal windows in the pre-policy period while the purple bars depict the distribution of *the same firms* in the post-policy period. In Panel A, McCrary (2008) density tests find no evidence of a discontinuity at the cutoff in either period. Panel B shows a statistically significant pre-policy drop in the density of firms just above the threshold relative to just below—consistent with bunching—but the post-policy estimate is again insignificant.

The absence of a post-policy discontinuity suggests that the higher discrete compliance cost was *not* binding for most firms, consistent with the converse prediction of Theorem 1. Appendix B.1 infers that this absence of bunching can arise for several reasons: firms may be productive enough to absorb the cost shock without distortion, enforcement may be weak, the financial penalty may be perceived as low, or firms may adjust along alternative labor margins. In the Vietnamese setting, two mechanisms appear the most salient. First, stricter enforcement (see Figure 3) combined with a higher penalty for non-compliance may have shifted incentives such that firms crossed the threshold transparently rather than intentionally employing less than 10 FP full-time workers.[25] Second, firms may have substituted toward informal and unpaid workers—consistent with Theorem 2—allowing them to avoid the higher cost without reducing their overall size or operational capacity. Together, the results illustrate the two cases anticipated by Theorem 1: bunching occurs when the threshold binds while continuity prevails when it does not. Section 6.2 investigates these adjustment margins directly, distinguishing whether firms relied primarily on transparency or substitution in practice.[26]

the McCrary (2008) density test yields $T = -2.00$ and $p = 0.045$ pre-policy and $T = -0.59$ and $p = 0.559$ post-policy.

[25]The financial penalty for non-compliance under the 2012 employee threshold policy is 5 to 10 million VND—five to ten times higher than in the earlier iteration.

[26]The absence of post-policy manipulation also supports the data reliability of the Vietnam SME database. If firms had strategically underreported FP full-time workers during survey administration, we would expect excess mass just below the 10-FP employee threshold. Instead, the smooth distribution suggests accurate reporting (see Section 3).

6.2 Difference-in-Discontinuities

Table 4 presents estimates from Equation 6 using the symmetric and asymmetric optimal windows, only underscoring outcomes with statistically significant effects.[27] The results show that firms just *above* the 10-FP employee threshold responded to the higher costs of informality introduced by the 2012 employee threshold policy compared to firms just *below* it. They are consistent with the behavioral margins predicted by the firm profit maximization model in Section 4. Specifically, they support Theorems 2 and 3 which predict threshold-induced substitution toward unregulated labor and strategic and selective compliance with multi-dimensional formality requirements, respectively. These patterns are robust to multiple-hypothesis adjustments: controlling the false discovery rate within pre-specified outcome families using the Benjamini-Hochberg and the two-stage Benjamini-Krieger-Yekutieli procedures yields the same qualitative inferences (see Appendix Table C.6).[28]

Table 4: Impact of Threshold-Based Labor Policy on Selected Firm Outcomes

	Unpaid Full-time Workers		Formal Firm But Operates Informally		Gross Profit Per Full-time Worker		Labor Productivity	
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Post_t \times Above10_i$	-0.404	-0.352*	0.316**	0.182**	44,694	139,326**	23,263	123,164*
	(0.344)	(0.209)	(0.136)	(0.088)	(60,592)	(67,209)	(72,688)	(69,049)
Group mean	0.444	0.763	0.056	0.038	65,617	55,148	100,840	86,681
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.691	0.767	0.666	0.570	0.560	0.508	0.565	0.514
Optimal window	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]
Observations	126	374	126	374	126	374	126	374

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the firm-level."Group mean" reports the pre-policy mean of the outcome variable for the control group (i.e., firms just below the threshold) in each model specification. Results only underscore outcomes with statistically significant effects. Appendix Figure C.2 presents binned scatter plots with fitted linear regressions for each selected firm outcome, illustrating the change in discontinuity at the 10-FP employee threshold between the pre- and post-policy periods. P-values and FDR-adjusted q-values for these estimates are listed in Appendix Table C.6.

Regarding the primary outcomes, Column 2 shows that the number of unpaid full-time workers decreased by approximately 0.35 workers for firms just above the threshold compared to those just below it following the policy change. The effect is only significant for firms in the optimal asymmetric window. Interpreted through Theorem 2, this discontinuity indicates that firms just *below* the threshold substituted toward unpaid labor arrangements—such as relying on family members—to

[27] Results for all the outcomes delineated in Section 5.2 are reported in Appendix Tables C.2, C.3, and C.4.

[28] Appendix Table C.6 reports false discovery rate (FDR)-adjusted q-values computed within pre-specified outcome families using the Benjamini-Hochberg and the two-stage Benjamini-Krieger-Yekutieli procedures for outcomes whose estimates appear in Appendix Tables C.2, C.3, and C.4. Under the optimal symmetric window, the estimate for *formal firm but operates informally* has $p = 0.0235$ and $q \approx 0.0704$. Under the optimal asymmetric window, the estimate for *unpaid full-time workers* ($p = 0.0939$) is $q \approx 0.1878$; that for *formal firm but operates informally* ($p = 0.0397$) is $q \approx 0.1192$; that for *gross profit per full-time worker* ($p = 0.0395$) is $q \approx 0.0791$; and, that for *labor productivity* ($p = 0.0761$) is $q \approx 0.1522$. Thus the qualitative pattern in Table 4 largely holds post-adjustment.

offset the higher marginal cost of FP employment triggered at the threshold while still maintaining operational capacity. The negative coefficient for firms just above the threshold indicates that firms just below it increased their reliance on unpaid labor, consistent with the Theorem 2 prediction that $\frac{\partial L_{U,i,t+1}}{\partial \tau} \geq 0$ when threshold-induced costs increase. This finding supports the model's expectation that the 2012 employee threshold policy distorted labor composition. Descriptive evidence further reinforces this behavioral pattern: Appendix Figure C.1 shows that about 78 percent of firms in the analytical sample—which includes firms within and outside of the optimal windows—employ zero FP full-time workers across the pre- and post-policy periods. This mass point conveys widespread avoidance of the regulatory threshold, suggesting that rather than risk triggering compliance obligations most firms structure their labor force to exclude *any* workers that satisfy the legal definition of "employee." These results indicate that firms adapt to the policy's higher financial penalty and stricter enforcement (see Figure 3) by adjusting their labor margins to reduce regulatory exposure.

For secondary outcomes, Column 6 shows that firms just above the threshold experienced sizable gains in profitability relative to firms just below it in the post-policy period. However, these effects are significant only under the optimal asymmetric window. For that case, gross profit per full-time worker increased by 139 million VND (about 5,300 USD). These improvements suggest that despite higher compliance costs, firms just above the threshold were able to maintain or even enhance their profit levels. This could stem from scale effects, more efficient input allocation, or cost-saving labor substitutions in response to the policy. Column 8 also indicates an increase in labor productivity[29] of 123 million VND (about 4,700 USD) for firms just above the threshold post-policy. This effect, significant only under the optimal asymmetric window, is consistent with how shifting from unpaid to paid labor can improve operational efficiency. Though not directly targeted by the 2012 employee threshold policy, these outcomes suggest positive spillovers beyond its intended compliance margins.

Finally, results in Columns 3 and 4 speak directly to the mechanism of strategic non-compliance. The probability that a firm is "formal but operates informally" (i.e., registered with the government but does not provide all full-time workers with formal, written labor contracts) increased by 18 to 32 percentage points among firms just above the threshold relative to those just below it post-policy. Relative to a pre-policy mean of only 3.8 to 5.6 percent among firms just *below* the threshold, this represents a substantial behavioral shift. The finding provides direct empirical support for Theorem 3, which predicts that once firms cross the threshold they often adopt a partially formal status—satisfying visible requirements such as government registration while avoiding costlier or harder-to-monitor ones like universal labor contracts. The observed increase in partial firm formality reflects the strategic adjustment predicted by the model: firms balance legitimacy and cost minimization by selectively complying on the cheapest margin. In practice, this means relying on informal labor even after registering. Such behavior undermines the full intent of the 2012 employee threshold policy and underscores the broader challenge of enforcing multi-dimensional formalization mandates.

These findings suggest a nuanced but strategic firm response to this threshold-based labor policy:

[29]Labor productivity is measured as value added per full-time worker, where value added refers to the difference between sales revenue and production costs (including raw materials and indirect expenses).

firms internalize the costs of formalization and adapt along multiple margins—labor composition, profitability, efficiency, and compliance—by shifting toward the least costly forms of adjustment. The 2012 employee threshold policy does succeed in incentivizing government registration. However, it concurrently enables partial compliance. This pattern raises important concerns for policy design. Without stronger enforcement across *all* dimensions of formality, threshold-based regulations risk encouraging only nominal compliance among firms, falling short of substantive formalization.

Since the estimates in Table 4 reflect *average* effects across firms with varying levels of pre-policy compliance, they may mask important sources of underlying heterogeneity. Appendix F presents subgroup analyses by firms' pre-policy formality status. The results indicate that fully formal and partially formal firms respond in systematically distinct ways to the 2012 employee threshold policy.

7 Conclusion

This research presents causal evidence that size-dependent formalization policies can engender multi-dimensional behavioral responses among firms in developing economies. I exploit Article 119 of Vietnam's Labor Code 2012 which builds upon a pre-existing firm-size threshold by imposing a substantially higher financial penalty on firms who employ at least 10 formal, paid (FP) workers that fail to register internal work regulations. Empirical results show that firms respond along multiple margins rather than exhibiting canonical "bunching" behavior. To interpret these findings, I derive three predictions from a firm profit maximization model.

Consistent with the converse of Theorem 1 from the model, McCrary (2008) density tests reveal no evidence of bunching just below the 10-FP employee threshold, indicating that the discrete cost shock was not binding for most firms. Instead, as predicted by Theorems 2 and 3, firms adapted on alternative margins. Firms just below the threshold substituted toward unpaid full-time workers—like family members—to avoid triggering costly compliance obligations. In contrast, firms just above the threshold registered with the government but continued to rely on informal labor arrangements. These "formal but operating informally" firms adopted hybrid compliance strategies that reduced exposure while maintaining capacity. Firms just above the threshold also realized profit and labor productivity gains, implying that partial compliance can coexist with efficiency improvements.

Together, these findings suggest that threshold-based labor policies can reshape *how* firms comply rather than *whether* they comply. Increased government registration among firms just above the threshold demonstrates a visible response to enforcement. However, limited improvements in labor contract compliance and continued reliance on unregulated labor show that firms selectively satisfy the most visible regulatory obligations while circumventing costlier or less-easily-monitored ones. Such patterns reflect strategic non-compliance: deterrence through increased discrete penalties can encourage formalization along observable margins but does not ensure comprehensive compliance.

Methodologically, this paper advances the application of the difference-in-discontinuities design with local randomization to a context with a discrete and highly skewed running variable. Conceptually, it operationalizes a nuanced definition of firm formality by distinguishing between a firm's government registration status and compliance with formal labor practices. This distinction reveals gradations of firm informality that binary measures mask—showing how firms can restructure their labor force or employ hybrid compliance strategies to minimize regulatory exposure.

From a policy perspective, these results highlight that effective formalization strategies hinge on *how* firms perceive and navigate regulatory costs. When compliance thresholds trigger sharp cost jumps, firms may neither fully comply with the policy nor evade it outright, but instead reconfigure their operations to meet visible enforcement requirements while continuing informal practices elsewhere. This behavior underscores a broader challenge for developing economies: when firms comply for the sake of visibility rather than substance, formalization policies risk undermining institutional credibility and perpetuating the very informality they seek to eliminate. More broadly, these findings suggest that formalization policies should be viewed not only as enforcement instruments but as development tools that influence employment quality, productivity, and long-term upgrading.

While this study offers robust evidence for Vietnamese manufacturing firms, its external validity is bounded by data and contextual limitations. First, restricting the measurement of employment to FP full-time workers likely underestimates firms' *true* formal labor force, as part-time and casual workers with formal, written contracts are excluded. Because this measurement differs from the employee definition used in the policy—which counts all formally contracted, paid workers—measured firm size may be biased downward compared to the 10-FP employee threshold. Consequently, the results may understate the extent to which such firms actually face or respond to the policy. Second, Vietnam's enforcement environment has recently been characterized by higher financial penalties and stronger inspection capacity. Hence it may differ from settings with very weak or very strong institutions. Future research should assess how variation in enforcement intensity, monitoring credibility, and policy thresholds influences firms' behavior across sectors and countries, and if similar hybrid compliance equilibria emerge. Ultimately, understanding these mechanisms is essential for designing policies that expand the benefits of formality without prompting firm-level distortions.

References

Aghion, P., Bergeaud, A., & Van Reenen, J. (2023). The impact of regulation on innovation. *American Economic Review*, *113*(11), 2894–2936.

Amaral, P. S., & Quintin, E. (2006). A competitive model of the informal sector. *Journal of monetary Economics*, *53*(7), 1541–1553.

Amin, M., & Okou, C. (2020). Casting a shadow: Productivity of formal firms and informality. *Review of Development Economics*, *24*(4), 1610–1630.

Amirapu, A., & Gechter, M. (2020). Labor regulations and the cost of corruption: Evidence from the indian firm size distribution. *Review of Economics and Statistics*, *102*(1), 34–48.

Angelino, A., Tassinari, M., Barbieri, E., & Di Tommaso, M. R. (2021). Institutional and economic transition in vietnam: Analysing the heterogeneity in firms' perceptions of business environment constraints. *Competition & Change*, *25*(1), 52–72.

Benedek, M. D., Deb, M. P., Gracia, M. B., Saksonovs, M. S., Shabunina, M. A., & Budina, M. N. T. (2017). *The right kind of help? tax incentives for staying small*. International Monetary Fund.

Benhassine, N., McKenzie, D., Pouliquen, V., & Santini, M. (2018). Does inducing informal firms to formalize make sense? experimental evidence from benin. *Journal of Public Economics*, *157*, 1–14.

Benjamin, N. C., & Mbaye, A. A. (2012). The informal sector, productivity, and enforcement in w est a frica: A firm-level analysis. *Review of Development Economics*, *16*(4), 664–680.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507.

Bennedsen, M., Simintzi, E., Tsoutsoura, M., & Wolfenzon, D. (2022). Do firms respond to gender pay gap transparency? *The Journal of Finance*, *77*(4), 2051–2091.

Bertrand, M., Hsieh, C.-T., & Tsivanidis, N. (2021). *Contract labor and firm growth in india* (tech. rep.). National Bureau of Economic Research.

