Causal Evaluation of a Small Business Training Program: Evidence from Regression Discontinuity and Difference-in-Differences

Nicholas Wong
Massachusetts Institute of Technology
nicwjh@mit.edu

Abstract

We evaluate the causal impact of a manager training program targeting small businesses using administrative panel data covering 100 firms from 2010 to 2019. The program, rolled out on January 1, 2013, was available to firms with 100 or fewer employees. We employ two complementary identification strategies: regression discontinuity exploiting the sharp eligibility cutoff, and difference-in-differences leveraging panel variation and simultaneous adoption timing. Both methods indicate large, positive, and persistent effects on firm sales and productivity. Our primary estimate from difference-in-differences shows that eligibility increased real sales by 0.982 log points (approximately 167%), with corresponding productivity gains of $3,356 per employee, while employment remained unchanged. Effects emerged immediately upon adoption and persisted throughout the seven-year observation window. Regression discontinuity estimates confirm positive local treatment effects at the cutoff, though with wider confidence intervals due to smaller sample size. Results are robust to alternative specifications, bandwidth choices, and outcome measures. The findings suggest that managerial training can generate substantial and durable improvements in firm performance through efficiency gains rather than scale expansion.

1 Introduction

Managerial capital is increasingly recognized as a critical determinant of firm productivity and growth. Governments worldwide invest in training programs designed to improve management practices at small firms, yet rigorous causal evidence on their effectiveness remains limited. On January 1, 2013, a manager training program was introduced for firms with 100 or fewer employees. Eligible firms could choose whether to enroll, while larger firms were ineligible. This policy design creates natural variation in treatment assignment that can be exploited for causal inference.

We address the central research question: what is the causal effect of program eligibility on business outcomes? Identifying causal effects is challenging due to selection bias. Firms that choose to participate may differ systematically from non-participants in unobserved ways that also affect performance. We overcome this challenge using two complementary quasi-experimental strategies. First, we implement a regression discontinuity (RD) design exploiting the sharp 100-employee eligibility threshold. Second, we employ a difference-in-differences (DiD) approach leveraging the panel structure of administrative data and the simultaneous timing of program introduction across all eligible firms.

Our primary estimate from the DiD specification indicates that program eligibility increased firm sales by 0.982 log points (standard error 0.110, $p < 0.001$), corresponding to approximately 167%

higher sales. This effect is accompanied by substantial productivity gains ($3,356 per employee) with no significant change in employment, suggesting the program improved operational efficiency rather than inducing scale expansion. The RD design confirms large positive local treatment effects at the cutoff (1.620 log points, $p = 0.015$), though estimates are less precise due to smaller sample size near the threshold. Both methods pass key validity tests and results are robust across multiple specifications.

2 Data

2.1 Data Sources and Panel Structure

We construct a firm-month panel from three administrative datasets. The first, `firm_information.csv`, contains time-invariant firm characteristics for 100 firms including unique identifiers, names, sector classifications, and a treatment group indicator. This file serves as the master reference for all merges and group definitions. The second, `aggregate_firm_sales.csv`, reports monthly sales by firm from 2010 to 2019, with each row representing a firm-month observation. The third source comprises 120 monthly auxiliary files (e.g., `2010_1.csv`) recording employment, wage bills, revenue, and program adoption indicators for each firm-month.

Our merge procedure prioritizes data integrity and transparency. We first construct a balanced skeleton dataset containing all possible firm-month combinations (100 firms × 120 months), ensuring no observations are inadvertently dropped during merges. We then sequentially left-join sales data, firm characteristics, and stacked auxiliary files onto this skeleton using firm identifiers and month indicators as keys. All merges employ defensive validation checks: we verify one-to-one or many-to-one cardinality, confirm no unmatched observations remain in transaction datasets, and assert that final panel dimensions match expected values (12,000 rows). This systematic approach guarantees complete temporal coverage for all firms and enables immediate detection of any data inconsistencies or structural breaks.

Merging these sources yields a balanced panel of 100 firms observed over 120 months (January 2010 to December 2019), totaling 12,000 firm-month observations. The program became available on January 1, 2013, creating a natural pre-period (36 months) and post-period (84 months). Eligibility was determined by employment in January 2013: firms with 100 or fewer employees could enroll, while larger firms could not.

2.2 Data Quality Issues and Resolution

Administrative data often contain reporting errors. We identified and addressed two main issues through systematic auditing and validation procedures.

Issue 1: Date parsing errors. Monthly auxiliary files exhibited date format inconsistencies. In randomly sampled files, 2 to 4% of rows encoded dates as YYYY-DD-MM instead of YYYY-MM-DD, causing parsing failures. We implemented a swap-parse repair policy: when standard parsing failed, we attempted YYYY-DD-MM format and validated the result against the filename's expected month. Rows where the repaired date matched the filename month were retained; others were dropped. This procedure corrected all date errors without manual intervention.

Issue 2: Firm ID inconsistencies. During merge validation, we discovered that some sales observations failed to match firms in the master information file due to malformed identifiers. Visual inspection revealed a systematic pattern: some firm IDs contained extraneous trailing zeros (e.g., ABCD-1200 instead of the canonical ABCD-12 format). To resolve this without manual intervention, we implemented a three-step correction procedure. First, we applied regex pattern matching to identify all malformed four-digit IDs in the sales data. Second, we constructed a deterministic mapping by programmatically stripping trailing zeros and validated that this mapping was one-to-one (no collisions) and complete (all malformed IDs mapped to valid firm identifiers in the master file). Third, we standardized all identifiers across datasets by converting to uppercase and removing whitespace before applying the validated mapping. Post-correction diagnostics confirmed 100% merge success with no orphaned records, eliminating potential data loss in the merge.

