What Do Treatment Effects Measure? Marginal Responses and Financial Constraints in Corporate Finance

Murray Z. Frank*

February 24, 2026

Abstract

Three prominent studies (Rauh, 2006; Lemmon and Roberts, 2010; Bakke and Whited, 2012) all estimate the causal effect of financing on investment. They find sharply different values of zero, 0.65, and one. This paper explains why all three are correct. In a canonical investment model, the marginal response of investment to financing is binary. It is one for constrained firms, zero for unconstrained firms. The local average treatment effect recovers the share of constrained firms among compliers. It is not a structural elasticity. Those three studies all estimate the same structural relationship. What differs is the respective sets of complier firms. Evidence from the broader quasi-experimental literature confirms the pattern. Estimates approach zero for unconstrained subsamples and one for constrained subsamples. An intermediate treatment effect is not a structural parameter. It is evidence of population heterogeneity. Generalizing estimates across settings requires explicit accounting for constraint shares. This paper also develops diagnostics distinguishing the binary specification from smooth alternatives and examines extensions for large shocks and precautionary savings.

Keywords: Treatment effects, LATE, financial constraints, investment, structural estimation
JEL Codes: G31, G32, C21

*University of Minnesota, E-mail: murra280@umn.edu. Thanks to Ruiqi Chen, Gabriel Chodorow-Reich, Michael Roberts, Max Maksimovic, and Toni Whited for helpful comments. In the process of writing this I used the AI systems called Claude Opus, Deepseek, Gemini Pro, and Grok, much like referees. I asked each of them to identify mistakes, and to help improve the writing. I alone am responsible for all remaining errors. There is no research support, nor any financial relationships or other potential conflicts of interest that apply to the author writing this paper.

1 Introduction

The credibility revolution (Angrist and Pischke, 2009; Goldsmith-Pinkham, 2024) has transformed empirical corporate finance. Researchers now routinely exploit quasi-experimental variation using methods such as instrumental variables, difference-in-differences, and regression discontinuity designs to establish causal relationships between financing and real outcomes. The resulting estimates are internally valid and statistically well defined. But what do they mean in financial economic terms?

Consider a concrete example. Rauh (2006) uses regulatory nonlinearities in pension funding rules as a source of exogenous variation and estimates that corporate investment declines by about \$0.60–\$0.70 for each dollar of mandatory pension contributions. Lemmon and Roberts (2010) estimate that investment declined nearly one-for-one with the contraction in debt issuance following the collapse of the junk bond market in 1989-1990. The puzzle is actually deeper still. Bakke and Whited (2012) revisit the Rauh (2006) setting using a regression discontinuity design and find essentially no effect. So we have three estimates of the same structural relationship: zero, 0.65, and one. All come from seemingly plausible sources of exogenous variation.

This paper supplies the missing structural interpretation. In a canonical investment model with financial constraints, the marginal response of investment to financing is one for constrained firms (at corner solutions) and zero for unconstrained firms (with slack financing capacity). Treatment effects are therefore weighted averages of zeros and ones. The local average treatment effect identifies the share of constrained firms among compliers. It is not the structural marginal response itself. An IV estimate of 0.65 does not mean that constrained firms invest sixty-five cents per dollar of financing. It means that approximately 65% of compliers are financially constrained, each with a marginal response of one.

The same logic applies to all standard treatment effects estimands. The Average Treatment Effect (ATE) identifies the population constraint share. The Conditional Average Treatment Effect (CATE) identifies the constraint probability conditional on observables (Chernozhukov et al., 2025). The Marginal Treatment Effect (MTE) of Heckman and Vytlacil (2005) identifies the constraint probability at each quantile of unobserved resistance to treatment. In each case the estimand recovers a probability of being constrained for a specific subpopulation.

Applying the framework to Rauh (2006), Bakke and Whited (2012), and Lemmon and Roberts (2010) gives a clean reconciliation. Rauh's complier population spans a range of credit qualities.

His 0.60–0.70 estimate means that roughly 60–70% of compliers are at corner solutions. Bakke and Whited restrict attention to firms near the pension funding threshold. These are larger, more liquid firms that are far more likely to have bond ratings. They get a near-zero constraint share. Lemmon and Roberts study below investment grade firms. This is a population that is predominantly constrained by construction, yielding a near unity estimate. All three results are consistent with the same structural marginal response of one. What differs across the studies is the composition of the complier firms.

A survey of the broader quasi-experimental literature confirms that the pattern holds across countries, decades, and identification strategies. Studies consistently find larger effects for subsamples more likely to be constrained and effects near zero for likely unconstrained subsamples (e.g., Khwaja and Mian (2008), Chodorow-Reich (2014), Duchin et al. (2010), Campello et al. (2010), Chaney et al. (2012), Almeida et al. (2012)). The within study evidence from Khwaja and Mian (2008) is particularly informative. Treatment effects by firm size decile range from approximately zero for the largest firms to 0.6–0.8 for the smallest. That traces out exactly the variation in constraint shares predicted by the model.

Beyond the baseline binary model we examine two extensions. Financing shocks could be large enough to push previously unconstrained firms across the constraint boundary. In that case the binary decomposition remains the correct first-order approximation. We derive bounded deviations for finite shocks.

Constrained firms might retain precautionary cash. Using Lemmon and Roberts (2010) as an anchor population, cross study variation implies precautionary savings of only about three cents per dollar. That is consistent with their direct evidence that treated firms did not accumulate cash.

We also develop a diagnostic that distinguishes the binary specification from a smooth alternative with continuously varying marginal responses. Under the binary model, subsample estimates should converge toward zero or one as conditioning on constraint relevant observables becomes finer. Under the smooth model they should stabilize at interior values. The available evidence favors the binary specification.

The paper contributes to two literatures. It connects the econometric literature on treatment effect interpretation (Heckman and Vytlacil, 2005; Imbens, 2010; Mogstad and Torgovitsky, 2024; Blandhol et al., 2022) to a canonical problem in corporate finance. As Haile (2025) emphasizes, the economic content of treatment-effect estimands depends on the structural model used to interpret them. We provide a concrete accounting for the evidence on the investment financing

relationship.

The paper also contributes to the financial constraints literature initiated by Fazzari et al. (1988). Kaplan and Zingales (1997) argued that investment cash flow sensitivity does not monotonically reflect constraint severity. Farre-Mensa and Ljungqvist (2016) raised concerns about standard proxies. Hoberg and Maksimovic (2015) turned to language based measures instead. Our framework sidesteps these issues. Treatment effects supply a revealed preference measure of constraint shares among compliers. This is an object that is well defined regardless of how constraint severity varies across firms and that requires no ex-ante classification.[1]

For empirical practice, the critical implication is that intermediate treatment effects reflect population heterogeneity, not structural parameters. Heterogeneity analysis by constraint proxies and explicit accounting for constraint shares become essential. They serve both for recovering structural content, and for generalizing estimates across populations.