Bruhn, M., & McKenzie, D. (2014). Entry regulation and the formalization of microenterprises in developing countries. *The World Bank Research Observer*, *29*(2), 186–201.

Candela, M. A. (2013). *Size-dependent policies and firm behavior*. University of California, Berkeley.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2024). *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press.

Dabla-Norris, M. E., Jaramillo, L., Lima, F., & Sollaci, A. (2018). *Size dependent policies, informality and misallocation*. International Monetary Fund.

De Soto, H. (1989). *The other path: The invisible revolution in the third world*. Harper & Row New York.

Ferguson, J., & Kim, O. (2023). *Reassessing china's rural reforms: The view from outer space* (tech. rep.). Mimeo. University of California at Berkeley.

Garibaldi, P., Pacelli, L., & Borgarello, A. (2004). Employment protection legislation and the size of firms. *Giornale degli economisti e annali di economia*, 33–68.

Garicano, L., Lelarge, C., & Van Reenen, J. (2016). Firm size distortions and the productivity distribution: Evidence from france. *American Economic Review*, *106*(11), 3439–3479.

Gourio, F., & Roys, N. (2014). Size-dependent regulations, firm size distribution, and reallocation. *Quantitative Economics*, *5*(2), 377–416.

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

Guner, N., Ventura, G., & Yi, X. (2006). How costly are restrictions on size? *Japan and the World Economy*, *18*(3), 302–320.

Institute of Labour Science and Social Affairs & the University of Copenhagen. (2009, July). *Manual - guidelines: Survey of small and medium scale manufacturing enterprises (smes) in vietnam* (tech. rep.).

Kugler, A., & Pica, G. (2008). Effects of employment protection on worker and job flows: Evidence from the 1990 italian reform. *Labour Economics*, *15*(1), 78–95.

La Porta, R., & Shleifer, A. (2008). *The unofficial economy and economic development* (tech. rep.). National Bureau of Economic Research.

La Porta, R., & Shleifer, A. (2014). Informality and development. *Journal of economic perspectives*, *28*(3), 109–126.

Lalive, R. (2008). How do extended benefits affect unemployment duration? a regression discontinuity approach. *Journal of econometrics*, *142*(2), 785–806.

Lewis, W. A., et al. (1954). Economic development with unlimited supplies of labour.

McCaig, B., & Pavcnik, N. (2013). *Moving out of agriculture: Structural change in vietnam* (tech. rep.). National Bureau of Economic Research.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, *142*(2), 698–714.

McKenzie, D. (2017). Identifying and spurring high-growth entrepreneurship: Experimental evidence from a business plan competition. *American Economic Review*, *107*(8), 2278–2307.

Morikawa, M. (2023). Compliance costs and productivity: An approach from working hours. *Journal of Regulatory Economics*, *63*(3), 117–137.

Mulligan, C. B. (2020). The employer penalty, voluntary compliance, and the size distribution of firms: Evidence from a survey of small businesses. *Tax Policy and the Economy*, *34*(1), 139–171.

Padmakumar, K. (2022). Small by choice? reassessing the aggregate implications of size-based regulations.

Perry, G. (2007). *Informality: Exit and exclusion*. World Bank Publications.

Prado, M. (2011). Government policy in the formal and informal sectors. *European Economic Review*, *55*(8), 1120–1136.

Qian, L., & Vereshchagina, G. (2022). *Capital financing constraints, size-dependent distortions, and aggregate productivity* (tech. rep.). mimeo.

Rand, J., & Tarp, F. (2020). *Micro, small, and medium enterprises in vietnam*. Oxford University Press.

Schivardi, F., & Torrini, R. (2008). Identifying the effects of firing restrictions through size-contingent differences in regulation. *Labour Economics*, *15*(3), 482–511.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751–754.

Takahashi, M. (2024). Difference-in-discontinuities design. *Available at SSRN 4686891*.

Ulyssea, G. (2018). Firms, informality, and development: Theory and evidence from brazil. *American Economic Review*, *108*(8), 2015–2047.

Appendix

A Formal, Paid Worker Calculation Details

This appendix details the derivation for Equation 1 given the following definitions for each variable:

- $C_{it} \in [0, 1]$: percentage of full-time workers with a formal, written labor contract for firm i at time t
- FT_{it}: number of full-time workers
- PT_{it}: number of part-time workers
- Cas_{it}: number of casual workers
- $U_{it} \in [0, 1]$: percentage of total workers (i.e., summation of full-time, part-time, and casual workers) that are unpaid

First, the number of unpaid total workers is:

$$Unpaid_{it} = U_{it} \times (FT_{it} + PT_{it} + Cas_{it})$$

Assuming that unpaid workers are distributed proportionately across all worker types, the share of unpaid workers that are full-time would be expressed as:

$$\frac{FT_{it}}{FT_{it} + PT_{it} + Cas_{it}}$$

Thus the number of unpaid full-time workers is:

$$UnpaidFT_{it} = U_{it} \times (FT_{it} + PT_{it} + Cas_{it}) \times \frac{FT_{it}}{FT_{it} + PT_{it} + Cas_{it}} = U_{it} \times FT_{it}$$

This means that the number of **paid**, full-time workers can be expressed as:

$$PaidFT_{it} = FT_{it} - (U_{it} \times FT_{it}) = FT_{it} \times (1 - U_{it})$$

Then, the number of formally contracted, paid (FP) full-time workers is:

$$FP_{it} = C_{it} \times FT_{it} \times (1 - U_{it}), \text{ which is the same as Equation 1.}$$

B Proofs of Theoretical Predictions

This appendix provides the full derivations and intermediate results used to prove Theorems 1, 2, and 3. These steps are omitted from Section 4 for brevity but are included here for completeness.

B.1 Proof of Theorem 1

Proof. Without loss of generality, let $L_{IP,i,t} \equiv L_{IP,i,t+1}, L_{U,i,t} \equiv L_{U,i,t+1}$, and $f_{i,t} \equiv f_{i,t+1}$, holding fixed these choice variables at their pre-policy levels. As a result, Equation 3 is now defined as:

$$L^{eff}_{i,t+1}(L_{FP,i,t+1}) = \gamma_{FP} L_{FP,i,t+1} + \bar{L}_{i,t}, \text{ where } \bar{L}_{i,t} \equiv \gamma_{IP} L_{IP,i,t} + \gamma_U L_{U,i,t}$$

Given this, Equation 4 now only depends on the number of formal, paid (FP) workers:

$$C_{i,t+1}(L_{FP,i,t+1}) = \begin{cases} C_{FP} L_{FP,i,t+1}, & \text{if } L_{FP,i,t+1} < 10 \\ C_{FP} L_{FP,i,t+1} + C_T, & \text{if } L_{FP,i,t+1} \geq 10 \text{ and } S_{i,t+1} = 1 \\ C_{FP} L_{FP,i,t+1} + \phi(R_{i,t+1}) P_{t+1}, & \text{if } L_{FP,i,t+1} \geq 10 \text{ and } S_{i,t+1} = 0 \end{cases}$$

Thus, firm i's profit maximization problem (i.e., Equation 5) becomes:

$$\max_{L_{FP,i,t+1}} A_i \cdot \left(\gamma_{FP} L_{FP,i,t+1} + \bar{L}_{i,t}\right)^{\alpha} - C_{i,t+1}(L_{FP,i,t+1})$$

I evaluate whether it is profitable for firm i to increase $L_{FP,i,t+1}$ from 9 to 10 workers. Although C_T and $\phi(R_{i,t+1})P_{t+1}$ are not marginal in the derivative sense, they induce a discrete cost increase at $L_{FP,i,t+1} = 10$. In increasing $L_{FP,i,t+1}$ from 9 to 10, the increase in effective labor is γ_{FP}, so a first-order Taylor expansion of Equation 2 at $L_{FP,i,t+1} = 9$ yields the approximate output gain:

$$A_i \cdot \left(L^{eff}_{i,t+1}(10)\right)^{\alpha} \approx A_i \cdot \left(L^{eff}_{i,t+1}(9)\right)^{\alpha} + A_i \alpha \cdot \left(L^{eff}_{i,t+1}(9)\right)^{\alpha-1} \cdot \gamma_{FP}$$

Thus the approximate change in output for firm i from employing the 10^{th} FP worker in period $t+1$ holding all else constant is:

$$\Delta y_{i,t+1} \approx A_i \alpha \cdot \left(L^{eff}_{i,t+1}(9)\right)^{\alpha-1} \cdot \gamma_{FP}$$

Having derived the marginal benefit in output terms, I now compare this to the corresponding discrete cost increase, which includes both the marginal cost of the 10^{th} FP employee (i.e., C_{FP}) *plus* a fixed cost associated with crossing the policy threshold:

$$\Delta C_{i,t+1} = \begin{cases} C_{FP} + C_T, & \text{if } S_{i,t+1} = 1 \\ C_{FP} + \phi(R_{i,t+1}) P_{t+1}, & \text{if } S_{i,t+1} = 0 \end{cases}$$

Firm i will refrain from employing a 10^{th} FP worker if the additional revenue from doing so is smaller than this discrete increase in cost. That is, firm i prefers to operate with $L_{FP,i,t+1} = 9$ if:

$$A_i \alpha \cdot \left(L^{eff}_{i,t+1}(9)\right)^{\alpha-1} \cdot \gamma_{FP} < \begin{cases} C_{FP} + C_T, & \text{if } S_{i,t+1} = 1 \\ C_{FP} + \phi(R_{i,t+1}) P_{t+1}, & \text{if } S_{i,t+1} = 0 \end{cases}$$

This inequality defines a region of the firm productivity distribution in which those with lower productivity (i.e., A_i) or smaller complementary labor inputs (i.e., $\bar{L}_{i,t+1}$) find the marginal benefit from hiring the 10th FP worker too small to justify the fixed costs induced by the 2012 employee threshold policy. That is, less productive firms or those relying less on informal or unpaid labor are more likely to find the policy threshold binding. Considering this cost-benefit comparison, it directly follows that some firms will find it optimal to employ 9 FP workers. In equilibrium, this generates a kink in firm i's labor demand function, resulting in excess mass in the distribution of FP workers at $L_{FP,i,t+1} = 9$ post-policy—a phenomenon known as "bunching." This prediction is consistent with the theoretical and empirical literature on threshold effects in labor policy (Garibaldi et al., 2004; Kugler & Pica, 2008; Gourio & Roys, 2014; Garicano et al., 2016; Amirapu & Gechter, 2020) and provides a clear, testable implication of the model. Yet if no bunching is observed, it suggests that the higher discrete cost at the threshold is *not* binding for most firms. This could arise from high firm productivity, weak policy enforcement, a low perceived financial penalty for non-compliance, or effective substitution toward informal or unpaid labor to absorb the cost increase without reducing output. In such cases, firms can cross the 10-FP employee threshold without substantial profit loss and the cost shock does not distort the firm size distribution. □

B.2 Proof of Theorem 2

Proof. Let firm i select labor inputs $\ell_{i,t+1} = \{L_{FP,i,t+1}, L_{IP,i,t+1}, L_{U,i,t+1}\}$, a firm formality status choice set $f_{i,t+1} = \{R_{i,t+1}, K_{i,t+1}\}$, and a threshold compliance decision $S_{i,t+1} \in \{0, 1\}$ to maximize profits as defined in Equation 5.

Given the production function in Equation 2 and the effective labor formulation in Equation 3, the first-order conditions (FOCs) for an interior solution are:

$$A_i \alpha \cdot (L^{eff}_{i,t+1})^{\alpha-1} \cdot \gamma_{FP} = C_{FP} + \tau \tag{B.1}$$

$$A_i \alpha \cdot (L^{eff}_{i,t+1})^{\alpha-1} \cdot \gamma_{IP} = C_{IP} \tag{B.2}$$

$$A_i \alpha \cdot (L^{eff}_{i,t+1})^{\alpha-1} \cdot \gamma_{U} = \psi'(L_{U,i,t+1}) \tag{B.3}$$

where τ is the marginal threshold cost associated with crossing the 10-FP employee threshold, which is defined as:

$$\tau = \frac{\partial}{\partial L_{FP,i,t+1}} \left(C_T \cdot \mathbf{1}\{L_{FP,i,t+1} \geq 10, S_{i,t+1} = 1\} + \phi(R_{i,t+1}) P_{t+1} \cdot \mathbf{1}\{L_{FP,i,t+1} \geq 10, S_{i,t+1} = 0\} \right).$$

The 2012 employee threshold policy causes τ to increase discontinuously at $L_{FP,i,t+1} = 10$ by raising both the expected penalty from P_t to P_{t+1} (where $P_t < P_{t+1}$) and introducing a substantially higher compliance cost C_T. These threshold-induced costs apply when firm i employs its 10th FP worker. This discrete increase in τ raises the effective marginal cost of employing $L_{FP,i,t+1}$, which by the first FOC in Equation B.1 implies a *reduction* in the optimal level of $L_{FP,i,t+1}$:

$$\frac{\partial L_{FP,i,t+1}}{\partial \tau} < 0$$

Given a fixed output level and the presence of diminishing returns to labor, firm i seeks to maintain its optimal effective labor input $L^{eff*}_{i,t+t}$ by substituting toward the relatively cheaper worker types, $L_{IP,i,t+1}$ and $L_{U,i,t+1}$, until their marginal value products equal their marginal costs. From the FOCs, the marginal costs of both $L_{IP,i,t+1}$ and $L_{U,i,t+1}$ are unaffected by τ; thus holding $L^{eff*}_{i,t+1}$ constant, firm i reallocates labor away from FP workers toward one or both unregulated margins, implying non-negative comparative statics for the substituting worker type(s):[30]

$$\frac{\partial L_{IP,i,t+1}}{\partial \tau} \geq 0, \quad \frac{\partial L_{U,i,t+1}}{\partial \tau} \geq 0$$

These comparative statics are derived under a partial-equilibrium assumption in which optimal effective labor input $L^{eff*}_{i,t+1}$ is held constant. They capture within-firm substitution patterns between worker types in response to an increase in the marginal cost of FP labor—abstracting from any general equilibrium effects. Considering this assumption, the threshold-induced increase in FP labor costs post-policy leads firm i to substitute labor input toward informal or unpaid alternatives.