After these corrections, no duplicate firm-month keys or pattern violations remained. All 12,000 observations fall within the expected date range, confirming internal consistency of the merged dataset.

2.3 CPI Deflation

We deflate all nominal monetary variables using the CPIAUCSL series (Consumer Price Index for All Urban Consumers, seasonally adjusted) from the Federal Reserve Bank of St. Louis. Deflation is necessary to separate real productivity changes from nominal price-level effects: failing to adjust for inflation would spuriously attribute 18.9% growth to all firms over the sample period, obscuring true treatment effects. The raw CPI in January 2010 is 217.49 (1982-84 base), which we normalize to 100 as our deflation base. The December 2019 CPI is 258.56 (index 118.9), indicating cumulative inflation of 18.9% over the sample period.

The deflator, defined as $\text{deflator}_t = \text{CPI}_{2010M1}/\text{CPI}_t$, merged successfully for all 12,000 observations with no missing entries. We constructed real variables in constant January 2010 dollars: sales (real), revenue (real), and per-employee measures. Aggregate checks confirm internal consistency: mean nominal sales rose 132.5% from 2010 to 2019, while real sales increased 95.5%, a difference consistent with the 37% cumulative inflation adjustment over the period.

2.4 Outcome Variable Construction

We focus on three primary outcomes to evaluate treatment effects.

Log Sales (Real). The primary outcome is log(1 + sales real), where sales real denotes nominal sales deflated by CPI (January 2010 base). This transformation captures proportional changes in firm output, accommodates zero or near-zero sales observations, and reduces right-skewness in the distribution. Log-transformed sales are standard in productivity literature because coefficients are interpretable as approximate percentage changes and the transformation stabilizes variance across firms of different scales. We prefer sales to revenue because sales represent core operating turnover directly affected by managerial improvements, whereas revenue may include incidental or financial income outside the program's scope.

Sales per Employee (Real). The second primary outcome measures labor productivity, defined as sales real divided by employment. This variable directly indicates output per worker and is particularly relevant for evaluating training interventions aimed at improving operational efficiency. It is straightforward to interpret as average real sales generated per employee.

Log Employment. The secondary outcome is log(1 + employment). This measure helps assess whether treatment effects operate through changes in firm scale (extensive margin) or through improved efficiency (intensive margin). It indicates whether sales growth stems from workforce expansion or from higher output per worker.

For robustness checks, we also construct analogous revenue-based outcomes (log revenue real) following identical procedures. However, sales-based measures are emphasized because they more accurately reflect operational performance directly targeted by the training program.

3 Descriptive Analysis

3.1 Sample Composition and Compliance

The final panel includes 100 firms observed from 2010 to 2019. Based on January 2013 employment, 60 firms are classified as eligible (100 or fewer employees) and 40 as ineligible (more than 100 employees). Among all firms, 46 adopted the training program, corresponding to a compliance rate of 76.7% among eligible firms (46 of 60). One firm, MFUP-80, is identified as a non-complier: it was ineligible by the assignment rule (111 employees in January 2013; average $\approx$200 employees in 2012) but adopted the program. This single defier introduces minor deviation from perfect compliance but does not materially affect group composition.

3.2 Descriptive Statistics by Eligibility Status

Table 1 presents descriptive statistics separately for eligible and ineligible firms in pre-treatment (2010 to 2012) and post-treatment (2013 to 2019) periods. Before the policy, eligible firms were systematically smaller and less productive than ineligible firms. Average real log sales were about one unit lower (approximately 100% less in levels), and real sales per employee were less than two-thirds of those for larger firms, confirming meaningful baseline differences in scale and productivity.

Table 1: Descriptive statistics by eligibility status and period. Values are means with standard deviations in parentheses. Pre-period covers 2010 to 2012 (36 months); post-period covers 2013 to 2019 (84 months).

	Eligible ($\leq$100 emp)		Ineligible ($>$100 emp)	
Variable	Pre	Post	Pre	Post
Log Sales (real)	10.85 (0.91)	11.83 (1.02)	11.85 (0.21)	11.85 (0.37)
Sales per Emp ($)	1,581 (929)	4,993 (4,517)	2,808 (692)	2,864 (1,004)
Employment	73.18 (38.73)	72.04 (36.64)	132.28 (26.39)	131.32 (24.82)
Observations	3,600	8,400	3,600	8,400
Firms	60	60	40	40

From 2010-2012 to 2013-2019, eligible firms experienced substantial growth in both sales and productivity. Average log sales rose from 10.85 to 11.83, and real sales per employee more than tripled from $1,581 to nearly $5,000. In contrast, ineligible firms show minimal change over the same period, with log sales nearly flat and only modest increases in sales per employee.

The descriptive difference-in-differences (change in means across groups) is approximately 0.98 log points for sales and $3,400 for productivity. While not causal, these patterns are consistent with positive treatment effects that motivate our formal analysis.

3.3 Employment Distribution Around Cutoff

Figure 1 plots the distribution of firms by January 2013 employment, the running variable for RD analysis. Panel A displays the full sample histogram (18 to 238 employees) with 20 bins overlaid with a kernel density estimate. Panel B zooms in on 50 to 150 employees, shading eligible (green) and ineligible (tan) regions. A vertical dashed red line marks the eligibility cutoff at 100 employees.