Section 2 develops the structural model. Section 3 derives what treatment effects identify. Sections 4 and 5 apply the framework. Section 6 evaluates the binary specification. Section 7 concludes.

2 A Structural Model of Investment and Financing

This section provides a simple model of corporate investment with financial constraints. The model has predictions about how investment responds to changes in financing capacity. This provides the structural benchmark for interpreting the treatment effect estimates. The main result is that the marginal response of investment to financing equals one for constrained firms and zero for unconstrained firms.

2.1 Environment

Consider a firm choosing investment I to maximize profits subject to a financing constraint. The firm solves

$$\max_{I \geq 0} \ \Pi(K + I) - C(I) - pI \tag{1}$$

[1] A distinct approach builds on dynamic structural models of the firm (Hennessy and Whited, 2007; Strebulaev, 2007; Whited and Wu, 2006). Such models are richer but typically require numerical solution. They are often used to match moments in the data and to do counterfactuals. Our goal is different. Our purpose is to provide a common interpretive framework for the quasi-experimental literature.

subject to:

$$I \leq W + D \tag{2}$$

where $\Pi(\cdot)$ is a profit function satisfying $\Pi' > 0$ and $\Pi'' < 0$, $C(I) = \frac{\gamma}{2} I^2$ is a convex adjustment cost with $\gamma > 0$, p is the price of capital, K is the existing capital stock, W is internal funds, and D is debt capacity.

The financing constraint in equation (2) captures the essence of financial frictions. The firm can fund investment using internal funds W plus external debt up to capacity D. Equity issuance is sufficiently costly that we abstract from it, or equivalently, the constraint already incorporates the shadow cost of external equity. The model nests frictionless capital markets. When D is arbitrarily large, the constraint never binds and financing does not limit investment.

The model is intentionally stylized. It abstracts from dynamics, uncetainty, and endogenous debt capacity to isolate the constrained/unconstrained distinction. Concave production, convex adjustment costs, and a financing constraint that may or may not bind are standard from Fazzari et al. (1988) to Almeida and Campello (2007). Our contribution is to show how the marginal implications connect to the treatment effect estimands.

2.2 Optimal Investment

Let I^{unc} denote the unconstrained optimum, which satisfies the first-order condition:

$$\Pi'(K + I^{unc}) = \gamma I^{unc} + p. \tag{3}$$

This condition equates the marginal benefit of investment with its marginal cost. The unconstrained optimum depends on investment opportunities through $\Pi'(\cdot)$ but is independent of the firm's financial position.

Define constraint slack as

$$S \equiv W + D - I^{unc}. \tag{4}$$

When $S \geq 0$, the firm has sufficient funds to achieve its unconstrained optimum. When $S < 0$, desired investment exceeds available financing and the constraint binds. Optimal investment is

$$I^* = \begin{cases} I^{unc} & \text{if } S \geq 0 \quad \text{(unconstrained)} \\ W + D & \text{if } S < 0 \quad \text{(constrained).} \end{cases} \tag{5}$$

Unconstrained firms invest at the first-best level I^{unc}. Constrained firms invest all available funds $W + D$, which falls short of the first-best.

2.3 The Structural Marginal Response

The main structural object is the marginal response of investment to changes in financing capacity.

Proposition 1 (Structural Response to Financing) *The marginal effect of financing capacity on investment is*

$$MR \equiv \frac{\partial I^*}{\partial F} = \begin{cases} 0 & \text{if } S \geq 0 \quad \textit{(unconstrained)} \\ 1 & \text{if } S < 0 \quad \textit{(constrained)} \end{cases} \tag{6}$$

where $F \in \{W, D\}$ denotes either internal funds or debt capacity.

The proof follows immediately from equation (5). For unconstrained firms, $I^* = I^{unc}$, which does not depend on W or D. For constrained firms, $I^* = W + D$, so $\partial I^*/\partial W = \partial I^*/\partial D = 1$.

The economic intuition is straightforward. Unconstrained firms have slack, so an additional dollar does not change investment. Constrained firms are at a corner, so an additional dollar translates directly into investment.

For shocks to internal funds, this object is the marginal propensity to invest (MPI). For shocks to debt capacity, it is the marginal response to debt (MRD). Despite the different terminology, these are the same structural parameter:

$$MR_{structural} = MPI_{constrained} = MRD_{constrained} = 1. \tag{7}$$

This equivalence is central to our empirical applications. Rauh (2006) identifies MPI using shocks to internal funds. Lemmon and Roberts (2010) identifies MRD using shocks to debt capacity. Both estimate the same structural object for constrained firms, using different sources of variation.

Figure 1 illustrates the result. The investment function has a kink at I^{unc}. Below the kink, each dollar of financing translates to a dollar of investment. Above the kink, financing is slack.

2.4 Population Heterogeneity

Constraint heterogeneity is central to understanding what treatment effects identify. Let $\theta \equiv P(S < 0)$ denote the population share of constrained firms. The population average marginal