To further characterize this reallocation under the fixed-output condition (where optimal effective labor input $L^{eff*}_{i,t+1}$ is held constant), the per-effective-labor marginal cost of informal, paid labor is weakly less than that of unpaid labor at the local optimum (i.e., $c_{IP} \leq c_U(L^*_{U,i,t+1})$). Given this cost ranking and the higher effective marginal cost of $L_{FP,i,t+1}$ when τ rises, firm i reallocates labor toward the *cheapest* unregulated labor margin. Therefore $\frac{\partial L_{FP,i,t+1}}{\partial \tau} < 0$ and $\frac{\partial L_{IP,i,t+1}}{\partial \tau} \geq 0$, while $\frac{\partial L_{U,i,t+1}}{\partial \tau}$ remains residual under the constant-$L^{eff*}_{i,t+1}$ constraint and need not be strictly positive.

□

B.2.1 Comparative Statics Derivations

To derive the comparative statics of the different worker types with respect to τ, I first implicitly differentiate Equation B.1 with respect to τ, holding all other parameters fixed:

$$A_i\alpha(\alpha - 1) \cdot (L^{eff}_{i,t+1})^{\alpha-2} \cdot \gamma^2_{FP} \cdot \frac{\partial L_{FP,i,t+1}}{\partial \tau} = 1$$

Solving for $\frac{\partial L_{FP,i,t+1}}{\partial \tau}$ yields:

$$\frac{\partial L_{FP,i,t+1}}{\partial \tau} = \frac{1}{A_i\alpha(\alpha - 1) \cdot (L^{eff}_{i,t+1})^{\alpha-2} \cdot \gamma^2_{FP}} \tag{B.4}$$

Since $A_i > 0$, $\gamma_{FP} > 0$, and $(L^{eff}_{i,t+1})^{\alpha-2} > 0$, while $\alpha \in (0, 1)$ implies that $(\alpha - 1) < 0$, the denominator of Equation B.4 is negative. Thus:

[30] Mathematical derivations for these three comparative statics are shown in Appendix B.2.1.

$$\frac{\partial L_{FP,i,t+1}}{\partial \tau} < 0 \tag{B.5}$$

Therefore, a discrete increase in the threshold-induced cost τ raises the effective marginal cost of hiring FP workers, reducing the optimal level of $L_{FP,i,t+1}$ post-policy.

Implicitly differentiating Equation B.2 with respect to τ, I get:

$$A_i\alpha(\alpha-1)\cdot(L^{eff}_{i,t+1})^{\alpha-2}\cdot\gamma_{IP}\left(\gamma_{FP}\frac{\partial L_{FP,i,t+1}}{\partial \tau}+\gamma_{IP}\frac{\partial L_{IP,i,t+1}}{\partial \tau}+\gamma_{U}\frac{\partial L_{U,i,t+1}}{\partial \tau}\right)=0 \tag{B.6}$$

Because only $L_{FP,i,t+1}$ directly enters the threshold cost, τ affects Equation B.6 only *indirectly* through the change in $L_{FP,i,t+1}$. Solving for $\frac{\partial L_{IP,i,t+1}}{\partial \tau}$ yields:

$$\frac{\partial L_{IP,i,t+1}}{\partial \tau} = -\frac{\gamma_{FP}}{\gamma_{IP}}\frac{\partial L_{FP,i,t+1}}{\partial \tau} - \frac{\gamma_{U}}{\gamma_{IP}}\frac{\partial L_{U,i,t+1}}{\partial \tau} \tag{B.7}$$

Under the partial-equilibrium assumption that optimal effective labor input $L^{eff*}_{i,t+1}$ is held constant (i.e., firm output stays fixed), the second term can be omitted as $L^{eff}_{i,t+1}$ does not vary with τ. In Equation B.7, $\frac{\partial L_{U,i,t+1}}{\partial \tau}$ serves as a residual adjustment that endogenously preserves the constant-optimal-effective-labor condition, rather than independently determining the direction of within-firm substitution (i.e., $L^*_{U,i,t+1}$ adjusts to maintain $L^{eff*}_{i,t+1}$ even if $L^*_{IP,i,t+1}$ does not). In this setup, the negative result in Equation B.5 and the negative sign on the corresponding term in Equation B.7 imply that a reduction in FP labor induces an increase in informal, paid labor. However, since the constant-$L^{eff*}_{i,t+1}$ constraint can also be satisfied by adjustments in unpaid labor, substitution into informal, paid labor is not strictly required; thus, given $\frac{\partial L_{FP,i,t+1}}{\partial \tau} < 0$, the comparative static for $L_{IP,i,t+1}$ is weakly non-negative rather than strictly positive:

$$\frac{\partial L_{IP,i,t+1}}{\partial \tau} \geq 0 \tag{B.8}$$

Hence, when the cost of formal employment rises, firms may substitute toward informal, paid labor.

Lastly, implicitly differentiating Equation B.3 with respect to τ, I get:

$$\begin{aligned} A_i\alpha(\alpha-1)\cdot(L^{eff}_{i,t+1})^{\alpha-2}\cdot\gamma_{U}\left(\gamma_{FP}\frac{\partial L_{FP,i,t+1}}{\partial \tau}+\gamma_{IP}\frac{\partial L_{IP,i,t+1}}{\partial \tau}+\gamma_{U}\frac{\partial L_{U,i,t+1}}{\partial \tau}\right) \\ = \psi''(L_{U,i,t+1})\cdot\frac{\partial L_{U,i,t+1}}{\partial \tau} \end{aligned}$$

Solving for $\frac{\partial L_{U,i,t+1}}{\partial \tau}$ yields:

$$\frac{\partial L_{U,i,t+1}}{\partial \tau} = \frac{-A_i\alpha(\alpha-1)\cdot (L_{i,t+1}^{eff})^{\alpha-2}\cdot \gamma_U \left(\gamma_{FP}\frac{\partial L_{FP,i,t+1}}{\partial \tau} + \gamma_{IP}\frac{\partial L_{IP,i,t+1}}{\partial \tau}\right)}{A_i\alpha(\alpha-1)\cdot (L_{i,t+1}^{eff})^{\alpha-2}\cdot \gamma_U^2 - \psi''(L_{U,i,t+1})} \tag{B.9}$$

Given that $\psi''(L_{U,i,t+1}) \geq 0$ by the convexity of disutility and $A_i\alpha(\alpha-1) < 0$, the denominator of Equation B.9 is negative. In the numerator, both γ_{FP} and γ_{IP} are positive; however, $\gamma_{FP}\frac{\partial L_{FP,i,t+1}}{\partial \tau}$ is negative (see the result in Equation B.5) while $\gamma_{IP}\frac{\partial L_{IP,i,t+1}}{\partial \tau}$ is weakly non-negative (see the result in Equation B.8). Because the reduction in FP labor dominates under the constant-optimal-effective-labor condition, the entire expression inside the parentheses remains negative. Multiplying this negative term by $-A_i\alpha(\alpha-1) > 0$ renders the numerator negative as well. Consequently:

$$\frac{\partial L_{U,i,t+1}}{\partial \tau} \geq 0$$

The above derivations formally establish that a discrete increase in the threshold-induced cost τ leads to the following predictions:

$$\frac{\partial L_{FP,i,t+1}}{\partial \tau} < 0, \quad \frac{\partial L_{IP,i,t+1}}{\partial \tau} \geq 0, \quad \frac{\partial L_{U,i,t+1}}{\partial \tau} \geq 0$$

B.3 Proof of Theorem 3

Proof. In period $t+1$, the expected penalty for non-compliance at the threshold increases from P_t to P_{t+1}, where $P_t < P_{t+1}$. A firm i with $L_{FP,i,t+1} \geq 10$ chooses whether to register with the government (i.e., $R_{i,t+1} \in \{0,1\}$) and whether to submit IWRs (i.e., $S_{i,t+1} \in \{0,1\}$), subject to the institutional constraint that submitting IWRs requires government registration (i.e., $S_{i,t+1} = 1 \Rightarrow R_{i,t+1} = 1$). Components of firm i's cost function (see Equation 4) that depend on $R_{i,t+1}$ or $S_{i,t+1}$ include the registration cost $C_R R_{i,t+1}$, threshold-compliance cost $C_T \cdot 1\{S_{i,t+1} = 1\}$, and the expected penalty for non-compliance $\phi(R_{i,t+1})P_{t+1} \cdot 1\{S_{i,t+1} = 0\}$. All remaining costs are identical across choices.

If firm i does not submit IWRs and thus remains unregistered (i.e., $(R_{i,t+1}, S_{i,t+1}) = (0,0)$), then it incurs the expected penalty of $\phi(0)P_{t+1} = \phi_L P_{t+1}$. If instead it submits IWRs and thus registers with the government (i.e., $(R_{i,t+1}, S_{i,t+1}) = (1,1)$), it pays the compliance cost $C_R + C_T$ and avoids the penalty entirely. The strategy $(R_{i,t+1}, S_{i,t+1}) = (1,0)$ is strictly dominated as it yields the higher penalty $\phi(1)P_{t+1} = \phi_H P_{t+1}$ where $\phi_L < \phi_H$ without avoiding non-compliance. By definition, the strategy $(R_{i,t+1}, S_{i,t+1}) = (0,1)$ is infeasible.

However, whenever

$$\phi_L P_{t+1} > C_R + C_T,$$

firm i strictly prefers $(1,1)$ to $(0,0)$. Therefore, for any firm with $L_{FP,i,t+1} \geq 10$ satisfying this inequality, the cost-minimizing choice is to submit IWRs and thereby register with the government.

Having established that firms with $L_{FP,i,t+1} \geq 10$ optimally choose government registration when $\phi_L P_{t+1} > C_R + C_T$, I now characterize their post-policy labor composition decision.

Based on Equation 2, firm i's output only depends on $L_{i,t+1}^{eff}$. Hence, among any two feasible labor input bundles that compare $K_{i,t+1} = 1$ and $K_{i,t+1} = 0$ while delivering the same $L_{i,t+1}^{eff*}$, the profit comparison reduces to a total cost comparison. Let (marginal) per-effective-labor unit costs for paid (unpaid) worker types be defined compactly by:

$$c_k = \frac{MMC_k}{\gamma_k} = \begin{cases} C_{FP}, & \text{if } k = FP \text{ (formal, paid)} \\ \dfrac{w_{IP}}{\theta}, & \text{if } k = IP \text{ (informal, paid)} \\ \dfrac{\psi'(L_{U,i,t+1})}{\delta}, & \text{if } k = U \text{ (unpaid)} \end{cases} \quad \text{where } MMC_k = \frac{\partial VarCost_k}{\partial L_k}$$

where both paid labor inputs have linear money costs such that their marginal money cost (MMC) equals the unit price (i.e., $MMC_{FP} = C_{FP}$ and $MMC_{IP} = w_{IP}$) while unpaid labor has convex cost $\psi(L_{U,i,t+1})$ with $MMC_U = \psi'(L_{U,i,t+1})$. If $K_{i,t+1} = 1$, firm i must provide all workers with a formal, written contract (i.e., $L_{IP,i,t+1} = 0$) and consequently firm i incurs the fixed cost of labor contract compliance $C_K > 0$. If $K_{i,t+1} = 0$, at least *one* of firm i's workers does not have a formal, written contract and thus no fixed labor contract compliance cost is incurred.

First, fix a profit-maximizing labor input bundle under $K_{i,t+1} = 0$ that achieves $L_{i,t+1}^{eff*}$ with labor inputs $L_{FP,i,t+1}^{*}$, $L_{IP,i,t+1}^{*}$, and $L_{U,i,t+1}^{*}$. Then consider any feasible labor input bundle with $K_{i,t+1} = 1$ instead that also achieves $L_{i,t+1}^{eff*}$. Because $L_{IP,i,t+1} = 0$ when $K_{i,t+1} = 1$, firm i must replicate the effective labor input contributed by $\theta L_{IP,i,t+1}^{*}$ using FP workers, unpaid workers, or both.[31] Replacing $\theta L_{IP,i,t+1}^{*}$ units of effective labor with *just* FP workers raises variable cost by:

$$(c_{FP} - c_{IP}) \cdot \theta L_{IP,i,t+1}^{*}$$

Replacing $\theta L_{IP,i,t+1}^{*}$ units of effective labor with *just* unpaid workers raises variable cost by:

$$(c_U(L_{U,i,t+1}) - c_{IP}) \cdot \theta L_{IP,i,t+1}^{*}.$$

Under the condition $c_{IP} \leq c_{FP}$, the FP worker-only replacement is weakly more expensive on a per-effective-unit basis than retaining informal, paid workers. Therefore, any $K_{i,t+1} = 1$ labor input bundle that achieves $L_{i,t+1}^{eff*}$ cannot have lower variable costs than a $K_{i,t+1} = 0$ labor input bundle that is identical in both $L_{FP,i,t+1}$ and $L_{U,i,t+1}$ but retains $L_{IP,i,t+1}^{*}$. Moreover, even if variable costs were equalized by optimally adjusting the mix of FP and unpaid workers, moving from $K_{i,t+1} = 0$ to $K_{i,t+1} = 1$ strictly raises *total* costs by $C_K > 0$ without affecting output. Hence, any $K_{i,t+1} = 1$ labor input bundle yields strictly lower profits than the $K_{i,t+1} = 0$ optimum that attains $L_{i,t+1}^{eff*}$, meaning profit-maximizing firm i chooses $K_{i,t+1} = 0$. If firm i chooses $L_{FP,i,t+1} \geq 10$ and $S_{i,t+1} = 1$, which presumes $R_{i,t+1} = 1$, its observed post-policy formality status is $(R_{i,t+1}, K_{i,t+1}) = (1, 0)$. Per Table 2, post-policy firm i is government-registered yet still employs informal workers (i.e., is formal

[31] Note that since $\gamma_{IP} = \theta$ as defined in Section 4, $\gamma_{IP} L_{IP,i,t+1}^{*} = \theta L_{IP,i,t+1}^{*}$.

but operates informally). Here,"operating informally" can derive from hiring either informal, paid workers or unpaid workers without formal, written contracts. Both satisfy $K_{i,t+1} = 0$.