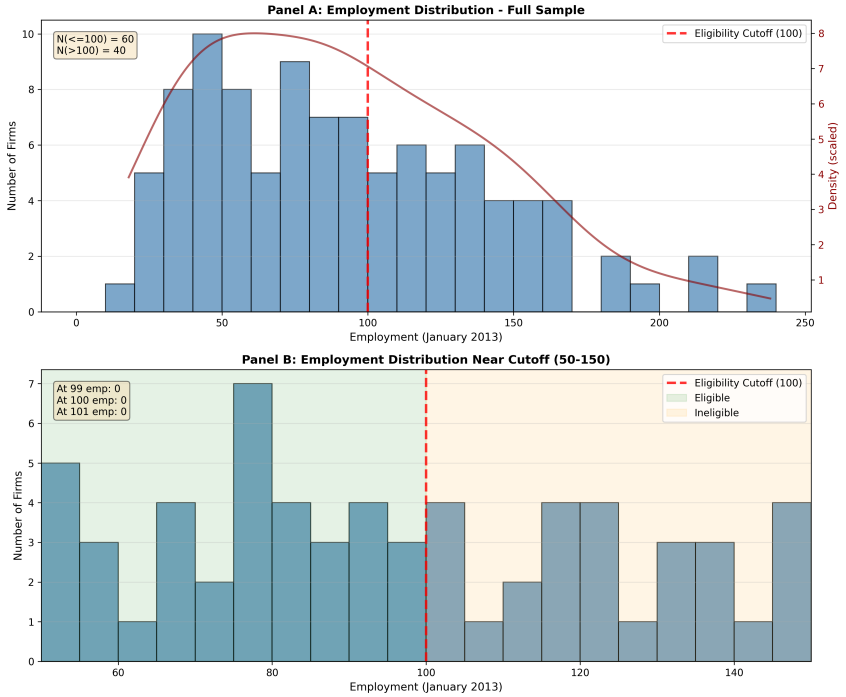

Figure 1: Employment distribution around the eligibility cutoff. Panel A shows the full sample distribution with 60 firms below or at 100 employees and 40 above. Panel B zooms in on the 50-150 range. The smooth distribution with no bunching at the threshold supports the continuity assumption required for regression discontinuity analysis.

The employment distribution is smooth across the 100-employee cutoff with no evidence of bunching or discontinuities. No firms report exactly 99, 100, or 101 employees, and only seven firms lie within the 95 to 105 range. A chi-square test for heaping (multiples of 5 or 10) yields $p = 0.141$, failing to reject uniformity. These results suggest firms did not manipulate employment counts to qualify for eligibility, supporting the continuity assumption required for RD identification.

3.4 Adoption Timing

Figure 2 illustrates program adoption patterns. Panel A shows the treatment timeline: all treated firms switched from control to treated status on January 1, 2013. Panel B displays treatment assignment by eligibility: 45 of 60 eligible firms adopted, while only one ineligible firm (the defier) adopted. Every adoption occurred simultaneously on the same date.

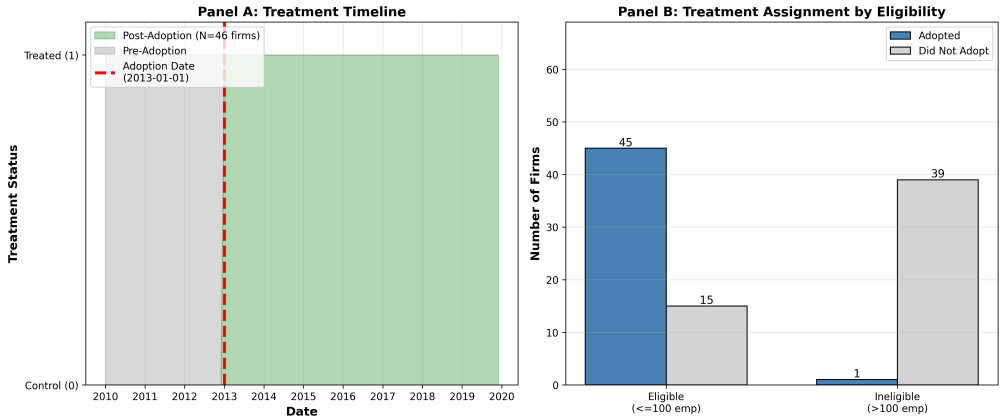

Figure 2: Program adoption patterns. Panel A shows perfectly synchronized adoption on January 1, 2013. Panel B confirms that 45 eligible firms and 1 ineligible firm adopted. The simultaneity creates a clean two-by-two difference-in-differences design.

This perfectly synchronized adoption creates a clean two-by-two DiD design with a single treatment group and single adoption date. Because there is no staggered rollout, the analysis can employ standard two-way fixed effects without concern for treatment effect heterogeneity by timing.

3.5 Outcome Trends Over Time

Figure 3 plots monthly averages of log sales (Panel A) and sales per employee (Panel B) for eligible and ineligible firms from 2010 to 2019. Eligible firms (blue) and ineligible firms (orange) are shown with 95% confidence bands. The vertical dashed red line marks January 2013.