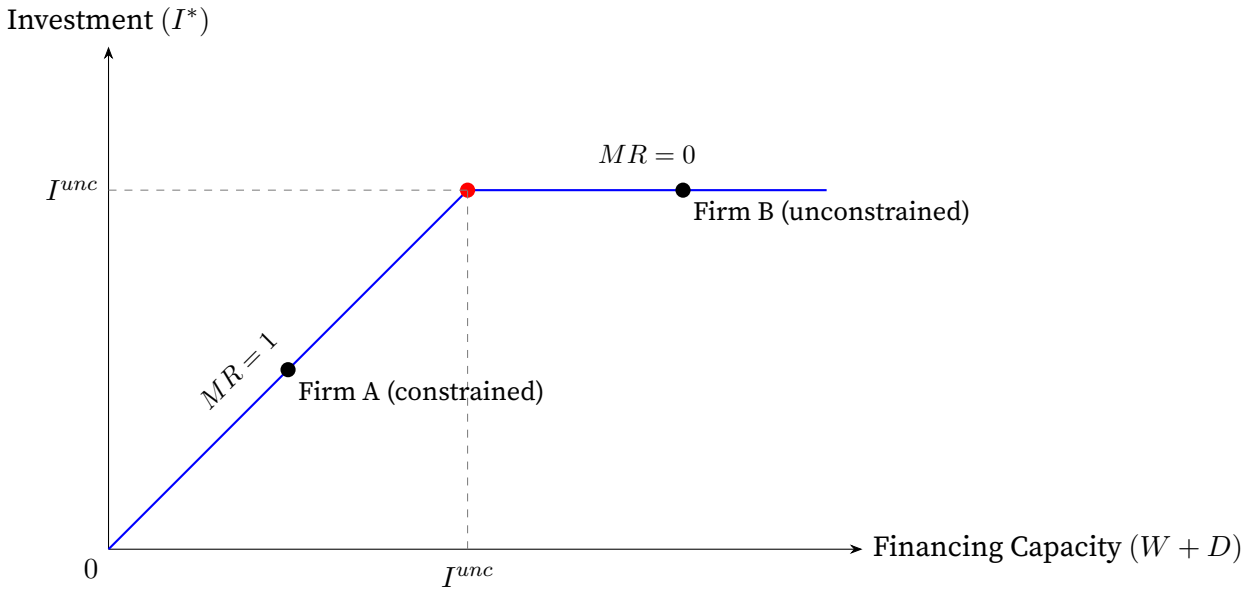


Figure 1: The corner solution in optimal investment. The figure plots optimal investment I^* as a function of financing capacity $(W+D)$ from equation (5). When financing capacity falls below the unconstrained optimum I^{unc}, the firm invests all available funds and the marginal response equals one. When financing capacity exceeds I^{unc}, the firm invests at the first-best level and additional financing is slack, so the marginal response equals zero. Firm A represents a constrained firm (e.g., a below-investment-grade firm in Rauh's sample) located on the 45-degree segment. Firm B represents an unconstrained firm (e.g., an investment-grade firm) located on the flat segment. Rauh's baseline estimate of 0.65 reflects a weighted average. Approximately 65% of compliers are like Firm A with $MR = 1$, and 35% are like Firm B with $MR = 0$.

response is

$$E\left[\frac{\partial I^*}{\partial F}\right] = \theta \cdot 1 + (1-\theta) \cdot 0 = \theta. \tag{8}$$

This average is strictly less than the structural marginal response for constrained firms whenever some firms are unconstrained ($\theta < 1$).

What determines whether a firm is constrained? From equation (4), a firm is constrained when investment opportunities (captured by Π') are strong relative to available financing ($W + D$). Firms with better investment opportunities, lower internal funds, or tighter debt capacity are more likely constrained. This maps naturally to observable characteristics: smaller firms, firms with lower credit ratings, and firms with less cash are more likely to face binding financing constraints.

For applications involving debt capacity shocks, there is also an intensive margin. The elasticity of investment with respect to debt capacity for constrained firms is

$$\varepsilon_D \equiv \frac{\partial I^*}{\partial D} \cdot \frac{D}{I^*} = \frac{D}{W+D}. \tag{9}$$

This elasticity equals the debt share of investment financing. Constrained firms that rely more heavily on debt have larger elasticities, but all have the same marginal response of one. We focus on the marginal response rather than the elasticity because the marginal response is more directly comparable across shock types.

2.5 Scaling and Empirical Specification

In empirical practice, researchers commonly scale variables by lagged assets to control for firm size. Define $i = I/A$, $w = W/A$, and $d = D/A$ where A is a predetermined scaling factor such as lagged assets. For constrained firms, $i^* = w + d$ implies $\partial i^*/\partial w = \partial i^*/\partial d = 1$. For unconstrained firms, $i^* = i^{unc}$ where i^{unc} does not depend on w or d, so both derivatives equal zero. The structural interpretation is therefore valid under the standard scaling conventions used in the empirical literature. The treatment effect identifies the share of constrained firms among compliers regardless of whether the regression is estimated in levels or scaled by assets.

2.6 Discussion

Several features of the model deserve comment. The model structure isolates the constrained versus unconstrained distinction which drives the interpretation of the treatment effects. Section 6

evaluates this assumption against the data.

The model treats constraint status as given. In practice, constraint status may be endogenous. Firms choose financial policies that affect future financing capacity. This endogeneity does not invalidate the framework as long as the instruments used in empirical applications are orthogonal to unobserved determinants of constraint status. The requirement is that variation in financing is exogenous, not that constraint status itself is exogenous.

The model assumes constrained firms invest all available funds. If constrained firms hold precautionary cash, the marginal response would be below one. Lemmon and Roberts (2010) find no evidence of increased cash holdings among their treated firms and report that substitution to alternative financing was "extremely limited." Appendix B addresses this extension formally.

3 What Do Treatment Effects Identify?

The previous section characterized the structural marginal response of investment to financing. It is one for constrained firms, zero for unconstrained firms. This section asks what treatment effects identify in terms of this structural parameter. The central result is that treatment effects are weighted averages of structural responses. The weights are determined by the composition of the complier population.

3.1 Setup

Consider a shock that affects financing for some firms. Let $Z_i \in \{0, 1\}$ indicate whether firm i is exposed to the shock. Exposure changes either debt capacity D_i or internal funds W_i by an amount Δ_i. For concreteness, suppose $Z_i = 1$ indicates a negative financing shock, as in both of our main empirical applications.

The standard instrumental variables assumptions are 1) Exclusion. Z_i affects investment only through the financing channel. 2) Relevance. $E[\Delta_i|Z_i = 1] \neq E[\Delta_i|Z_i = 0]$. 3) Monotonicity. The shock weakly decreases (or increases) financing for all affected firms. There are no defiers.

Under these assumptions, the instrumental variables estimator identifies the local average treatment effect (LATE) for the subpopulation of compliers. These are the firms whose financing is affected by the instrument (Imbens and Angrist, 1994). The LATE is a well defined statistical object, but its economic content depends on the structural model used to interpret the estimates (Heckman and Vytlacil, 2007; Haile, 2025).

3.2 LATE in the Structural Model

Within our framework, we can show exactly what LATE identifies.

Proposition 2 (LATE Decomposition) *The local average treatment effect of financing on investment equals:*

$$\tau_{LATE} = E\left[\frac{\partial I^*}{\partial F} \;\middle|\; \textit{Complier}\right] = P(S < 0 \mid \textit{Complier}) \cdot 1 + P(S \geq 0 \mid \textit{Complier}) \cdot 0. \tag{10}$$

Simplifying,

$$\tau_{LATE} = \theta_C \tag{11}$$

where $\theta_C \equiv P(S < 0 \mid \textit{Complier})$ is the share of constrained firms among compliers.

The proof follows directly from the model. Among compliers, constrained firms have $\partial I^*/\partial F = 1$ and unconstrained firms have $\partial I^*/\partial F = 0$. The LATE is the average of these responses, weighted by the shares of each type among compliers. Since the responses are binary, this average equals the share with response equal to one; that is the share of constrained compliers.