It remains to give a sufficient condition for $L_{IP,i,t+1} > 0$ at the $K_{i,t+1} = 0$ optimum if $c_{IP} \leq c_{FP}$ and $c_{IP} \leq c_U(L^*_{U,i,t+1})$. Suppose, contrary to this, that $L^*_{IP,i,t+1} = 0$. Then consider the following two within-firm substitutions that would keep $L^{eff*}_{i,t+1}$ constant: (1) increasing $L_{IP,i,t+1}$ by $\varepsilon > 0$ and decreasing $L_{FP,i,t+1}$ by $\theta\varepsilon$, and (2) increasing $L_{IP,i,t+1}$ by $\varepsilon > 0$ and decreasing $L_{U,i,t+1}$ by $(\frac{\theta}{\delta})\varepsilon$.[32]

With $L^{eff*}_{i,t+1}$ fixed, the associated variable cost changes under (1) and (2), respectively, are:[33]

$$\Delta V_{FP\to IP}(\varepsilon) = \theta\varepsilon(c_{IP} - c_{FP}) \text{ if } c_{IP} < c_{FP} \tag{B.10}$$

$$\Delta V_{U\to IP}(\varepsilon) = \theta\varepsilon(c_{IP} - c_U(L^*_{U,i,t+1})) \text{ if } c_{IP} \leq c_U(L^*_{U,i,t+1}) \tag{B.11}$$

Substituting toward informal, paid workers from either FP workers (when $c_{IP} < c_{FP}$) or from unpaid workers (when $c_{IP} \leq c_U(L^*_{U,i,t+1})$)—while holding $L^{eff*}_{i,t+1}$ constant—weakly decreases variable cost, and strictly so if either of these two inequalities is strict. With output and prices fixed, either substitution increases profits, contradicting $L^*_{IP,i,t+1} = 0$. Thus the cost-minimizing solution satisfies $L_{IP,i,t+1} > 0$ whenever $c_{IP} \leq c_U(L^*_{U,i,t+1})$ with at least one strict inequality (i.e., $c_{IP} < c_{FP}$ or $c_{IP} < c_U(L^*_{U,i,t+1})$). If $c_{IP} = c_{FP} = c_U(L^*_{U,i,t+1})$, no strictly profitable substitution exists. □

B.3.1 Cost-Change Expressions Derivations

Given Equation 3, total variable labor cost post-policy for firm i is defined as:

$$V_{i,t+1} = C_{FP}L_{FP,i,t+1} + C_{IP}L_{IP,i,t+1} + \psi(L_{U,i,t+1}),$$

where $C_{FP} = w_{FP} + b_{FP}$ is the total per-worker cost of FP labor; $C_{IP} = w_{IP}$ is the per-worker cost of informal, paid labor; and, $\psi(L_{U,i,t+1})$ is the non-monetary cost function for unpaid workers (see Section 4 for more detailed definitions).

The corresponding per-effective-labor costs for (1) a FP worker, (2) an informal, paid worker, and (3) an unpaid worker—each normalized by its respective productivity weight—are as follows:

$$c_{FP} = \frac{C_{FP}}{\gamma_{FP}} = C_{FP}, \quad c_{IP} = \frac{C_{IP}}{\gamma_{IP}} = \frac{w_{IP}}{\theta}, \quad c_U(L_{U,i,t+1}) = \frac{\psi'(L_{U,i,t+1})}{\gamma_U} = \frac{\psi'(L_{U,i,t+1})}{\delta} \tag{B.12}$$

Case 1: Substitution from FP to IP Labor. This case examines within-firm substitution from FP to informal, paid labor while keeping $L^{eff*}_{i,t+1}$ fixed to determine whether such a shift reduces

[32]The coefficients $\theta\varepsilon$ and $(\frac{\theta}{\delta})\varepsilon$ follow from Equation 3: $L^{eff} = \gamma_{FP}L_{FP,i,t+1} + \gamma_{IP}L_{IP,i,t+1} + \gamma_U L_{U,i,t+1}$. An increase of $\varepsilon > 0$ in $L_{IP,i,t+1}$ adds $\gamma_{IP}\varepsilon = \theta\varepsilon$ units of effective labor. To keep $L^{eff*}_{i,t+1}$ constant, the firm must therefore either reduce $L_{FP,i,t+1}$ by $\theta\varepsilon$ (since $\gamma_{FP} = 1$) or reduce $L_{U,i,t+1}$ by $(\frac{\theta}{\delta})\varepsilon$ (since $\gamma_U = \delta$).

[33]Mathematical derivations for these two cost changes are shown in Appendix B.3.1.

firm i's total variable cost. To derive Equation B.10, consider increasing $L_{IP,i,t+1}$ by $\varepsilon > 0$ and decreasing $L_{FP,i,t+1}$ by $\theta\varepsilon$ to hold $L_{i,t+1}^{eff*}$ constant:

$$\Delta L_{IP} = \varepsilon, \qquad \Delta L_{FP} = -\theta\varepsilon.$$

Increasing $L_{IP,i,t+1}$ by $\varepsilon > 0$ adds $\theta\varepsilon$ units of effective labor. To keep $L_{i,t+1}^{eff*}$ (and thus output) constant, the firm must reduce $L_{FP,i,t+1}$ by $\theta\varepsilon$ (since $\gamma_{FP} = 1$).

The change in variable cost is:

$$\begin{aligned} \Delta V_{FP\to IP} &= C_{IP}\Delta L_{IP} + C_{FP}\Delta L_{FP} \\ &= C_{IP}(\varepsilon) + C_{FP}(-\theta\varepsilon) = \varepsilon(C_{IP} - \theta C_{FP}). \end{aligned}$$

Using $c_{FP} = C_{FP}$ and $C_{IP} = \theta c_{IP}$ from the per-effective-labor costs in Equation B.12 gives:

$$\Delta V_{FP\to IP}(\varepsilon) = \theta\varepsilon(c_{IP} - c_{FP}). \tag{B.13}$$

Case 2: Substitution from Unpaid to IP Labor. This case examines within-firm substitution from unpaid to informal, paid labor while holding $L_{i,t+1}^{eff*}$ constant to assess whether such a shift reduces firm i's total variable cost. To derive Equation B.11, consider increasing $L_{IP,i,t+1}$ by $\varepsilon > 0$ and decreasing $L_{U,i,t+1}$ by $(\frac{\theta}{\delta})\varepsilon$ to keep $L_{i,t+1}^{eff*}$ fixed:

$$\Delta L_{IP} = \varepsilon, \qquad \Delta L_{U} = -(\tfrac{\theta}{\delta})\varepsilon.$$

Increasing $L_{IP,i,t+1}$ by $\varepsilon > 0$ adds $\theta\varepsilon$ units of effective labor. To keep $L_{i,t+1}^{eff*}$ (and thus output) constant, the firm must reduce $L_{U,i,t+1}$ by $(\frac{\theta}{\delta})\varepsilon$ (since $\gamma_U = \delta$).

The change in variable cost is:

$$\begin{aligned} \Delta V_{U\to IP} &= C_{IP}\Delta L_{IP} + \psi'(L_{U,i,t+1})\Delta L_U \\ &= C_{IP}(\varepsilon) + \psi'(L_{U,i,t+1})\big(-\tfrac{\theta}{\delta}\varepsilon\big) = \varepsilon\big(C_{IP} - \tfrac{\theta}{\delta}\psi'(L_{U,i,t+1})\big). \end{aligned}$$

Substituting $C_{IP} = \theta c_{IP}$ and $\psi'(L_{U,i,t+1}) = \delta \cdot c_U(L_{U,i,t+1})$ from the per-effective labor costs in Equation B.12 yields:

$$\Delta V_{U\to IP}(\varepsilon) = \theta\varepsilon(c_{IP} - c_U(L_{U,i,t+1})). \tag{B.14}$$

C Supplemental Tables & Figures

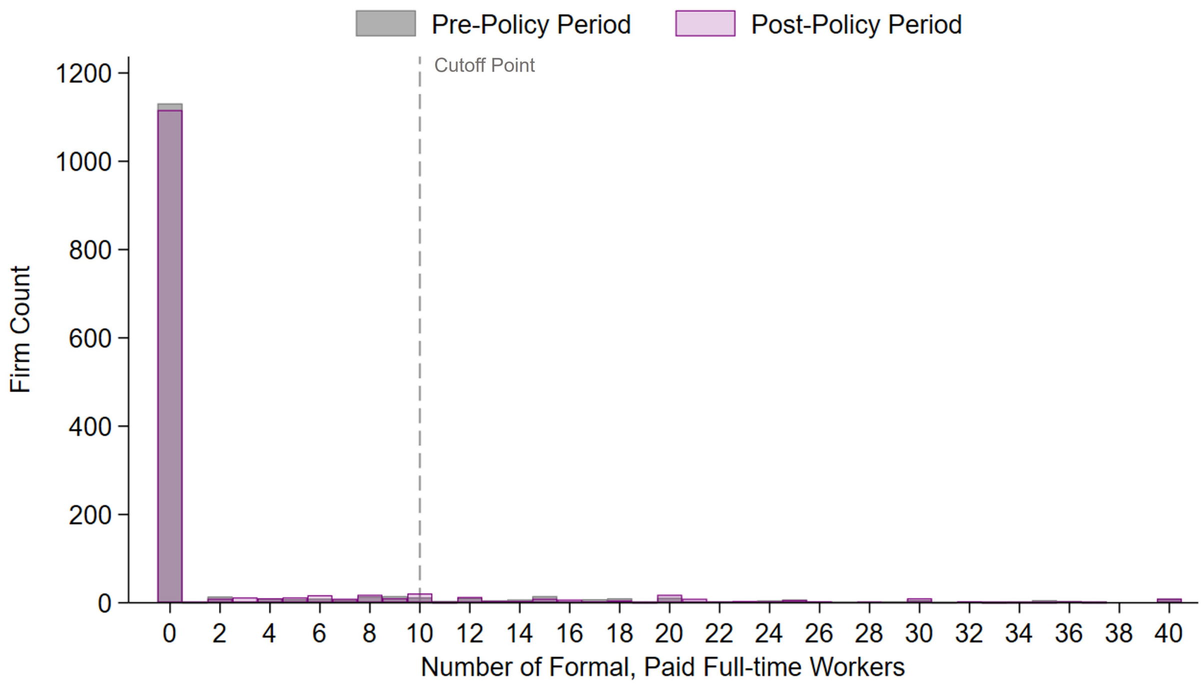


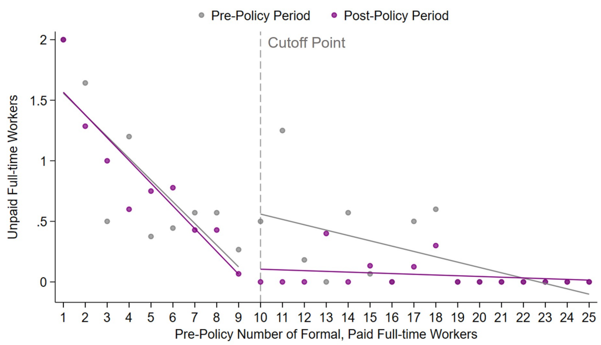
Figure C.1: Distribution of Running Variable Pre- and Post-Policy

Notes: This figure displays the distribution of the running variable (i.e., the number of formal, paid full-time workers) for the pre- and post-policy periods for the analyitcal sample (see Appendix E). Three issues are apparent: (1) the distribution is highly left-skewed, (2) the running variable is discrete and takes on few values, and (3) there is a mass point at zero where about 78 percent of firms cluster in both periods. These violate continuity assumptions required for conventional RDD estimation and instead motivate the use of a local randomization approach (see Section 5.3).

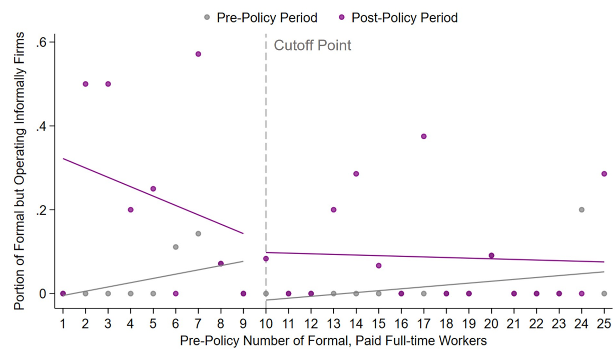
Table C.1: Firm Labor Contract Compliance Distribution by Year-End Period

Percentage of full-time workers with a formal, written labor contract	2008	2010	2012	2014	Notes
0%	65.8%	69.6%	68.2%	68.9%	Firms are **not** labor contract compliant
1–99%	14.7%	7.9%	11.1%	12.2%	Firms are partially labor contract compliant
100%	19.4%	22.5%	20.7%	18.9%	Firms are fully labor contract compliant
Observations	2,558	2,488	2,541	2,648	

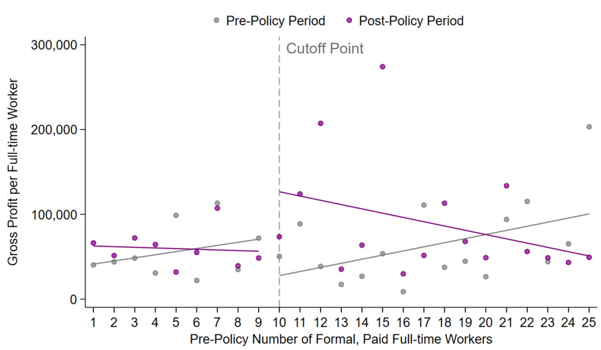
Note: Data reflect year-end values (i.e., December 31st).



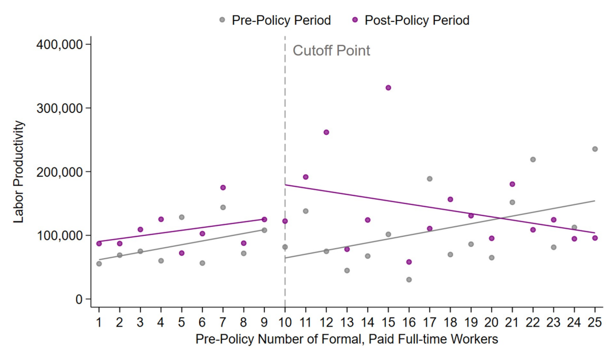
(A)



(B)



(C)



(D)

Figure C.2: Visualizing Changes in Discontinuity at the 10-Formal, Paid Employee Threshold: Pre- and Post-Policy Period Comparisons for Selected Firm Outcomes

Notes: Each panel shows a bin scatter plot overlaid with fitted linear regressions, comparing the association between the selected firm outcomes listed in Table 4 and the pre-policy number of formal, paid full-time workers across the pre- and post-policy periods. The vertical line denotes the cutoff introduced by the 2012 employee threshold policy. All firms above this threshold are subject to a higher financial penalty for non-compliance in the post-policy period. These plots illustrate changes in the discontinuity at the threshold over time within the optimal windows, providing visual evidence of the policy's localized temporal impact that complements the causal estimates reported in Table 4. Results in Panels C and D are reported in thousands of Vietnamese Dong (VND) and reflect year-end values.