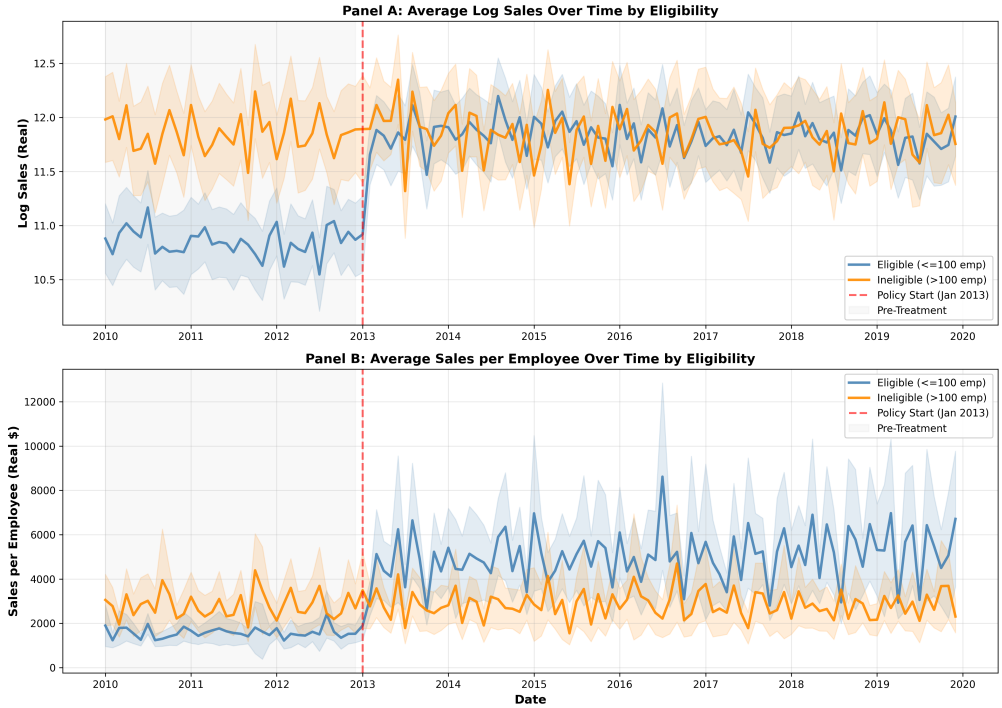

Figure 3: Outcome trends over time by eligibility status. Panel A shows log sales and Panel B shows sales per employee. Both groups exhibit parallel trends before 2013, supporting the parallel trends assumption. After January 2013, eligible firms diverge upward while ineligible firms remain stable.

Before 2013, both groups exhibit relatively stable and parallel trends. Eligible firms maintain average log sales around 10.8 to 11.0, while ineligible firms average 11.8 to 12.2, reflecting size differences but no systematic divergence in growth. Formal pre-trend tests confirm that the interaction between eligibility and time is statistically insignificant for both outcomes ($p = 0.821$ for log sales; $p = 0.936$ for sales per employee), providing strong support for the parallel trends assumption critical for DiD identification.

After January 2013, eligible firms experience a visible upward shift in both outcomes while ineligible firms remain largely unchanged. The post-treatment divergence suggests positive treatment effects consistent with productivity gains among eligible firms.

3.6 Outcome Distributions

Figure 4 compares outcome distributions before and after the program separately for eligible and ineligible firms. Panels A and B present box plots of log sales and sales per employee. Panels C and D show corresponding violin plots revealing full distributional shapes.

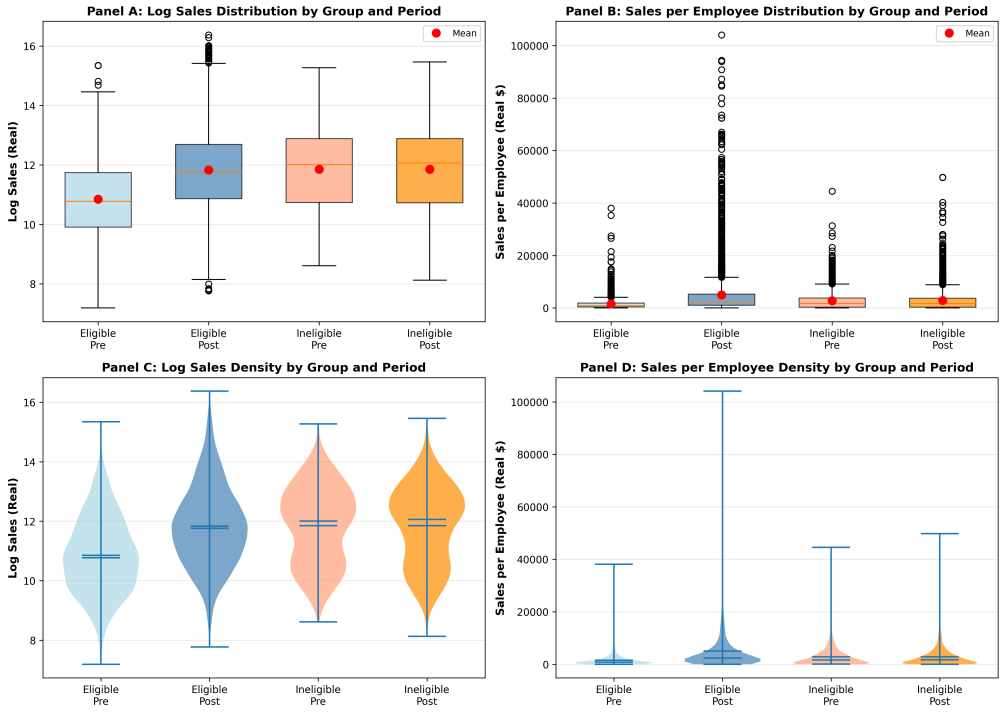

Figure 4: Outcome distributions by group and period. Panels A and B show box plots; Panels C and D show violin plots. Distributions for eligible firms shift markedly rightward post-treatment, while ineligible firm distributions remain stable. The patterns indicate substantial heterogeneous treatment effects concentrated among eligible firms.

Both visual and numerical evidence point to substantial gains among eligible firms after 2013. The eligible group's log sales mean rises from 10.85 to 11.83, while the ineligible group remains flat around 11.85. For sales per employee, eligible firms increase by approximately $3,400 compared with only $56 among ineligible firms. The violin plots confirm rightward shifts and greater spread in the eligible group's post-treatment distribution, consistent with heterogeneous but generally positive treatment effects on firm performance and productivity.