This result has a sharp interpretation. The LATE does not identify the structural marginal response for constrained firms. That parameter just equals one. It does not identify the population share of constrained firms. That parameter is θ. It identifies the constraint share among the specific subpopulation of compliers. It answers the following question. What fraction of firms whose financing is affected by the instrument are financially constrained?

The complier population may differ systematically from the overall population. Instruments that exploit financial distress such as pension underfunding, credit supply contractions, may disproportionately affect constrained firms. If so, $\theta_C > \theta$. Instruments that affect financing broadly may have complier populations that mirror the general population. If so $\theta_C \approx \theta$. Without knowing which case applies, the magnitude of a treatment effect cannot be interpreted structurally. The statistics does not by itself provide the meaning. That comes from the economic model grounded in a reasonable interpretation of the actual context.

3.3 Other Treatment Effect Parameters

The same logic applies to other common estimands.

Average Treatment Effect (ATE). Random assignment of financing shocks to all firms would identify

$$\tau_{ATE} = E\left[\frac{\partial I^*}{\partial F}\right] = \theta, \tag{12}$$

the population share of constrained firms. The ATE differs from LATE whenever compliers have a different constraint share than the population ($\theta_C \neq \theta$).

Conditional Average Treatment Effect (CATE). Estimating treatment effects conditional on observable characteristics X yields

$$\tau(X) = E\left[\frac{\partial I^*}{\partial F} \;\middle|\; X\right] = P(S < 0 \mid X). \tag{13}$$

The CATE for firms with characteristics X equals the probability that such firms are constrained. If X includes determinants of constraint status such as firm size, credit rating, and cash holdings, then variation in $\tau(X)$ reveals structural heterogeneity. Finding larger effects for smaller, lower-rated, or cash-poor firms is consistent with the model. These firms are more likely constrained, so a higher share of them have $\partial I^*/\partial F = 1$.

Marginal Treatment Effect (MTE). With continuous variation in treatment intensity, one can estimate the marginal treatment effect (Heckman and Vytlacil, 2005).

$$MTE(u) = E\left[\frac{\partial I^*}{\partial F} \;\middle|\; U = u\right] = P(S < 0 \mid U = u) \tag{14}$$

where U indexes unobserved resistance to treatment. If firms with low resistance are more likely constrained, the MTE is decreasing in u. The most responsive firms are also the most constrained.

Table 1 summarizes the mapping from estimands to structural parameters. Each estimand identifies the probability of being constrained for a specific subpopulation. None directly identifies the structural marginal response.

Table 1: What Treatment Effect Estimands Identify

Estimand	Structural interpretation	Identifies	Equals $MR_{structural}$ when
LATE	Constraint share among compliers	θ_C	All compliers constrained
ATE	Constraint share in population	θ	All firms constrained
CATE(X)	Constraint probability given X	$P(S<0 \mid X)$	X fully determines constraint status
MTE(u)	Constraint probability at quantile u	$P(S<0 \mid U=u)$	All firms at u constrained

Notes: The structural marginal response equals one for constrained firms and zero for unconstrained firms. Each estimand identifies the probability of being constrained for a particular subpopulation: compliers (LATE), the full population (ATE), firms with characteristics X (CATE), or firms at propensity score quantile u (MTE). Constraint slack is $S \equiv W + D - I^{unc}$. A firm is constrained when $S < 0$.

3.4 Recovering Structural Parameters

Given a LATE estimate $\hat{\tau}$, can we recover the structural marginal response for constrained firms? Yes, if we know or can estimate the constraint share among compliers.

$$\widehat{MR}_{structural} = \frac{\hat{\tau}}{\hat{\theta}_C}. \tag{15}$$

If the constraint share is unknown, we can bound the structural parameter. Since $0 \leq \theta_C \leq 1$ and $MR_{structural} = 1$ for constrained firms in our model,

$$\hat{\tau} \leq MR_{structural} \leq 1. \tag{16}$$

The lower bound is the LATE itself, attained when all compliers are constrained ($\theta_C = 1$). The upper bound is one. An estimate of $\hat{\tau} = 0.65$ implies that the structural marginal response lies between 0.65 and 1.

These bounds can be tightened using subsample estimates. If low-rated firms have $\hat{\tau}_{low} = 0.84$ and high-rated firms have $\hat{\tau}_{high} = 0.50$, the pattern suggests that low-rated firms are more constrained. Under the assumption that nearly all low-rated compliers are constrained ($\theta_{C,low} \approx 1$), the estimate $\hat{\tau}_{low} \approx 0.84$ provides a tighter lower bound. Alternatively, it implies that even among low-rated firms, approximately 16% are unconstrained.

3.5 A Monte Carlo Illustration

To illustrate Proposition 2, we simulate the IV estimator in the structural model and verify that it recovers θ_C rather than the structural marginal response.

For each value of θ_C on a grid from 0.05 to 0.95, we draw $N = 1{,}000$ firms. Each firm is independently assigned constraint status. They are constrained with probability θ_C and unconstrained with probability $1 - \theta_C$. Constrained firms have $MR_i = 1$; unconstrained firms have $MR_i = 0$. A unit financing shock is applied to all firms. The investment response is $\Delta I_i = MR_i \cdot \Delta F$, so constrained firms increase investment by one dollar and unconstrained firms do not respond. The IV (Wald) estimator is $\hat{\tau} = \overline{\Delta I}/\Delta F$, the sample mean of the marginal responses. We repeat this procedure 500 times at each grid point.

Figure 2 plots the results. The solid line is the theoretical prediction $\tau_{LATE} = \theta_C$. The blue circles are Monte Carlo means, and the shaded band is the 90% Monte Carlo interval. Two features are clear. The IV estimator is unbiased. The Monte Carlo means lie on the 45-degree line at every grid point, with bias below 0.001 everywhere. The estimator is precise. The 90% interval never exceeds ± 0.03, reflecting a maximum standard error of $\sqrt{\theta_C(1-\theta_C)/N} \approx 0.016$ at $\theta_C = 0.5$.

The red squares in Figure 2 locate four empirical studies along the 45-degree line. Chaney et al. (2012), has an estimate of 0.06 and it appears near the origin. Their sample of large publicly traded firms is mostly unconstrained. Khwaja and Mian (2008) at 0.60 and Rauh (2006) at 0.65 occupy the middle, reflecting mixed complier populations. Lemmon and Roberts (2010) at approximately 0.97 appears near the upper corner, reflecting a predominantly constrained sample of junk-rated firms. These studies differ in their position *along* the 45-degree line not in their distance from it. They differ in the composition of their complier populations. All are consistent with the same structural marginal response of unity.

Figure 3 displays the sampling distribution of the IV estimator at four selected values of θ_C: 0.10 (mostly unconstrained), 0.35 (mixed), 0.65 (comparable to Rauh's baseline), and 0.95 (comparable to Lemmon and Roberts). Each distribution is centered on the true θ_C with no detectable bias and is approximately normal, as expected from the central limit theorem applied to sample means of binary outcomes.