Table C.2: Impact of Threshold-Based Labor Policy on Firm Size and Labor Composition

Variable	Total Workers		Part-time Workers		Casual Workers		Informal, Paid Full-time Workers		Unpaid Full-time Workers	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$Post_t \times Above10_i$	-25.165	-15.845	1.397	1.179	-20.002	-12.276	1.057	0.571	-0.404	-0.352*
	(27.973)	(11.564)	(1.022)	(1.098)	(24.122)	(10.002)	(2.050)	(1.325)	(0.344)	(0.209)
Group mean	10.861	9.125	0.389	0.413	1.444	0.875	0.361	1.263	0.444	0.763
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.522	0.724	0.492	0.500	0.514	0.557	0.564	0.518	0.691	0.767
Optimal window	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]
Observations	126	374	126	374	126	374	126	374	126	374

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the firm-level. "Group mean" reports the pre-policy mean of the outcome variable for the control group (i.e., firms just below the threshold) in each model specification.

Table C.3: Impact of Threshold-Based Labor Policy on Firm Performance

	Sales Revenue Per Full-time Worker		Gross Profit Per Full-time Worker		Labor Costs Per Full-time Worker		Labor Productivity		Capital Productivity	
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$Post_t \times Above10_i$	133,530	2,047,275	44,694	139,326**	-21,431	-16,162	23,263	123,164*	0.090	0.451
	(312,150)	(1,770,930)	(60,592)	(67,209)	(35,254)	(14,618)	(72,688)	(69,049)	(0.555)	(0.516)
Group mean	331,964	298,647	65,617	55,148	35,223	31,533	100,840	86,681	0.752	0.560
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.557	0.504	0.560	0.508	0.570	0.585	0.565	0.514	0.502	0.534
Optimal window	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]
Observations	126	374	126	374	126	374	126	374	126	374

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the firm-level. "Group mean" reports the pre-policy mean of the outcome variable for the control group (i.e., firms just below the threshold) in each model specification. Results are reported in thousands of Vietnamese Dong (except for capital productivity) and reflect year-end values. Gross profit equals sales revenue minus total costs—including raw materials, indirect expenses, and labor costs (e.g., wages, allowances, recruitment, training, as well as health, social, and unemployment insurance). Labor productivity is defined as value added per full-time worker while capital productivity is defined as value added per asset. Value added is calculated as sales revenue minus production costs (including raw materials and indirect expenses).

Table C.4: Impact of Threshold-Based Labor Policy on Firm Formality Status

	Fully Formal Firm		Formal Firm But Operates Informally		Informal Firm But Operates Formally		Fully Informal Firm	
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Post_t \times Above10_i$	-0.079	0.016	0.316**	0.182**	-0.248	-0.227	0.011	0.030
	(0.264)	(0.138)	(0.136)	(0.088)	(0.281)	(0.162)	(0.228)	(0.141)
Group mean	0.250	0.175	0.056	0.038	0.611	0.588	0.083	0.200
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.726	0.648	0.666	0.570	0.619	0.609	0.457	0.549
Optimal window	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]
Observations	126	374	126	374	126	374	126	374

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the firm-level. "Group mean" reports the pre-policy mean of the outcome variable for the control group (i.e., firms just below the threshold) in each model specification. Firm formality status definitions are listed in Table 2.

Table C.5: Financial Penalty as a Percent of Firm Economic Account Metrics Pre- and Post-Policy for Optimal Windows

	Pre-Policy Period				Post-Policy Period			
	Control		Treatment		Control		Treatment	
Percent of Average Annual	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]
Sales Revenue	0.03%	0.04%	0.03%	0.01%	0.08-0.16%	0.12-0.24%	0.08%-0.15%	0.02%-0.05%
Gross Profit	0.17%	0.25%	0.17%	0.09%	1.09-2.18%	1.07-2.15%	0.35%-0.70%	0.31%-0.61%
Labor Costs	0.32%	0.39%	0.24%	0.13%	0.48-0.95%	0.76-1.52%	0.85%-1.71%	0.49%-0.99%
Observations	36	80	27	107	36	80	27	107

Notes: Firms with less than 10 formal, paid full-time workers pre-policy form the control group; those with 10 or more constitute the treatment group. The optimal symmetric and asymmetric windows are [7, 12] and [1, 25], respectively. The financial penalty for non-compliance increased from 1 million VND pre-policy to between 5 and 10 million VND under the 2012 employee threshold policy. Gross profit is the difference between sales revenue and total costs—including raw materials, indirect expenses, and labor costs (e.g., wages, allowances, recruitment, training, as well as health, social, and unemployment insurance). The revised financial penalty tied to the 2012 employee threshold policy accounts for a larger share of sales and gross profit in the control group than in the treatment group across both optimal windows. However, for labor costs, the control group's burden is larger in the asymmetric window while the treatment group's burden is larger in the symmetric window. Overall, this suggests that firms in the control group generally faced a substantially higher financial burden for non-compliance under the 2012 employee threshold policy.

Table C.6: P-values and FDR-Adjusted Q-values for the Impact of the Threshold-Based Labor Policy on All Outcomes

	[7,12]			[1,25]		
Outcome	P-values (1)	BH Q-values (2)	BKY Q-values (3)	P-values (4)	BH Q-values (5)	BKY Q-values (6)
Labor Force Size						
Total workers	0.3718	—	—	0.1723	—	—
Flexibility-Hours Employment						
Part-time workers	0.1766	0.3533	0.3535	0.2844	0.2844	0.2845
Casual workers	0.4102	0.4102	0.4105	0.2212	0.2844	0.2845
Unregulated Full-time Employment						
Informal, paid full-time workers	0.6082	0.6082	0.6085	0.6673	0.6673	0.3340
Unpaid full-time workers	0.2445	0.4891	0.4895	0.0939	0.1878	0.1880
Normalized Financial Returns						
Sales revenue per full-time worker	0.6703	0.6703	0.6705	0.2491	0.2491	0.1250
Gross profit per full-time worker	0.4635	0.6703	0.6705	0.0395	0.0791	0.0795
Normalized Labor Costs						
Labor costs per full-time worker	0.5455	—	—	0.2703	—	—
Productivity						
Labor productivity	0.7500	0.8723	0.8725	0.0761	0.1522	0.1525
Capital productivity	0.8723	0.8723	0.8725	0.3831	0.3831	0.1920
Full Formality						
Fully formal firm	0.7643	—	—	0.9101	—	—
Deviations from Full Formality						
Formal firm but operates informally	0.0235	0.0704	0.0705	0.0397	0.1192	0.1195
Informal firm but operates formally	0.3814	0.5720	0.3815	0.1624	0.2437	0.1625
Fully informal firm	0.9632	0.9632	0.6425	0.8338	0.8338	0.3225
Observations		126			374	

Notes: Columns 1 and 4 report the raw p-values for $\hat{\delta}_0$ in Equation 6, estimated in separate regressions for each outcome using the optimal symmetric and asymmetric windows of [7, 12] and [1, 25], respectively. P-values correspond to the results in Appendix Tables C.2, C.3, and C.4. Columns 2 and 5 report false discovery rate (FDR)-adjusted q-values using the Benjamini-Hochberg (BH) linear step-up procedure (Benjamini & Hochberg, 1995) and implemented using the Simes method (Simes, 1986). Columns 3 and 6 report "sharpened" q-values from the two-stage adaptive BH procedure of Benjamini et al. (2006) (i.e., BKY). The BH and BKY q-values are computed within each of the eight outcome families listed in the table. In some cases the BH and BKY q-values are nearly identical. When there are few very small p-values across outcomes within a family, the BKY "sharpening" provides little refinement; hence its q-values coincide with the BH q-values. In singleton outcome families, the multiple-testing adjustment is trivial since q-value = p-value so only the p-values are reported.

D Diagnostics for the Validity of Difference-in-Differences

This appendix presents diagnostic analyses that evaluate the suitability of a standard difference-in-differences (DiD) approach for identifying the causal effects of the 2012 employee threshold policy. I first examine the sensitivity of global DiD estimates to sample composition around the 10-formal, paid (FP) employee threshold. I then analyze pre-policy dynamics near the cutoff using event-study evidence to assess local continuity. The findings suggest that a global DiD approach is ill-suited for this application and motivate the use of a difference-in-discontinuities (DiDisc) design in Section 5.

D.1 Sensitivity of Global DiD Estimates to Sample Composition

As a first diagnostic, I evaluate how a standard global DiD estimator performs under alternative sample restrictions around the 10-FP employee threshold. This exercise is motivated by concerns about the interpretation of the parallel trends assumption when treatment is defined by a size-based threshold. In this setting, treatment assignment is endogenously determined by a discrete running variable (i.e., the firm's number of FP full-time workers in the pre-policy period) and firms on either side of the cutoff differ systematically in regulatory incentives and outcome-relevant characteristics. As a result, including firms far from the threshold can cause global DiD estimates to conflate policy effects with underlying size-related heterogeneity, even in the absence of differential pre-trends.

To illustrate this issue, I estimate a two-way fixed effects (TWFE) DiD specification that compares firms above and below the 10-FP employee threshold before and after the implementation of the 2012 employee threshold policy. Equation D.1 defines this regression specification.

$$Y_{it} = \beta_0(Above10_i \times Post_t) + \lambda_i + \gamma_t + \epsilon_{it} \tag{D.1}$$

In Equation D.1, Y_{it} is an outcome for firm i at time t. $Above10_i$ equals 1 if firm i has at least 10 FP full-time workers pre-policy and 0 otherwise. $Post_t$ equals 1 in the post-policy period (i.e., year-end 2014) and 0 in the pre-policy period (i.e., year-end 2010). The terms λ_i and γ_t represent firm and year fixed effects, respectively, and ϵ_{it} is the error term. The coefficient of interest β_0 captures the average differential change in outcomes for firms above the threshold relative to those below it post-policy. Standard errors are clustered at the firm-level. This model specification intentionally omits the running variable, reflecting the global comparison inherent in a DiD design.

Equation D.1 is estimated for each outcome described in Section 5.2 using a sequence of nested samples defined by firms' pre-policy FP full-time employment. The broadest window is the analytical sample (see Appendix E) which includes firms with at most 321 FP full-time workers pre-policy and reflects the global DiD comparison. The second window excludes firms with zero FP full-time workers, thereby removing the dominant mass point in the running variable depicted in Figure F.1 and mitigating concerns that estimates could be driven by comparisons between employing firms and owner-only firms. The third window further restricts the sample to firms with less than 50 FP full-time workers, approximating the World Bank's definition of micro- and small- enterprises based on total employment.[34] The fourth window is the optimal asymmetric window of $[1, 25]$, identified using the local randomization procedure in Section 5.3. This window maximizes sample size while maintaining pre-policy covariate balance around the cutoff. The fifth window imposes a symmetric restriction of ± 5 FP full-time workers around the threshold, providing a transparent benchmark that treats firms equidistant from the cutoff symmetrically. The sixth window is the optimal symmetric window of $[7, 12]$, the widest symmetric window that satisfies covariate balance under the local randomization approach (see Section 5.3).[35] This sequence of windows directly evaluates how global DiD estimates depend on sample composition and the inclusion of firms farther from the threshold. However, restricting the sample to firms closer to the threshold does *not* resolve this design limitation. This is because the DiD design ignores the running variable and therefore cannot account for pre-existing discontinuities or size-related trends at the cutoff.

Figure D.1 plots the estimated $\hat{\beta}_0$ coefficients for all outcomes analyzed in the DiDisc analysis. The estimates differ substantially in magnitude, sign, and statistical significance as the sample is restricted toward the threshold. This instability suggests that global DiD estimates are driven by comparisons with firms far from the threshold—where parallel trends are less plausible—rather than by firms near the threshold. Consequently, the global DiD estimator does not recover a well-defined causal effect as the observed sensitivity indicates that the DiD estimand itself is ill-defined.

Notably, when Equation D.1 is estimated on the same samples that satisfy covariate balance at the cutoff—namely the optimal asymmetric and symmetric windows—only one of the 28 estimated treatment effects is statistically significant and it does not correspond to any outcome identified as significant under the DiDisc design. This occurs despite these being the samples in which the DiDisc design identifies economically meaningful and statistically significant policy effects (see Section 6.2). The contrast indicates that DiD's failure is not attributable to limited statistical power nor sample size, but rather to its inability to condition on the policy-induced discontinuity at the threshold.

D.2 Pre-Policy Continuity at the Threshold

Appendix D.1 demonstrates that conventional TWFE DiD estimates are sensitive to the estimation sample when treatment is defined by a size-based threshold. One potential explanation for this is a violation of the parallel trends assumption. As a second diagnostic, I assess that possibility directly by analyzing whether firms above and below the 10-FP employee threshold exhibit differential outcome trends *prior to* the 2012 employee threshold policy. Using an event-study specification, I test for pre-policy differences in outcomes between firms on either side of the cutoff. Although the DiDisc design employed in Section 5 does not require parallel trends between treatment and control groups, its validity depends on a related but weaker condition: absent the policy change, outcomes must evolve smoothly across the threshold over time. This continuity condition is inherently local and differs from the parallel trends assumption required for the DiD design. To determine whether it is plausible in this setting, I analyze outcome dynamics in periods *preceding* the policy's passage

[34] See https://ieg.worldbankgroup.org/sites/default/files/Data/Evaluation/files/SME_Synthesis.pdf for details.
[35] The number of unique firms in the first through sixth windows is 1,439, 308, 243, 187, 107, and 63, respectively.