4 Regression Discontinuity Design

4.1 Identification Strategy

The regression discontinuity design exploits the sharp eligibility cutoff at 100 employees. Under the continuity assumption, firms just below and just above the threshold are similar in all relevant characteristics except eligibility. Any discontinuity in outcomes at the cutoff can be attributed to causal effects of program eligibility.

Let Y_i denote the outcome for firm i, X_i denote January 2013 employment (the running variable), and $c = 100$ denote the eligibility cutoff. Since RD requires a cross-sectional dataset with one observation per firm, we collapse the post-treatment panel by computing firm-level averages of residualized outcomes, where residuals are obtained by regressing outcomes on month fixed effects to remove time-varying aggregate shocks. The sharp RD estimand is:

$$\tau_{RD} = \lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x], \tag{1}$$

representing the local average treatment effect for firms at the threshold. We estimate this using local linear regression on either side of the cutoff with triangular kernel weights and MSE-optimal bandwidth selection. The rdrobust package provides bias-corrected robust inference accounting for estimation uncertainty in both the conditional mean functions and the bandwidth choice.

4.2 Validity Tests

Credible RD inference requires three conditions: (1) no manipulation of the running variable, (2) continuity of pre-treatment covariates at the cutoff, and (3) no discontinuities in pre-treatment outcomes. We assess each condition empirically.

McCrary density test. We test whether the density of the running variable is continuous at the cutoff using the McCrary test. The test statistic is $t = -0.262$ with $p = 0.793$, failing to reject the null hypothesis of density continuity. Combined with visual inspection showing smooth distributions and no heaping at multiples of 5 or 10 ($\chi^2 = 6.90$, $p = 0.141$), these results provide strong evidence against manipulation.

Covariate balance. Table 2 presents local balance tests for six pre-treatment covariates using the optimal bandwidth of 19.35 employees (yielding 14 firms below and 11 above the cutoff). None of the covariates exhibit statistically significant discontinuities at the threshold (all $p > 0.05$), supporting the identifying assumption that firms near the cutoff are similar in observable characteristics.

Table 2: Local covariate balance within optimal bandwidth (19.35 employees). None of the pre-treatment covariates differ significantly across the cutoff, supporting the regression discontinuity validity assumption.

Variable	Mean ($<$100)	Mean ($\geq$100)	RD Est	SE	p-value
Log Sales (Pre)	11.70	11.81	-0.059	0.403	0.884
Sales per Emp (Pre)	2,502	2,761	348	1,279	0.785
Employment (Pre)	104.91	126.11	-45.64	37.93	0.229
Log Employment (Pre)	4.57	4.77	-0.370	0.327	0.259
Log Revenue (Pre)	13.25	13.31	0.571	0.833	0.493
Revenue per Emp (Pre)	16,644	19,813	11,766	13,484	0.383

Placebo tests. We test for discontinuities in pre-treatment outcomes at the cutoff. All three placebo estimates are small and statistically insignificant: log sales (-0.057, $p = 0.830$), sales per employee (199.82, $p = 0.792$), and employment (-19.58, $p = 0.491$). The absence of pre-treatment discontinuities reinforces the validity of the RD design.

4.3 Main Results

Table 3 presents primary RD estimates for the three outcomes. We report bias-corrected robust point estimates, standard errors, p-values, 95% confidence intervals, and sample sizes within the optimal bandwidth.

Table 3: Primary regression discontinuity estimates using local linear regressions around the 100-employee cutoff. Treatment effects are sign-corrected to reflect the effect of eligibility (being below the cutoff). Standard errors are bias-corrected and robust. Bandwidth is MSE-optimal and varies by outcome.

Outcome	Treatment Effect	SE	p-value	95% CI	Bandwidth	N (left, right)
Log Sales (Real)	1.620	0.663	0.015	[0.321, 2.920]	19.35	14, 11
Sales per Emp ($)	12,288	5,686	0.031	[1,144, 23,432]	15.60	10, 7
Log Employment	0.110	0.263	0.675	[−0.404, 0.625]	16.83	10, 8

For log sales, the point estimate is 1.620 log points (SE 0.663, $p = 0.015$), corresponding to approximately 405% higher sales for eligible firms at the cutoff. For sales per employee, the estimate is \$12,288 (SE \$5,686, $p = 0.031$), indicating substantial productivity gains. For log employment, the estimate is small and insignificant (0.110, $p = 0.675$), suggesting employment levels did not change materially at the threshold.

Both sales-based outcomes show economically meaningful and statistically significant jumps at the cutoff, while employment remains continuous. This pattern is consistent with genuine productivity improvements rather than scale expansion. However, the RD estimates are substantially larger than corresponding DiD estimates (presented in Section 5), reflecting both the local nature of RD identification (effects for firms right at the cutoff) and considerable sampling variability given the small sample size (N=25).

4.4 Robustness Checks

We assess robustness to alternative specifications. Table 4 presents results for log sales under six variations: baseline optimal bandwidth, bandwidth 0.5h, bandwidth 1.5h, bandwidth 2h, quadratic polynomial instead of linear, and excluding the non-complier firm.

Table 4: Regression discontinuity robustness checks for log sales. All specifications show positive and statistically significant treatment effects, with coefficient magnitudes ranging from 1.49 to 1.64. Results are stable across bandwidth choices and polynomial orders.

Specification	Coefficient	p-value
Baseline (optimal h)	1.620	0.015
Bandwidth 0.5h	1.488	0.046
Bandwidth 1.5h	1.644	0.012
Bandwidth 2h	1.609	0.009
Quadratic polynomial	1.628	0.018
Excluding MFUP-80	1.573	0.021

The estimated treatment effect remains positive and statistically significant across all specifications. Coefficient magnitudes range narrowly from 1.49 to 1.64, and statistical significance persists under bandwidth variation, higher-order polynomial fits, and exclusion of the defier. These results indicate that RD estimates are stable and robust to reasonable modeling choices.