The simulation confirms the analytical result. The IV estimator does not recover the structural marginal response for constrained firms, nor for unconstrained firms. It recovers θ_C, the share of constrained firms among compliers. An estimate of $\hat{\tau} = 0.65$ is **not** a noisy estimate of one,

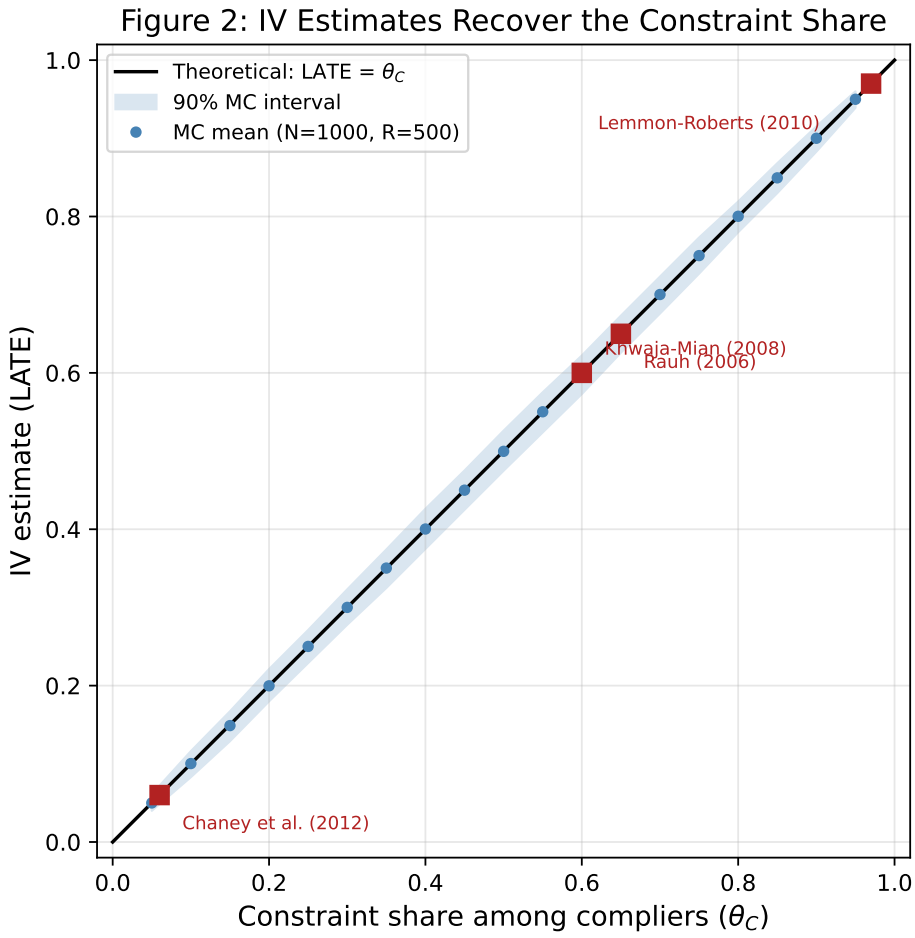


Figure 2: IV estimates recover the constraint share. The solid line is the theoretical prediction $\tau_{LATE} = \theta_C$ from Proposition 2. Blue circles are Monte Carlo means across 500 replications with $N = 1{,}000$ firms. The shaded band is the 90% Monte Carlo interval. Red squares locate empirical estimates: Chaney et al. (2012) at 0.06, Khwaja and Mian (2008) at 0.60, Rauh (2006) at 0.65, and Lemmon and Roberts (2010) at 0.97. Cross-study differences reflect variation in complier composition, not in structural parameters.

attenuated by econometric issues. It is a precise estimate of the probability that a complier is financially constrained. The next section applies this perspective to two influential papers.

4 Empirical Applications

This section applies the framework to two influential papers that estimate the causal effect of financing on corporate investment. Rauh (2006) finds that investment declines by approximately \$0.60 to \$0.70 for each dollar of mandatory pension contributions. Lemmon and Roberts (2010) find that investment declined nearly one-for-one with the contraction in debt issuance. Our framework

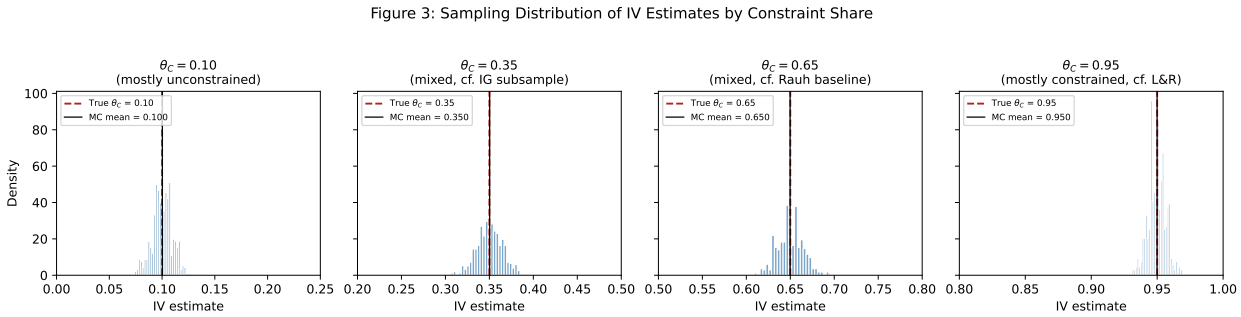


Figure 3: Sampling distribution of the IV estimator at four constraint shares. Each panel displays 500 Monte Carlo replications with $N = 1{,}000$ firms. The dashed line marks the true θ_C; the solid line marks the Monte Carlo mean. From left to right: $\theta_C = 0.10$, 0.35, 0.65, and 0.95. The estimator is unbiased and approximately normal at all values, with dispersion governed by $\sqrt{\theta_C(1-\theta_C)/N}$.

provides a unified interpretation.

4.1 Rauh (2006): Pension Contributions and Investment

4.1.1 Setting and Identification

Rauh (2006) studies how mandatory contributions to defined-benefit pension plans affect corporate investment. U.S. pension funding rules create sharp nonlinearities in required contributions. When a plan's funding ratio falls below regulatory thresholds, mandatory contributions increase discontinuously. The identification strategy exploits the interaction of funding ratios with unexpected asset returns that push plans across these thresholds. The exclusion restriction requires that pension funding shocks affect investment only through their impact on internal funds. The sample covers firms sponsoring defined-benefit pension plans during 1990-1998.