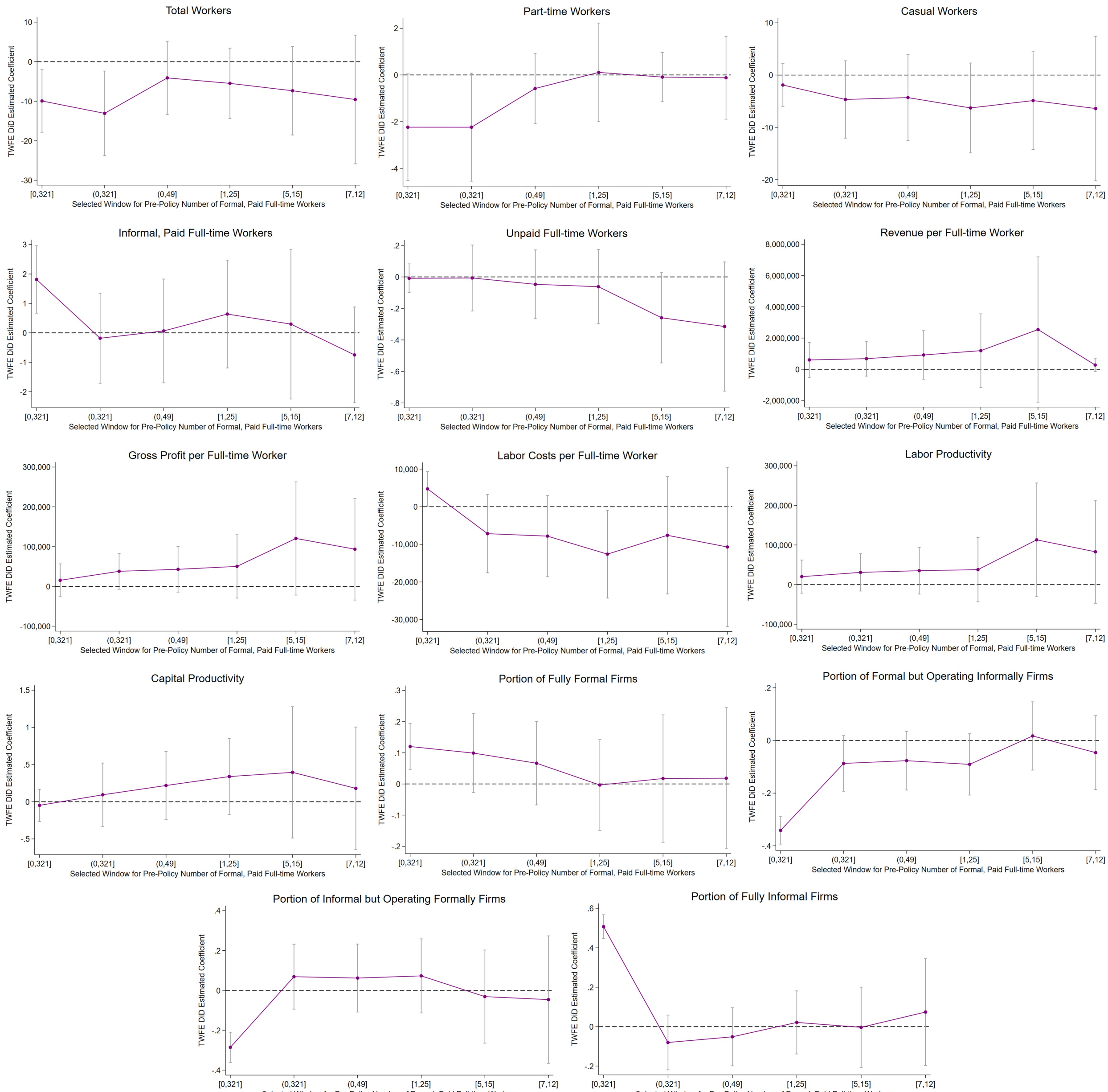


Figure D.1: Sensitivity of TWFE DiD Estimates Around the 10-Formal, Paid Employee Threshold

Notes: This figure plots the estimated two-way fixed effects (TWFE) difference-in-differences (DiD) $\hat{\beta}_0$ coefficient from Equation D.1 for all the outcomes analyzed in the difference-in-discontinuities (DiDisc) analysis (see Section 5). Each point reports $\hat{\beta}_0$ on progressively narrower samples defined by a firm's number of formal, paid (FP) full-time workers pre-policy. Whiskers denote 95% confidence intervals. Estimates vary substantially across selected windows, indicating that global DiD results are sensitive to sample composition around the threshold. The maximum number of FP full-time workers pre-policy in the analytical sample (see Appendix E) is 321. All size-normalized firm economic account metrics (see Section 5.2) are reported in thousands of Vietnamese Dong (except for capital productivity).

and implementation. Given the structure of the UNU-WIDER Vietnam SME database (see Section 3) and the policy's timeline (see Figure 1), I focus on year-end 2008 and 2010 as pre-policy periods.

To examine local pre-trends, I estimate the following event-study regression separately for each outcome variable:

$$Y_{it} = \sum_{k \neq 2010} \beta_k \left(Above10_i \times \mathbb{1}\{t = k\} \right) + \lambda_i + \gamma_t + \epsilon_{it} \tag{D.2}$$

In Equation D.2, Y_{it} is an outcome for firm i at time t. $Above10_i$ equals 1 if firm i has more than 10 FP full-time workers pre-policy and 0 otherwise, and $k \in \{2008, 2010, 2012, 2014\}$ denotes year-end periods. The terms λ_i and γ_t capture firm and year fixed effects, respectively, and ϵ_{it} is the error term. The 2010 year-end period is omitted and serves as the reference category. Standard errors are clustered at the firm-level.

Although the regression includes observations from multiple year-end periods, inference regarding pre-policy dynamics focuses only on the coefficient β_{2008}, which captures differences in outcomes between firms above and below the threshold in year-end 2008 relative to 2010. Observations from year-end 2008 and 2012 are included only for firms within the analytical sample (see Appendix E), which are observed in both the pre-policy (i.e., year-end 2010) and post-policy (i.e., year-end 2014) periods.[36] This restriction ensures that resulting comparisons reflect within-firm dynamics rather than changes in sample composition. Estimation is conducted using three samples: the analytical sample, the optimal asymmetric window of $[1, 25]$, and the optimal symmetric window of $[7, 12]$.[37]

Figure D.2 plots the estimated coefficient $\hat{\beta}_{2008}$ for each outcome variable and sample definition. Across all outcomes and samples, the estimated pre-policy differences are generally small in magnitude and statistically indistinguishable from zero. Moreover, these estimates become more tightly centered around zero as the sample is restricted from the analytical sample to the optimal windows. This pattern is consistent with the interpretation that firms closer to the cutoff are more comparable along both observed and unobserved dimensions. Accordingly, there is no systematic evidence of pre-policy discontinuities or differential trends at the threshold across the optimal windows.

Importantly, these findings do *not* validate a global DiD design. Rather, they highlight a crucial limitation of standard pre-trends diagnostics in threshold-based settings. Though firms just above and below the cutoff exhibit similar pre-policy outcome dynamics, Appendix D.1 shows that global TWFE DiD estimates remain highly sensitive to window selection and sample composition. This instability arises not from violations of the parallel trends assumption, but from how the DiD design averages across heterogeneous firms whose outcomes and incentives differ systematically with size. When treatment is defined by an endogenous firm-size threshold, the DiD estimand itself is not well defined globally. This is because firms on either side of the cutoff do not form a stable counterfactual comparison group even when conventional event-study tests fail to reject parallel trends.

By contrast, while the absence of pre-policy differences is uninformative for validating a global

[36] These periods were selected based on analyses carried out in Section 3.

[37] The optimal asymmetric and symmetric windows are selected using the local randomization procedure described in Section 5.3. The number of unique firms in these three samples is 1,439, 187, and 63, respectively.

DiD design, it is precisely what supports an identification strategy that explicitly conditions on the running variable at the cutoff. The DiDisc design exploits continuity at the threshold and identifies policy effects from changes in the discontinuity over time. As a result, the DiDisc design avoids the pitfalls inherent in averaging across heterogeneous and arbitrarily defined comparison groups.

Together, Appendices D.1 and D.2 demonstrate that the limitations of a standard DiD approach in this setting derive from the estimand it targets rather than from an assumption violation. When treatment is induced by a size-based threshold, the policy effect is inherently local because incentives change discontinuously at the cutoff. A global DiD estimator aggregates comparisons across firms that differ systematically in size and regulatory exposure; hence it cannot isolate this margin even when pre-policy dynamics are continuous near the threshold. In contrast, the DiDisc design isolates this policy-relevant margin, yielding a well-defined estimand aligned with the institutional structure of the policy. In turn, the DiD design does not recover the causal effect of interest.

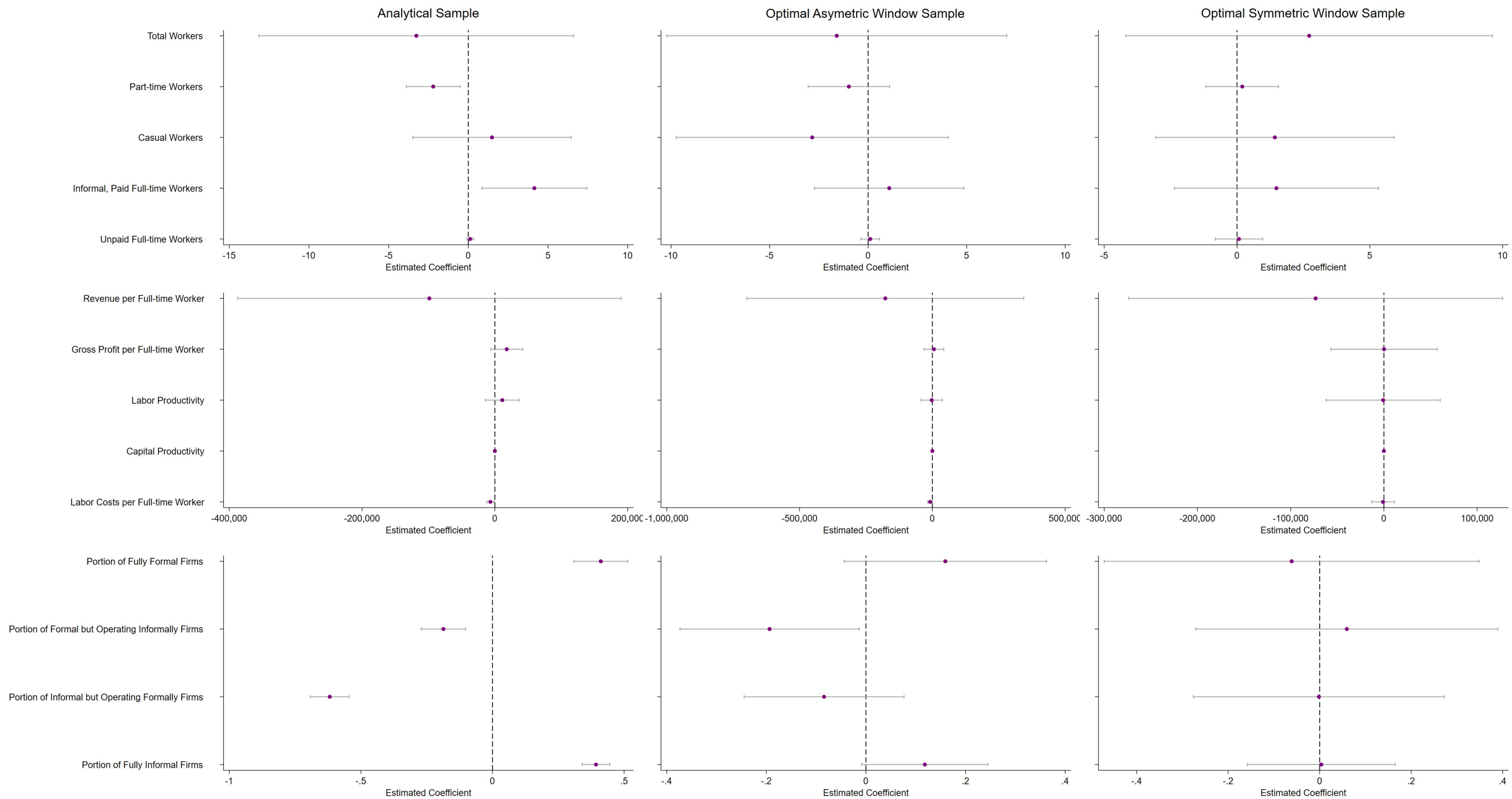


Figure D.2: Estimated Pre-Policy Discontinuities Across Samples

Notes: This figure plots estimated coefficient $\hat{\beta}_{2008}$ from Equation D.2 for all outcomes across three model specifications that correspond to the following samples: the analytical sample (see Appendix E), the optimal asymmetric window, and the optimal symmetric window (see Section 5.3). The coefficient $\hat{\beta}_{2008}$ captures the interaction between an indicator for firms with more than 10 formal, paid (FP) full-time workers in the pre-policy period (i.e., year-end 2010) and year-end 2008, relative to the omitted base of year-end 2010. Whiskers denote 95% confidence intervals. Estimates are predominantly small and statistically indistinguishable from zero. All size-normalized firm economic account metrics (see Section 5.2) are reported in thousands of Vietnamese Dong (except for captial productivity).

E Rounding Procedure for Analytical Sample

This appendix details how the analytical sample (see Sections 5.3 and 6.1) is constructed from the 2011 and 2015 waves of the UNU-WIDER Vietnam SME database, which correspond to the pre-policy (i.e., year-end 2010) and post-policy (i.e., year-end 2014) periods, respectively. Because several labor variables in these surveys are recorded as proportions, converting them into headcounts can generate fractional worker counts whereas headcounts must be integers. Thus I apply integrity checks and a bounded rounding procedure to the raw samples, removing firm-years that fail these checks. The resulting cleaned worker counts form the analytical sample referenced in this paper.

To provide context for these transformations, survey documentation and dataset variable labels indicate that the "percentage of the regular full-time labor force [that] has a formal (written down) labor contract" pertains to *full-time* workers only. Let FT_{it} be the number of full-time workers and $C_{it} \in [0, 1]$ denote the percentage of full-time workers with a formal, written labor contract for firm i at time t (see the same definitions in Appendix A). The implied number of formally-contracted full-time workers (i.e., $FT_{it} \times C_{it}$) should be an integer. To preserve the meaning of C_{it}, I do *not* round at this stage. Instead, I drop firm-years where this product is non-integer, which I treat as data entry errors. Table E.1 reports attrition by year-end period from this integrity check.

Referencing the conclusions from Appendix A, let $U_{it} \in [0, 1]$ denote the percentage of full-time workers that are unpaid. I then decompose full-time workers into four mutually exclusive groups:

$$FP_{it} = C_{it} \times FT_{it} \times (1 - U_{it}) \tag{F.1}$$

$$IP_{it} = (1 - C_{it}) \times FT_{it} \times (1 - U_{it}) \tag{F.2}$$

$$FU_{it} = C_{it} \times FT_{it} \times U_{it} \tag{F.3}$$

$$IU_{it} = (1 - C_{it}) \times FT_{it} \times U_{it} \tag{F.4}$$

so that $FP_{it} + IP_{it} + FU_{it} + IU_{it} = FT_{it}$ by construction.[38] Equations F.1-F.4 estimate the number of (i) formal, paid (FP), (ii) informal, paid, (iii) formal, unpaid, and (iv) informal, unpaid full-time workers, respectively, for a given firm and period.[39]

Because C_{it} and U_{it} are proportions, Equations F.1-F.4 may yield fractional full-time worker counts. To conservatively recover integer counts for the four full-time worker types where defensible, I use a bounded rounding procedure that reconciles small numerical discrepancies in three stages:

1. A full-time worker type value is snapped to the nearest integer only when it lies within a very small tolerance of an integer (i.e., ≤ 0.101) and the *other* three full-time worker type values are already integers.