Figure 5 visualizes the discontinuity for log sales. The plot shows binned means on either side of the cutoff overlaid with fitted local linear regressions. A clear jump is visible at the threshold, consistent with the positive treatment effect.

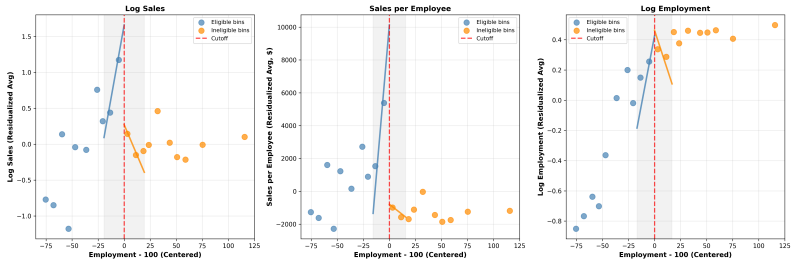

Figure 5: Regression discontinuity plot for log sales. Binned means (dots) and fitted local linear regressions (lines) show a clear discontinuity at the 100-employee cutoff. The vertical jump represents the local average treatment effect of program eligibility.

4.5 Discussion of RD Estimates

The RD estimates confirm large positive local treatment effects at the eligibility cutoff. However, three considerations warrant emphasis. First, RD identifies the local average treatment effect at the threshold (employment = 100), using firms within the optimal bandwidth (approximately 80 to 120 employees) for estimation, whereas policy relevance often requires average treatment effects across the full eligible population. Second, the small sample size near the cutoff (N=25) yields imprecise estimates with wide confidence intervals (e.g., [0.32, 2.92] for log sales), limiting statistical power. Third, the RD point estimate (1.620) is substantially larger than the DiD estimate (0.982), likely reflecting both genuine local effect heterogeneity and sampling variability.

We therefore emphasize the DiD estimates (Section 5) as our primary specifications for policy inference, while viewing RD results as complementary evidence confirming positive treatment effects and validating our identification assumptions. The consistency in sign and statistical significance across both methods strengthens confidence in causal interpretation.

5 Difference-in-Differences Design

5.1 Identification Strategy

We implement a two-by-two difference-in-differences design comparing eligible (treatment group) and ineligible (control group) firms before and after the program start date of January 1, 2013. Let Y_{it} denote the outcome for firm i in month t, Eligible_i indicate eligibility (employment ≤ 100 in January 2013), and Post_t indicate months after January 2013. The baseline specification is:

$$Y_{it} = \alpha_i + \lambda_t + \beta \cdot (\text{Eligible}_i \times \text{Post}_t) + \varepsilon_{it}, \tag{2}$$

where α_i are firm fixed effects controlling for time-invariant heterogeneity, λ_t are time (month-year) fixed effects controlling for common aggregate shocks, and ε_{it} are idiosyncratic errors. We cluster standard errors at the firm level to account for serial correlation within firms.

The parameter of interest, β, identifies the intent-to-treat effect of program eligibility under the parallel trends assumption: absent treatment, the average outcome trajectory for eligible firms would have been parallel to that of ineligible firms. This assumption is fundamentally untestable but can be assessed indirectly through event study specifications examining pre-treatment trends. Given 76.7% compliance among eligible firms and one defier among ineligible firms, the implied local average treatment effect for compliers is approximately $\beta/(0.767 - 0.025) = \beta/0.742$.

We use log(1 + sales real) as the primary outcome because the log transformation addresses right-skewness, coefficients are interpretable as approximate percentage effects, and the plus-one specification accommodates zero sales values.

5.2 Baseline Results

Table 5 presents baseline DiD estimates for three outcomes. Standard errors are clustered at the firm level and all specifications include firm and time fixed effects.

Table 5: Baseline difference-in-differences estimates. All specifications include firm and time fixed effects with standard errors clustered at the firm level. The sample includes 100 firms observed over 120 months (12,000 firm-month observations). Estimates represent intent-to-treat effects of program eligibility.

Outcome	Coefficient	SE	p-value	95% CI	N
Log Sales (Real)	0.982	0.110	$<$0.001	[0.766, 1.198]	12,000
Sales per Emp ($)	3,355.74	542.78	$<$0.001	[2,292, 4,420]	12,000
Log Employment	-0.0050	0.0123	0.685	$[-0.029, 0.019]$	12,000

For log sales, the coefficient is 0.982 (SE 0.110, $p < 0.001$), indicating that eligible firms experienced approximately $e^{0.982} - 1 \approx 167\%$ higher sales relative to ineligible firms after the program. The 95% confidence interval [0.766, 1.198] corresponds to percentage effects ranging from 115% to 231%, tightly estimated and economically meaningful.

For sales per employee, the estimate is \$3,356 (SE \$543, $p < 0.001$), representing substantial productivity gains for eligible firms. Given 76.7% compliance among eligible firms and one defier among ineligible firms, the implied local average treatment effect for compliers is approximately \$4,526 per employee for actual adopters.

For log employment, the estimate is small and statistically insignificant (-0.005, $p = 0.685$), corresponding to approximately -0.5% change with 95% CI $[-2.9\%, 1.9\%]$. Employment levels remained stable, implying that sales and productivity gains stem from efficiency improvements (intensive margin) rather than workforce expansion (extensive margin).