The IV estimates indicate that a dollar increase in mandatory contributions reduces capital expenditures by approximately \$0.60 to \$0.70.[2] Crucially, the paper reports substantial heterogeneity by credit rating. The sensitivity of investment to mandatory contributions is larger for firms without an investment-grade rating.

4.1.2 Interpretation Through the Framework

Rauh estimates the marginal propensity to invest out of internal funds, $\hat{\tau}_{Rauh} \approx 0.65$. The model predicts that this marginal propensity equals one for constrained firms and zero for unconstrained

[2] Rauh (2006), reports IV coefficients ranging from -0.60 to -0.72 across specifications. OLS coefficients on cash flow are approximately 0.10, presumably due to attenuation bias from measurement error (Erickson and Whited, 2000) and simultaneity.

firms. Applying Proposition 2, the IV estimate identifies the share of constrained firms among compliers.

$$\hat{\tau}_{Rauh} \approx 0.65 \quad \Longrightarrow \quad \hat{\theta}_C \approx 0.65. \tag{17}$$

Approximately 65% of compliers are financially constrained, each with $MPI = 1$. The remaining 35% are unconstrained with $MPI = 0$.

The heterogeneity by credit rating provides a consistency check. Firms without investment-grade ratings are more likely constrained, so our framework predicts a higher constraint share among compliers in this subgroup; $\hat{\theta}_{C,low\text{-}rated} > \hat{\theta}_{C,high\text{-}rated}$. This is exactly what the data show. For the subsample most likely constrained, the estimate moves closer to one, consistent with a higher fraction of compliers at corner solutions.

Under our model, the structural marginal propensity to invest for constrained firms is $MPI_{constrained} = 1$. Rauh's baseline estimate of 0.65 understates this structural parameter because the complier population includes unconstrained firms who contribute zeros to the average.

4.2 Lemmon and Roberts (2010): Credit Supply Shocks

4.2.1 Setting and Identification

Lemmon and Roberts (2010) study how exogenous contractions in below-investment-grade credit supply affect corporate financing and investment. They exploit three near concurrent shocks in 1989-1990. These are the collapse of Drexel Burnham Lambert, the passage of FIRREA (which required savings and loans to liquidate junk bond holdings), and changes in NAIC guidelines that led insurance companies to retreat from below-investment-grade private placements. Lemmon and Roberts (2010) provide a detailed discussion of each event.

The identification strategy uses difference-in-differences, comparing below-investment-grade firms (treatment) to propensity-score-matched unrated firms (control) before and after 1989. Net debt issuances by treated firms decreased by approximately 5 or 6% of assets relative to the control group.[3] The paper finds "almost no substitution" to alternative financing sources, and the consequence is "an almost one-for-one decline in net investment with the decline in net debt issuances."

[3]The model is expressed in terms of debt capacity D, a stock. Lemmon and Roberts measure debt issuance, a flow. For constrained firms at the corner, changes in debt capacity translate one-for-one into changes in issuance, so the distinction does not affect the structural interpretation.

4.2.2 Interpretation Through the Framework

Lemmon and Roberts estimate the marginal response to debt capacity, $\hat{\tau}_{LR} \approx 1$. The treatment group consists entirely of below investment grade firms. This is a population that, by definition, faces significant financing frictions. In our framework, the probability of being constrained among these firms is high.

$$\theta_{C,junk} \approx 1 \quad \Longrightarrow \quad \hat{\tau}_{LR} \approx \theta_{C,junk} \cdot 1 \approx 1. \tag{18}$$

When nearly all compliers are constrained, the treatment effect approximates the structural marginal response.

The limited substitution finding reinforces this interpretation. Lemmon and Roberts emphasize that treated firms could not substitute to bank debt, equity, or internal funds. In the model, this is precisely what we expect from constrained firms. They have already exhausted cheaper financing sources, so a reduction in debt capacity translates directly into reduced investment. Unconstrained firms, by contrast, would have slack in other sources and could substitute away from the affected margin.

4.3 Reconciliation

Bakke and Whited (2012) provide a third estimate that completes the spectrum. They revisit Rauh (2006) using a regression discontinuity design (RDD) that restricts attention to firms near the pension funding threshold. Mandatory pension contributions do not cause investment declines for these firms. The reconciliation is immediate. Near-threshold firms are systematically different from the compliers in Rauh's IV design. Bakke and Whited (2012) document that they are larger, nearly twice as likely to have bond ratings, and issue substantially more short-term debt than the deeply underfunded firms that drive Rauh's result. Under our model, these characteristics imply that most near-threshold firms have sufficient financing capacity to absorb a modest pension funding shock without cutting investment. The constraint share among RDD compliers is low, so $\hat{\tau}_{\text{RDD}} \approx 0$.

Bakke and Whited (2012) also document how near-threshold firms respond to cash shortfalls. Those firms adjust on financial margins such as receivables factoring, payout reductions, short-term borrowing. This is is precisely the behavior of unconstrained firms with slack financing capacity. By contrast, the firms studied by Rauh (2006) include deeply underfunded compliers

cannot easily substitute. For them the pension funding shock translates into reduced investment. One group adjusts on financial margins, the other on real margins. This is direct evidence of heterogeneous constraint status within the population of pension-sponsoring firms.

Table 2 presents all three studies ordered by implied constraint share. Rauh's sample spans a range of credit qualities, producing a complier population that mixes constrained and unconstrained firms; his estimate of 0.65 implies that roughly 65% are at corner solutions. Lemmon and Roberts study junk-rated firms exclusively. That is a population that is predominantly constrained by construction. The estimates span the full unit interval, and all three are consistent with a structural marginal response of one.

Table 2: Reconciling Three Estimates of the Financing Investment Relationship

	Bakke-Whited (2012)	Rauh (2006)	Lemmon-Roberts (2010)
Financing shock	Mandatory pension contributions	Mandatory pension contributions	Junk bond market collapse
Design	RDD	IV	DiD
Compliers	Near-threshold firms	All DB sponsors (all credit ratings)	Below-investment-grade firms only
Complier characteristics	Large, rated, liquid	Mixed credit quality	Below-investment-grade
Baseline estimate	≈ 0	0.60–0.70	≈ 1.0
Heterogeneity	Adjustment on financial margins	Larger effects for lower-rated firms	Limited substitution to alternative financing
Implied $\hat{\theta}_C$	≈ 0 (mostly unconstrained)	0.60–0.70 (mixed population)	≈ 1.0 (predominantly constrained)
Structural MR	1.0	1.0	1.0

Notes: All three studies estimate the causal effect of financing on investment using plausibly exogenous variation. Under the framework, all three are consistent with $MR_{\text{structural}} = 1$ for constrained firms. Cross-study differences reflect variation in the constraint share among compliers. The Rauh and Bakke-Whited estimates exploit the same pension funding rules applied to overlapping populations of defined-benefit sponsors; their different estimates arise entirely from the identification strategy selecting different complier populations. DB = defined benefit; RDD = regression discontinuity design; DiD = difference-in-differences; FIRREA = Financial Institutions Reform, Recovery, and Enforcement Act of 1989.