[38]Note that Equation F.1 equals Equation 1 from Section 3.

[39]The sum of FU_{it} and IU_{it} equals the number of unpaid full-time workers, an outcome described in Section 5.2.

2. Full-time worker type values are rounded up only when the fractional part is clearly above one-half (i.e., ≥ 0.501), rounded down only when it is clearly below one-half, and unchanged when it is exactly one-half (i.e., $= 0.500$).

3. Observations with unreconcilable half-type fractional patterns (e.g., the fractional part of all full-time worker type values are equal to one-half, that for two out of the four worker type values are equal to one-half, etc.) are dropped.

After implementing stages 1 through 3, I again verify that the rounded full-time worker type values sum to the total number of full-time workers. These stages apply to the pre- and post-policy period samples. Table E.1 reports the step-by-step sample changes at each stage by year-end period.

Table E.1: Integrity Checks & Rounding Procedure Applied to Construct the Analytical Sample

Description	Pre-Policy Period (2010)	Post-Policy Period (2014)
Baseline sample	2,512	2,648
Integrity checks (pre-rounding)		
Drop if mismatched full-time worker counts across modules	8	—
Drop if missing C_{it}	16	—
Drop if non-integer $FT_{it} \times C_{it}$	122	208
Rounding procedure		
Stage 1: near-integer snapping	126	111
Stage 2: half-rule rounding	293	267
Stage 3: dropping unreconcilable half-type patterns	53	68
Integrity check (post-rounding)		
Drop if $FP_{it} + IP_{it} + FU_{it} + IU_{it} \neq FT_{it}$	1	—
Total rounded (stages 1–2)	419	378
Total dropped (integrity checks + stage 3)	200	276
Rounded sample	2,312	2,372
Merged analytical sample	1,439	

Notes: Entries report how many firms are dropped or rounded at each step when converting the raw survey samples into the analytical sample. Pre-rounding integrity checks drop observations with (i) inconsistent full-time worker counts across separate survey modules, (ii) missing contract shares that could not be reliably imputed using data from the previous survey year, or (iii) non-integer implied full-time worker counts where the product of a headcount and a proportion should be integer-valued. The three-stage rounding procedure is only applied to the four full-time worker types. "Rounded sample" reports the remaining observations by period and "merged analytical sample" is the balanced panel used to select the optimal symmetric and asymmetric windows as described in Section 5.3.

To determine if the above rounding procedure biases sample composition, I first compare firms with *any* rounded full-time worker type values to firms with no rounded values within the optimal symmetric and asymmetric windows. Figure E.1 indicates that the prevalence of rounding within these optimal windows is modest and fairly balanced across periods and on either side of the cutoff, making it unlikely to influence sample composition. For example, Panel A shows that six of the 14 firms with two formal, paid (FP) full-time workers pre-policy had *at least one* rounded full-time worker type. Panel B provides the post-policy analogue for these same firms; it shows that among the four firms with two FP full-time workers post-policy, only one had *at least one* rounded full-time worker type. Because such cases are limited, they are not expected to influence the research design. To confirm that the rounding procedure does not change pre- or post-policy period assignment, I re-select these periods using the unrounded sample.[40] The patterns in Figure E.2 closely mirror those in Figure 2: the average number of FP and informal, paid full-time workers decreases smoothly while the number of unpaid full-time workers remains flat. Adjacent year-end period comparisons for the mean number of FP full-time workers do not differ significantly and only the decline from year-end 2010 to year-end 2014 is significant.[41] This establishes that designating year-end 2010 and year-end 2014 as the pre- and post-policy periods, respectively, is *not* an artifact of rounding.

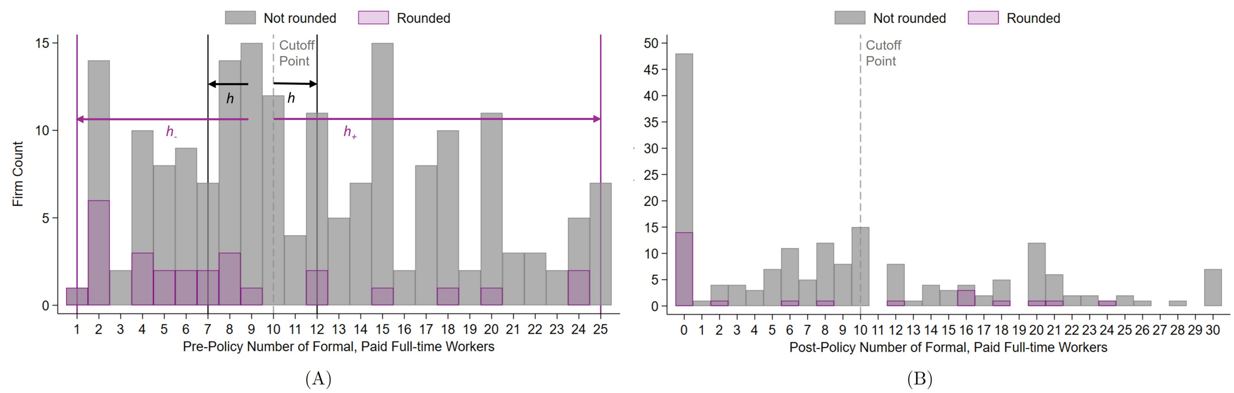


Figure E.1: Distribution of Running Variable (Rounded versus Unrounded) Pre- and Post-Policy within Optimal Windows

Notes: This figure shows the distribution of the running variable disaggregated by whether firms' number of formal, paid full-time workers is affected by the rounding procedures. Panels A and B illustrate the pre- and post-policy periods, respectively. The black and purple vertical lines in Panel A indicate the boundaries of the optimal symmetric and asymmetric windows, respectively (see Section 5.3). Rounded firm values in both periods are apparent but small.

The empirical findings delineated in Section 6 are also stable to excluding any rounding. Figure E.3 shows the distribution of firms by the running variable (i.e., number of FP full-time workers) for both periods leveraging the optimal windows derived from the unrounded sample. Applying the local randomization approach (Cattaneo et al., 2024) to this unrounded sample yields an optimal symmetric window of [8,11] and an optimal asymmetric window of [1,13]. McCrary (2008) density tests reveal no evidence of spurious discontinuity at the cutoff post-policy. Namely, the bunching conclusions drawn from the unrounded sample are consistent with those reported in Section 6.1.[42]

Re-estimating the selected firm outcomes in Table 4 using Equation 6 for the unrounded sample

[40]The unrounded sample excludes firms with *at least one* non-integer value among the four full-time worker types after those values are computed. After performing the pre-rounding integrity checks listed in Table E.1 and removing firms with non-integer worker types, the sample consists of 1,915 firms in the pre-policy period and 2,010 firms in the post-policy period. Restricting to firms observed in both periods yields a balanced, unrounded sample of 1,006 firms.

[41]Statistical details for these year-end comparisons are reported in the notes to Figure E.2.

[42]Statistical details for these McCrary (2008) density tests are reported in the notes to Figure E.3.

(see Table E.2) yields coefficients with similar signs and significance levels. Specifically, the impact of the 2012 employee threshold policy on the number of unpaid full-time workers remains negative and significant for the optimal asymmetric window (i.e., $p = 0.0388$), the portion of formal firms that operate informally remains positive and significant in at least one window (i.e., $p = 0.0550$), and gross profit per full-time worker remains positive and marginally significant under the optimal asymmetric window ($p = 0.1197$). However, labor productivity is no longer significant. These checks indicate that the results in Section 6.2 are robust to excluding the rounding procedure. Table E.3 reports the associated p-values and false-discovery-rate-adjusted q-values for the unrounded sample estimates; the qualitative conclusions are also largely unchanged from Appendix Table C.6.

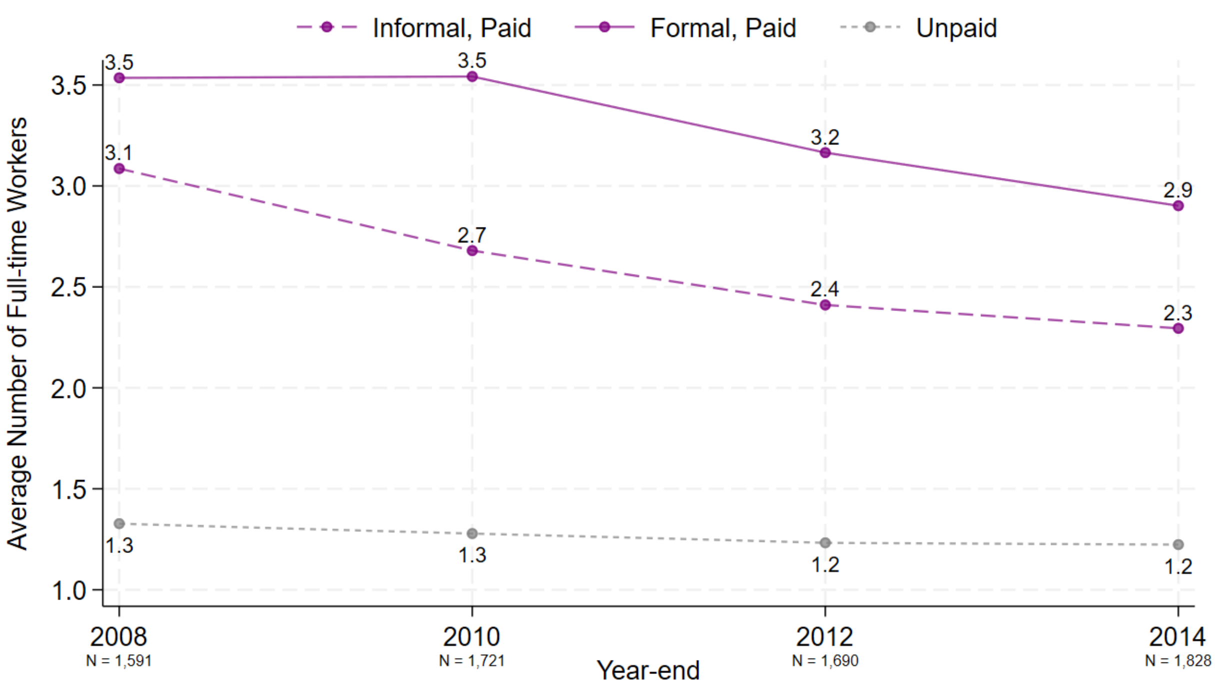
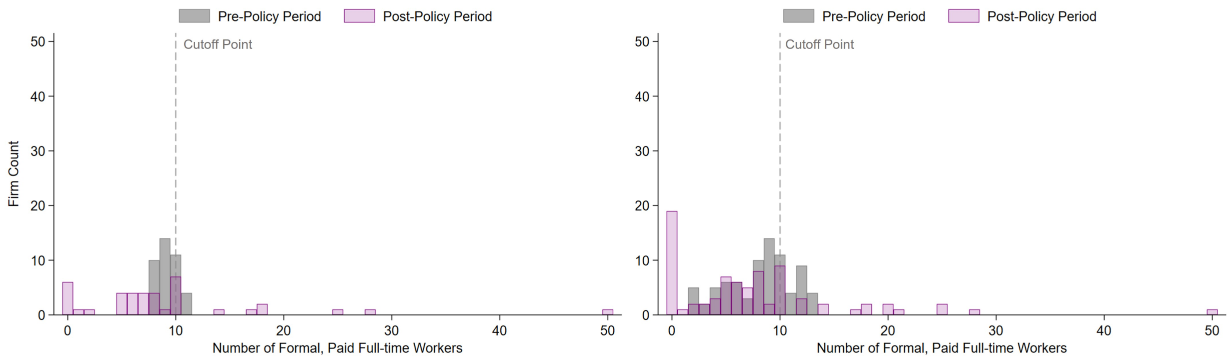


Figure E.2: Mean Number of Worker Types for Restricted, Unrounded Sample by Year-End Period

Notes: This figure shows the average number of (1) informal, paid, (2) formal, paid (FP), and (3) unpaid full-time workers across multiple year-end periods for the unrounded sample. The unrounded sample excludes firms with *at least one* non-integer value among the four full-time worker types. The sample used in this figure is also restricted to firms with fewer than 50 total workers, consistent with the World Bank's definition of micro- and small-enterprises. Each year-end period's sample size is indicated below the corresponding year. Examining FP full-time workers, mean differences are not statistically significant between year-end 2008 and 2010 ($p = 0.9812$) or year-end 2012 and 2014 ($p = 0.2981$). The decline from year-end 2010 to 2012 is marginally significant ($p = 0.1594$) and that from year-end 2010 to 2014 is statistically significant ($p = 0.0133$). These patterns are consistent with those shown in Figure 2.



(A) (B)

Figure E.3: Distribution of Running Variable Pre- and Post-Policy within the Optimal Windows Based on the Unrounded Sample

Notes: This figure shows the distribution of firms around the 10-formal, paid (FP) employee threshold, using optimal symmetric (Panel A) and asymmetric (Panel B) windows derived from the unrounded sample. This sample excludes firms with non-integer values in *any* of the four full-time worker types after estimation. In Panel A, the McCrary (2008) density test detects a significant discontinuity above the cutoff in the pre-policy period ($T = 5.465$, $p = 0.0000$) but no evidence of manipulation post-policy ($T = -1.144$, $p = 0.2526$). Panel B indicates a significant pre-policy drop in the density of firms just above the cutoff relative to just below ($T = -2.955$, $p = 0.0031$) while the post-policy estimate is insignificant ($T = -0.514$, $p = 0.6075$). The post-policy results align with those reported in Section 6.1. Using the *full* unrounded sample—which includes 197 treatment firms and 809 control firms—shows the same pattern: McCrary (2008) density tests yield $T = -1.84$ and $p = 0.066$ pre-policy and $T = -0.54$ and $p = 0.599$ post-policy.