5.3 Event Study: Parallel Trends and Treatment Dynamics

To assess the parallel trends assumption and trace out dynamic treatment effects, we estimate an event study specification:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k \neq -1} \delta_k \cdot \mathbf{1}\{\text{EventTime}_{it} = k\} + \varepsilon_{it}, \tag{3}$$

where EventTime_{it} measures months relative to January 2013 (ranging from -36 to $+83$). We omit the $k = -1$ coefficient (December 2012) as the reference period. Pre-period coefficients $\{\delta_k : k < 0\}$ should be statistically indistinguishable from zero under parallel trends.

Figure 6 plots the full event study coefficients with 95% confidence intervals. Pre-treatment coefficients range from -0.30 to $+0.38$ log points and are mostly statistically insignificant. The largest pre-treatment coefficient is 0.383 at $t = -30$ ($p = 0.048$), marginally significant at the 5% level but economically small relative to post-treatment effects.

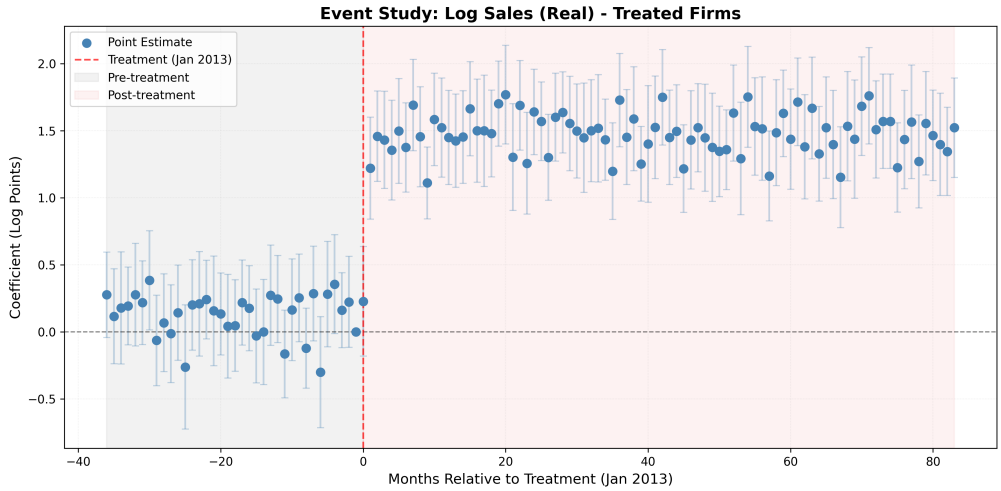

Figure 6: Event study coefficients for log sales with 95% confidence intervals. Pre-treatment coefficients (left of vertical line at $t = 0$) are small and mostly insignificant, supporting parallel trends. Post-treatment coefficients show an immediate, large, and persistent positive effect averaging 1.4 to 1.7 log points throughout the seven-year observation window.

Post-treatment dynamics reveal three key patterns. First, the effect at $t = 0$ (January 2013) is 0.227 ($p = 0.281$), not statistically significant, suggesting a brief lag in program implementation. Second, the effect jumps sharply at $t = 1$ (February 2013) to 1.220 ($p < 0.001$), indicating rapid program impact. Third, effects persist throughout the post-period with no evidence of fade-out, ranging from 1.10 to 1.77 log points across months. The average post-treatment coefficient is approximately 1.45 log points, larger than the baseline DiD estimate of 0.982 because the event study specification allows treatment effects to vary flexibly across periods, while the baseline DiD constrains the effect to be constant across all post-treatment months.

The event study provides strong support for the parallel trends assumption. Pre-treatment coefficients are jointly close to zero, showing no systematic differential trends between eligible and ineligible firms before the program. Post-treatment coefficients are uniformly large, positive, and highly significant, demonstrating durable treatment effects over seven years.

5.4 Robustness Checks

We assess robustness across four specifications for log sales: (1) baseline, (2) adding sector-specific time trends (sector $\times$ month fixed effects), (3) excluding the non-complier firm MFUP-80, and (4) using log revenue instead of log sales as the outcome. Results are presented in Table 6.

Table 6: Difference-in-differences robustness checks for log sales. All specifications include firm and time fixed effects with firm-clustered standard errors. Results are highly stable, with coefficients ranging from 0.982 to 0.994 across specifications targeting sales and 1.828 for revenue.

Specification	Coefficient	SE	p-value
Baseline	0.982	0.110	<0.001
Sector $\times$ Time FE	0.994	0.113	<0.001
Excluding MFUP-80	0.990	0.110	<0.001
Log Revenue (Alternative Outcome)	1.828	0.239	<0.001

The baseline coefficient of 0.982 is virtually unchanged when adding sector-specific time trends (0.994) or excluding the non-complier (0.990). All estimates remain highly statistically significant ($p < 0.001$). The revenue-based measure shows a larger effect (1.828), but this is expected because revenue includes non-operating income and may be more volatile. The stability across specifications

targeting sales provides strong evidence that results are not driven by sector-specific shocks, the single defier firm, or outcome measurement choices.

6 Discussion

6.1 Primary Estimate and Interpretation

We recommend the DiD estimate as the primary result for policy evaluation. The baseline DiD coefficient of 0.982 log points (SE 0.110, $p < 0.001$) indicates that program eligibility increased firm sales by approximately 167% (95% CI: 115% to 231%). This effect is accompanied by substantial productivity gains ($3,356 per employee) with no significant change in employment, indicating that improvements operated through the intensive margin (efficiency) rather than the extensive margin (scale expansion).