The Bakke and Whited comparison is particularly informative because it holds the institutional setting approximately constant. Both their RDD and Rauh's IV exploit pension funding rules applied to the same underlying population of defined-benefit sponsors. A skeptic could attribute the difference between Rauh and Lemmon-Roberts to differences in country, time period, or shock type. That objection does not apply here. The variation in estimates arises from the identifica-

tion strategy selecting different complier populations. When the running variable in an RDD is a measure of financial health, near-threshold firms are selected on characteristics correlated with constraint status. The RDD estimate is valid for this local population but need not generalize to populations with different constraint shares. This is a feature of all LATE-type estimands, but particularly acute when the running variable is itself related to the structural parameter of interest.

5 The Framework Applied to the Broader Literature

This section shows that the same logic organizes findings across a wider body of quasi-experimental research on financing and real outcomes. The central prediction is that treatment effects should be larger when complier populations contain higher shares of constrained firms. We first work through Khwaja and Mian (2008) in detail. We can do that because their unusually rich heterogeneity analysis by firm size provides clean information. We then survey additional studies more briefly. Table 4 summarizes the evidence.

Each paper uses somewhat different language to describe their estimates. Our purpose is not to dispute any individual description. These are all serious studies. Our purpose is to provide a unified economic interpretation.

5.1 A Detailed Example: Khwaja and Mian (2008)

Khwaja and Mian (2008) exploit liquidity shocks to Pakistani banks induced by unanticipated nuclear tests in May 1998. Banks with dollar denominated deposits experienced runs when the government restricted dollar withdrawals. Using loan level data with firm fixed effects, they estimate how a bank's liquidity shock affects its lending to a given firm, holding constant borrowing from other banks. The baseline estimate is that a 1 percent larger decline in bank liquidity reduces lending to the same firm by 0.6 percent.

The paper's main result for our purposes appears in Figure VII, which plots the treatment effect separately for each firm size decile. The pattern is striking and maps directly into the model. For the largest firms (top three deciles), the effect on total borrowing is "almost zero." For smaller firms, effects are substantial and increase as firm size decreases. The pattern is monotonically declining in firm size.

Under the framework, the treatment effect for size decile d identifies the constraint share in

that decile.

$$\hat{\tau}_d = \hat{\theta}_{C,d} \cdot 1 + (1 - \hat{\theta}_{C,d}) \cdot 0 = \hat{\theta}_{C,d}. \tag{19}$$

Table 3 reports the implied constraint shares. The near-zero effect for large firms implies $\hat{\theta}_C \approx 0$ in this group. That is essentially all large compliers are unconstrained. They have slack financing capacity and substitute across lenders when one bank contracts. The substantial effects for small firms imply $\hat{\theta}_C$ in the range of 60 to 80%. Most small compliers are constrained. They lack alternative financing sources, so the bank specific shock translates directly into reduced total borrowing. The full sample estimate of 0.6 is a weighted average.

$$\hat{\tau}_{full} = \sum_{d=1}^{10} \omega_d \cdot \hat{\tau}_d = 0.6 \tag{20}$$

where ω_d is the share of compliers in decile d.

This example shows several basic features. The aggregate estimate of 0.6 does not identify a structural parameter. It identifies the constraint share among compliers in the full sample. The monotonic relationship between firm size and treatment effects traces out variation in θ_C across subpopulations, exactly as the model predicts. The subsample estimates for the largest firms ($\hat{\tau} \approx 0$) and smallest firms ($\hat{\tau} \approx 0.6$–0.8) approach the structural bounds of zero and one. That is consistent with the binary marginal response.

Table 3: Implied Constraint Shares by Firm Size: Khwaja and Mian (2008)

Size Group	Deciles	Estimated $\hat{\tau}$	Implied $\hat{\theta}_C$	Distance from nearest bound
Largest firms	8–10	≈ 0	$\approx 0\%$	≈ 0
Medium-large	5–7	≈ 0.2–0.4	≈ 20–40%	0.2–0.4
Medium-small	3–4	≈ 0.5–0.6	≈ 50–60%	0.4–0.5
Smallest firms	1–2	≈ 0.6–0.8	≈ 60–80%	0.2–0.4
Full sample	1–10	0.6	60%	0.4

Notes: Treatment effects are approximate values from Figure VII of Khwaja and Mian (2008). Implied constraint shares follow from equation (19) under the assumption that $MR_C = 1$ and $MR_U = 0$. The full-sample estimate of 0.6 is from page 1413 of that paper. Distance from nearest bound is $\min(\hat{\tau}_g, 1 - \hat{\tau}_g)$. Under the binary model, this distance converges to zero as subgroups become homogeneous in constraint status. See Section 6.2 for discussion.

5.2 Additional Evidence

Table 4 surveys additional quasi-experimental studies. Rather than discuss each in detail, we highlight the common pattern predicted by the framework. We see larger effects for likely constrained

subsamples, and effects near zero for likely unconstrained subsamples.

Bank lending and employment. Chodorow-Reich (2014) extends the bank lending channel to employment during the 2008-09 crisis. Using pre-crisis lender health as a source of exogenous variation, he finds economically and statistically significant effects on employment at small and medium firms. He cannot reject zero effect at the largest or most transparent firms. This binary pattern matches the model's prediction that firms are either at corner solutions with $MR = 1$ or at interior solutions with $MR = 0$.

Credit supply contractions. Several papers study real effects of the 2007-08 financial crisis. Duchin et al. (2010) find that investment declines were greatest for firms with low cash reserves, with the bottom tercile experiencing declines roughly twice as large as the top tercile. Campello et al. (2010) survey 1,050 CFOs during December 2008 and report that constrained U.S. firms planned to cut capital spending by approximately 9%, compared to 0.4% for unconstrained firms; 86% of self-identified constrained U.S. CFOs reported bypassing attractive investment projects. Almeida et al. (2012) exploit variation in debt maturity at the crisis onset. Firms with long-term debt maturing right after the third quarter of 2007 cut their investment-to-capital ratio by 2.5 percentage points more than otherwise similar firms, with larger effects for smaller and unrated firms.