Table E.2: Impact of Threshold-Based Labor Policy on Selected Firm Outcomes Based on the Unrounded Sample

	Unpaid Full-time Workers		Formal Firm But Operates Informally		Gross Profit Per Full-time Worker		Labor Productivity	
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Post_t \times Above10_i$	-0.460	-0.672**	0.191	0.195*	51,494	37,780	-9,022	21,672
	(0.394)	(0.320)	(0.134)	(0.100)	(51,263)	(24,012)	(74,841)	(32,179)
Group mean	0.333	0.412	0.000	0.039	60,335	60,449	97,705	94,440
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.692	0.680	0.536	0.659	0.591	0.539	0.565	0.546
Optimal window	[8,11]	[1,13]	[8,11]	[1,13]	[8,11]	[1,13]	[8,11]	[1,13]
Observations	78	158	78	158	78	158	78	158

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the firm-level."Group mean" reports the pre-policy mean of the outcome variable for the control group (i.e., firms just below the threshold) in each model specification. Results only underscore outcomes listed in Table 4. P-values and FDR-adjusted q-values for these estimates are listed in Appendix Table E.3. The unrounded sample excludes firms with non-integer values in *any* of the four full-time worker types after estimation.

Table E.3: P-values and FDR-Adjusted Q-values for the Impact of the Threshold-Based Labor Policy on All Outcomes Based on the Unrounded Sample

	[8,11]			[1,13]		
Outcome	P-values (1)	BH Q-values (2)	BKY Q-values (3)	P-values (4)	BH Q-values (5)	BKY Q-values (6)
Labor Force Size						
Total workers	0.3180	—	—	0.3311	—	—
Flexibility-Hours Employment						
Part-time workers	0.1788	0.3497	0.3500	0.5062	0.5062	0.5065
Casual workers	0.3497	0.3497	0.3500	0.3722	0.5062	0.5065
Unregulated Full-time Employment						
Informal, paid full-time workers	0.9010	0.9010	0.9010	0.8161	0.8161	0.4085
Unpaid full-time workers	0.2501	0.5001	0.5005	0.0388	0.0777	0.0780
Normalized Financial Returns						
Sales revenue per full-time worker	0.1225	0.2451	0.2455	0.0142	0.0284	0.0285
Gross profit per full-time worker	0.3215	0.3215	0.3215	0.1197	0.1197	0.0600
Normalized Labor Costs						
Labor costs per full-time worker	0.2521	—	—	0.4084	—	—
Productivity						
Labor productivity	0.9047	0.9047	0.5410	0.5026	0.5195	0.5200
Capital productivity	0.1755	0.3511	0.3515	0.5195	0.5195	0.5200
Full Formality						
Fully formal firm	0.5918	—	—	0.8900	—	—
Deviations from Full Formality						
Formal firm but operates informally	0.1615	0.3768	0.3770	0.0550	0.1649	0.1650
Informal firm but operates formally	0.3768	0.3768	0.3770	0.5257	0.7886	0.5260
Fully informal firm	0.2954	0.3768	0.3770	0.8827	0.8827	0.5885
Observations	78			158		

Notes: Columns 1 and 4 report the raw p-values for $\hat{\delta}_0$ in Equation 6, estimated in separate regressions for each outcome using the optimal symmetric and asymmetric windows of [8,11] and [1,13], respectively, derived from the unrounded sample. The unrounded sample excludes firms with non-integer values in *any* of the four full-time worker types after estimation. P-values correspond to the results in Table E.2. See Table C.6 notes for details about the false discovery rate (FDR)-adjusted q-values reported in Columns 2, 3, 5, and 6.

F Heterogeneity by Pre-Policy Firm Formality Status

The average causal effects presented in Section 6.2 potentially mask important differences in how firms with distinct *pre-policy* formality statuses responded to the 2012 employee threshold policy. To analyze this heterogeneity, this appendix focuses on two firm subgroups that are well represented on both sides of the cutoff within the optimal windows identified in Section 5.3: fully formal firms and those that were informal but operating formally. To reiterate, fully formal firms were already both registration and labor contract compliant prior to the implementation of the policy whereas informal but operating formally firms provided all their full-time workers with formal, written labor contracts but did not register with the government. Since each group occupies a different position along the formality spectrum, their incentives under the policy differ. Examining these subgroups separately determines if the average effects identified earlier are driven by already-compliant firms or also extend to partially-compliant firms. This helps clarify the mechanisms underlying the policy's impact and identifies which firm *types* are most responsive to such threshold-based labor policies.

Figure F.1 shows the distribution of firms around the cutoff by their pre-policy formality status within the optimal symmetric and asymmetric windows. Firms that were fully formal and informal but operating formally appear in sizable and balanced numbers on both sides of the 10-FP employee threshold, making them credible candidates for subgroup DiDisc analysis. This balance helps ensure that any observed differences in outcomes reflect genuine heterogeneity in pre-policy firm compliance behavior rather than compositional shifts.

Table F.1 reports the subgroup heterogeneity estimates obtained from Equation 6. Among fully formal firms, the estimates reveal significant adjustments in labor composition and performance outcomes. Specifically, Columns 1 and 2 of Panel A show that firms just *above* the 10-FP employee threshold had between 1.1 to 1.8 fewer unpaid full-time workers relative to firms just *below* it post-policy. This decline implies that fully formal firms just *below* the threshold—facing a higher cost of informality—substituted toward unpaid labor arrangements, consistent with Theorem 2. Column 4 shows that fully formal firms just above the threshold were 33 percentage points more likely to be "formal but operating informally" than firms below it post-policy, as predicted by Theorem 3. This suggests that some firms strategically relaxed compliance along one dimension by reducing formal labor contract coverage while still remaining registered. Fully formal firms just above the threshold also exhibited gross profit per full-time worker and labor productivity gains of 154 million VND (about 5,800 USD) and 158 million VND (about 6,000 USD), respectively. Though only significant under the optimal symmetric window, these patterns suggest efficiency improvements rather than distortions. Thus, in response to the 2012 employee threshold policy, fully formal firms substituted away from unregulated labor and reshaped their compliance margins by maintaining government registration while selectively adjusting their broader labor strategies.

By contrast, informal firms operating formally pre-policy display a distinct adjustment pattern. Column 3 of Panel B shows that for this subgroup, the probability of being "formal but operating informally" increases by 38 percentage points for firms just above the threshold compared to those just below it post-policy. This indicates that the policy's significant effect on this outcome is not

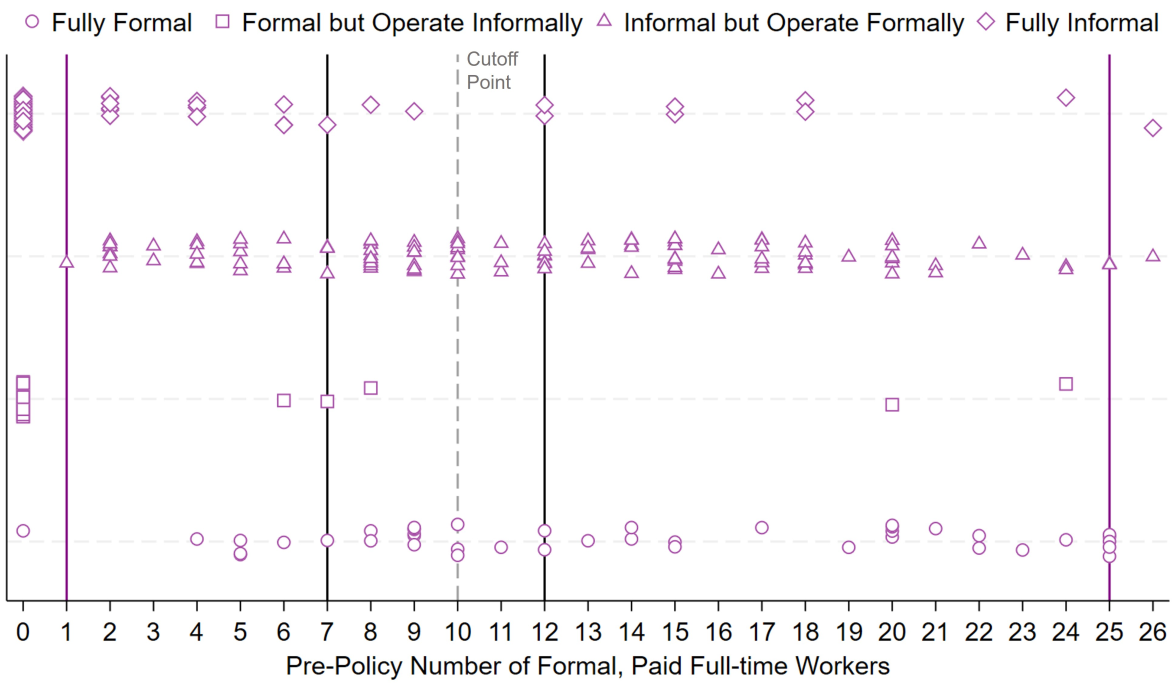


Figure F.1: Distribution of Firms within Optimal Windows by Pre-Policy Formality Status

Notes: This figure shows the distribution of firms within the optimal symmetric and asymmetric windows identified in Section 5.3 by their pre-policy formality status. Each formality status is represented by a different marker shape. The black and purple vertical lines denote the boundaries of the optimal symmetric and asymmetric windows, respectively.

confined to fully formal firms but also extends to firms that, prior to the 2012 employee threshold policy, were labor contract compliant but not registered. When pushed beyond the threshold, these firms responded by registering with the government while adopting hybrid labor strategies rather than providing universal labor contracts. This firm behavior is predicted by Theorem 3: selective or partial compliance with multi-dimensional formality requirements. Instead of absorbing the costs of providing formal, written contracts to all their full-time workers, firms balanced the fixed cost of government registration against newfound reliance on informal labor margins.

As a robustness check, I estimate subgroup-specific optimal symmetric and asymmetric windows for each pre-policy formality status. Table F.2 shows that the patterns observed in Table F.1 remain unchanged under these subgroup-specific optimal windows.

These findings suggest that the 2012 employee threshold policy amplified pre-existing differences in firms' positions along the formality spectrum rather than producing a uniform shift toward full formality. Fully formal firms converted their compliance into efficiency gains while firms that were informal but operating formally engaged in selective compliance and hybrid labor practices. This heterogeneity highlights that the impact of threshold-based labor policies depends more so on firms' initial formality status than the regulation itself, a nuance that average treatment effects obscure.

Table F.1: Impact of Threshold-Based Labor Policy on Selected Firm Outcomes by Pre-Policy Formality Status: Balanced Subgroups within Optimal Windows

	Unpaid Full-time Workers		Formal Firm But Operates Informally		Gross Profit Per Full-time Worker		Labor Productivity	
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(A) Fully Formal Firms								
$Post_t \times Above10_i$	-1.803**	-1.143*	—[a]	0.333**	153,992**	78,014	157,536**	79,218
	(0.835)	(0.566)	—	(0.162)	(62,858)	(72,612)	(63,069)	(73,986)
Group mean	0.222	0.214	—	0.000	79,002	101,043	120,268	139,814
R^2	0.741	0.601	—	0.574	0.742	0.532	0.772	0.537
Observations	30	82	30	82	30	82	30	82
(B) Informal Firms Operating Formally								
$Post_t \times Above10_i$	0.134	—[b]	0.378*	—[b]	5,954	—[b]	-40,233	—[b]
	(0.375)	—	(0.186)	—	(96,335)	—	(118,761)	—
Group mean	0.545	—	0.000	—	68,892	—	103,402	—
R^2	0.812	—	0.617	—	0.568	—	0.548	—
Observations	82	236	82	236	82	236	82	236
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Optimal window	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]	[7,12]	[1,25]

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the firm-level. "Group mean" reports the pre-policy mean of the outcome variable for the control group (i.e., firms just below the threshold) in each model specification for each subgroup. Fully formal firms are registered with the government and provide formal, written contracts to all full-time workers. Informal firms that operate formally are also labor contract compliant but are *not* registered. Results come from subgroup analyses by firms' pre-policy formality status *within* the optimal symmetric and asymmetric windows identified in Section 5.3. Estimates are robust to using subgroup-specific optimal symmetric and asymmetric windows for *each* pre-policy firm formality status (see Table F.2). No estimate is reported for Column 3 of Panel A because the dependent variable does not vary across firms or periods.[a] Results for informal but operating formally firms in the optimal asymmetric window (Panel B) are omitted due to covariate imbalance: firm owners or managers just *above* the threshold were significantly more likely to have completed upper secondary school than those just below it ($p = 0.054$).[b] This table restricts attention to the selected firm outcomes from Table 4 for the formality statuses listed here. Coefficients for other outcomes—even when significant—are not shown. For formality statuses not listed, none of the selected firm outcomes are significant so estimates are omitted.

Table F.2: Impact of Threshold-Based Labor Policy on Selected Firm Outcomes by Pre-Policy Formality Status: Subgroup-Specific Optimal Windows

	Unpaid Full-time Workers		Formal Firm But Operates Informally		Gross Profit Per Full-time Worker		Labor Productivity	
Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(A) Fully Formal Firms								
$Post_t \times Above10_i$	-2.015**	-1.665**	—[a]	-0.035	117,359**	91,496**	127,960**	119,104**
	(0.879)	(0.705)	—	(0.087)	(55,011)	(41,790)	(55,103)	(45,366)
Group mean	0.200	0.200	—	0.000	71,352	94,307	110,891	130,493
R^2	0.659	0.626	—	0.667	0.720	0.572	0.779	0.579
Optimal window	[6,13]	[0,16]	[6,13]	[0,16]	[6,13]	[0,16]	[6,13]	[0,16]
Observations	34	52	34	52	34	52	34	52
(B) Informal Firms Operating Formally								
$Post_t \times Above10_i$	0.134	-0.241	0.378*	0.121	5,954	27,056	-40,233	-907
	(0.375)	(0.276)	(0.186)	(0.124)	(96,335)	(35,166)	(118,761)	(43,826)
Group mean	0.545	0.830	0.000	0.000	68,892	49,987	103,402	79,557
R^2	0.812	0.753	0.617	0.645	0.568	0.572	0.548	0.573
Optimal window	[7,12]	[0,14]	[7,12]	[0,14]	[7,12]	[0,14]	[7,12]	[0,14]
Observations	82	150	82	150	82	150	82	150
Firm fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors (in parentheses) are clustered at the firm-level. "Group mean" reports the pre-policy mean of the outcome variable for the control group (i.e., firms just below the threshold) in each model specification for each subgroup. Fully formal firms are registered with the government and provide formal, written contracts to all full-time workers. Informal firms that operate formally are also labor contract compliant but are *not* registered. Results use subgroup-specific optimal symmetric and asymmetric windows for *each* pre-policy firm formality status. No estimate is reported for Column 3 of Panel A because the dependent variable exhibits minimal change across firms and periods, which allows coefficients to be calculated but leaves too little information across clusters to compute standard errors or test statistics.[a]