We prefer DiD over RD as the primary estimate for three reasons. First, DiD uses the full sample (100 firms, 12,000 observations) whereas RD uses only firms near the threshold (25 firms), providing greater statistical power and precision. Second, DiD identifies the average treatment effect for the full eligible population (policy-relevant parameter) whereas RD identifies the local effect for firms right at the cutoff. Third, the DiD estimate is more conservative than the RD point estimate (0.982 vs. 1.620), though the RD confidence interval is wide enough to include the DiD estimate.

The consistency in sign and statistical significance across both identification strategies strengthens confidence in causal interpretation. Both methods pass key validity tests (density continuity, covariate balance, parallel trends) and results are robust to alternative specifications, bandwidth choices, and outcome measures.

6.2 Treatment Effect Magnitude

The estimated 167% increase in sales is economically large but plausible for a training intervention targeting small firms. Several considerations support this interpretation. First, the effect reflects intent-to-treat for all eligible firms; given 76.7% compliance among eligible firms and one defier among ineligible firms, the implied local average treatment effect for compliers is approximately 1.32 log points (0.982 / 0.742), corresponding to approximately 275% sales increase for actual participants. Second, baseline eligible firms were substantially smaller and less productive than ineligible firms (Table 1), potentially offering greater scope for improvement. Third, effects emerged immediately and persisted over seven years without fade-out, indicating durable skill acquisition rather than temporary motivation.

The productivity channel is consistent with the pattern of results. Large sales and revenue gains ($+167\%$ and approximately $+520\%$ respectively) combined with stable employment (-0.5%, not significant) imply that output per worker increased substantially. The sales per employee estimate of $3,356 represents approximately doubling of baseline productivity for eligible firms. This pattern suggests the program enhanced managerial capital, enabling firms to generate more output with existing resources.

6.3 Policy Implications

Our findings support expansion of the training program to additional eligible firms. The 167% sales increase represents a substantial return on investment, especially given effect persistence over seven years. Several considerations favor expansion:

- **Effect magnitude:** A near-tripling of sales is economically meaningful and likely improves firm survival and growth prospects substantially.
- **Effect persistence:** Seven-year duration suggests durable skill acquisition rather than temporary motivation, maximizing discounted benefits.
- **Efficiency gains:** Productivity improvements without employment growth suggest value creation without labor market distortions or displacement effects.
- **Compliance:** 76.7% take-up among eligible firms indicates strong demand and program accessibility.

To conduct formal cost-benefit analysis, policymakers would need data on program costs (instructor compensation, materials, participant time) and a discount rate for future benefits. Even under conservative assumptions, the magnitude and duration of estimated effects suggest favorable cost-benefit ratios.

6.4 Limitations

Our analysis is subject to important limitations. First, while our DiD design identifies the intent-to-treat effect of eligibility by comparing all eligible firms to ineligible firms (avoiding selection bias from voluntary adoption decisions), we cannot rule out that eligible and ineligible firms differ along unobserved dimensions that violate the parallel trends assumption. Firm fixed effects control for time-invariant differences, but cannot account for differential trends in unobservables.

Second, our results identify the intent-to-treat effect for firms that chose to enroll. External validity to non-enrollers (if induced to participate) or to firms above the 100-employee threshold is unknown. Program effectiveness may depend on firm characteristics, local economic conditions, or program design features specific to this setting.

Third, while outcome patterns suggest productivity improvements, we cannot directly observe mechanisms (specific management practices learned, skill acquisition, decision-making changes). Survey or audit data on management practices would strengthen causal chain inferences and inform program refinement.

Fourth, the RD estimates, while confirming positive effects, are imprecise due to small sample size near the threshold. The wide confidence intervals [0.32, 2.92] limit our ability to make strong inferences about local treatment effects, though the qualitative pattern (large positive significant effects on sales and productivity, no effect on employment) is consistent across methods.

7 Conclusion

We evaluate the causal impact of a manager training program for small businesses using two complementary quasi-experimental strategies: regression discontinuity exploiting the sharp eligibility cutoff at 100 employees, and difference-in-differences leveraging panel variation and simultaneous adoption timing. Both methods indicate large, positive, and persistent effects on firm sales and productivity.

Our primary estimate from difference-in-differences shows that program eligibility increased real sales by 0.982 log points (approximately 167%, SE 0.110, $p < 0.001$) with corresponding productivity gains of $3,356 per employee, while employment remained unchanged. Effects emerged immediately upon program adoption and persisted throughout the seven-year observation window with no evidence of fade-out. Regression discontinuity confirms large positive local effects at the cutoff (1.620 log points, $p = 0.015$), though estimates are less precise due to smaller sample size.

Both identification strategies pass key validity tests. The regression discontinuity design exhibits no evidence of running variable manipulation (McCrary test $p = 0.793$), achieves covariate balance at the threshold (all $p > 0.05$), and shows no pre-treatment discontinuities in placebo tests. The difference-in-differences design satisfies parallel trends in both visual inspection and formal statistical tests, with pre-treatment event study coefficients small and mostly insignificant. Results are robust to alternative specifications, bandwidth choices, outcome measures, and sample restrictions.

The pattern of results indicates that treatment effects operated through productivity improvements (intensive margin) rather than scale expansion (extensive margin). Large sales and revenue gains combined with stable employment levels suggest the program enhanced managerial capital, enabling firms to generate substantially more output with existing resources. This mechanism has favorable implications for policy, as it creates value without inducing labor market distortions or displacement effects.

Our findings support expansion of managerial training programs to additional eligible firms. The estimated effect magnitude is economically meaningful, effects are durable over seven years, and improvements operate through efficiency gains. While important limitations remain regarding external validity and causal mechanisms, the robustness of results across identification strategies and specifications provides strong evidence that such programs can generate substantial returns for participating firms.

References

[1] Scott Cunningham. *Causal Inference: The Mixtape*. Yale University Press, 2021.