Collateral channels. Gan (2007) studies Japanese land price declines as a collateral shock, finding larger effects for high-leverage and low-liquidity firms. Chaney et al. (2012) estimate that U.S. corporations invest only $0.06 per dollar of real estate collateral. This is an estimate well below unity. It implies most publicly traded firms are unconstrained with respect to this margin. Effects are concentrated among smaller firms without bond market access, consistent with higher θ_C in those subsamples. Peek and Rosengren (2000) exploit the Japanese banking crisis as a loan supply shock transmitted to U.S. markets, finding larger effects in markets with greater Japanese bank penetration.

6 Testing the Binary Model Specification

Our framework interprets treatment effect coefficients as constraint shares among compliers rather than structural marginal responses. This interpretation depends on the model. A natural objection is that some alternative model might justify reading the coefficient as a true marginal response

Table 4: Heterogeneity in Treatment Effects Across the Literature

Paper	Baseline estimate	Constraint proxy	Constrained subsample	Unconstrained subsample
Rauh (2006)	0.60–0.70	Credit rating	Low-rated: higher	IG-rated: lower
Lemmon-Roberts (2010)	≈ 1.0	Credit rating	Junk sample	—
Khwaja-Mian (2008)	0.6	Firm size	Small: 0.6–0.8	Large: ≈ 0
Chodorow-Reich (2014)	Significant	Firm size	Small/med: sig.	Large: zero
Duchin et al. (2010)	Significant	Cash reserves	Low cash: $2\times$	High cash
Campello et al. (2010)	9% cut planned	Self-reported status	86% bypassed projects	0.4% cut
Almeida et al. (2012)	2.5 ppt decline	Rating, size	Unrated: larger	Rated
Gan (2007)	0.08	Leverage, liquidity	High lev.: larger	Low lev.
Chaney et al. (2012)	0.06	Size, bond market access	Small: larger	Large
Peek-Rosengren (2000)	Significant	Bank exposure	High exp.: larger	Low exp.

Notes: "Estimate" reports the baseline treatment effect. "Constraint proxy" identifies the observable characteristic used to split the sample into likely constrained and unconstrained subgroups. "Constrained" and "Unconstrained" summarize heterogeneity by these proxies. All studies find larger effects for likely constrained subsamples, consistent with the prediction that treatment effects identify constraint shares among compliers. The same small set of firm characteristics (credit rating, firm size, cash reserves, leverage, and bond market access) predicts the direction of heterogeneity across all studies. Campello et al. is a CFO survey. All other studies are quasi-experimental. IG = investment grade; lev. = leverage; exp. = exposure.

$\partial I/\partial F$. This section argues that the constraint interpretation is not merely one possible reading of the evidence. It is the interpretation most consistent with the empirical patterns documented in Sections 4 and 5.

6.1 Qualitative Evidence: Three Facts

The studies surveyed in Table 4 establish three facts that any interpretation must account for.

Fact 1: Effects for unconstrained subsamples are approximately zero. Across all studies, estimates for likely unconstrained firms (large firms, firms with high cash holdings, investment-grade ratings, or bond market access) are consistently close to zero and often statistically indistinguishable from it. Under the smooth alternative, these firms should still respond positively. They have valuable investment opportunities, and the financing shock raises their cost of funds. The zero response is instead what the binary model predicts for firms at interior solutions where financing is slack.

Fact 2: Effects for constrained subsamples approach unity. Estimates for likely constrained firms (small firms, junk-rated firms, firms with low cash or no bond market access) are consistently larger, and in the most constrained samples approach one. Under the smooth model, there is no reason for this particular value to happen. Investment opportunities vary across firms and industries, so a marginal response of exactly one would be coincidental. Under the binary model, the result follows directly. Constrained firms invest every available dollar, so investment moves one-for-one with financing regardless of the marginal product.

Fact 3: The pattern is monotonic and consistent across settings. The relationship between constraint proxies and treatment effects is monotonic. More constrained subsamples always have larger effects. This holds across countries (U.S., Pakistan, Japan), decades (1990s, 2000s, 2008 crisis), identification strategies (IV, DiD, matching), and outcome variables (investment, employment, lending). Under the smooth model, this cross setting consistency would require the marginal product of investment to covary with firm size, cash holdings, and credit ratings in exactly the same way in each context. An unlikely coincidence. Under the binary model, the same characteristics that predict financial constraints predict larger treatment effects. That is because the treatment effect measures the constraint share.

6.2 Distinguishing the Binary and Smooth Models

The qualitative evidence in the preceding subsection is suggestive but informal. This subsection develops sharper predictions that distinguish the binary model from the smooth alternative and evaluates them against the available evidence. The idea is that the two models make different predictions about what happens to treatment effect estimates as researchers condition more finely on constraint-relevant observables.

6.2.1 The Convergence Prediction

Under the binary model, individual treatment effects take values in $\{0, 1\}$. A subgroup estimate $\hat{\tau}_g$ is a sample average of zeros and ones, so it can take any value in $[0, 1]$. But as the partition of firms into subgroups becomes sufficiently fine that each subgroup is approximately homogeneous in constraint status, subgroup estimates must approach the boundary values:

$$\hat{\tau}_g \longrightarrow c_g \in \{0, 1\} \quad \text{as conditioning refines.} \tag{21}$$

The within-group variance $\hat{\tau}_g(1-\hat{\tau}_g)$ converges to zero because each subgroup increasingly consists of one type.

Under the smooth model, individual effects take values in the interior of $[0,1]$. Finer conditioning reduces heterogeneity within subgroups, but subgroup means converge to conditional expectations of a continuous distribution. There is no reason for these conditional expectations to approach zero or one. A firm with moderate constraint slack has a marginal response that reflects its specific position along the smooth cost curve, and grouping it with similar firms produces a subgroup mean near this interior value. Convergence to the binary bounds requires that the underlying structural responses are themselves concentrated at the bounds, which is the binary model's distinctive prediction.

The economic content of this distinction is worth emphasizing. Under the smooth model, large investment-grade firms have positive marginal products of investment and active capital expenditure programs. An additional dollar of financing reduces the marginal cost of funds and should increase investment at the margin, even if the effect is small. The prediction of an exactly zero response requires that these firms are at exact corner solutions where financing is completely slack. But this is the binary model, not the smooth model. Similarly, the prediction of a response approaching one for constrained firms requires that these firms invest every marginal dollar, leaving nothing for precautionary savings or partial adjustment. Again, a corner solution. The smooth model predicts responses that are high for constrained firms and low for unconstrained firms, but bounded away from both zero and one.

6.2.2 Within-Study Evidence: Khwaja and Mian (2008)

The cleanest test of the convergence prediction uses within-study variation, which avoids the comparability problems inherent in pooling estimates across different countries, identification strategies, and outcome variables. Khwaja and Mian (2008) provide the most informative within-study evidence because their Figure VII reports treatment effects separately for each of ten firm-size deciles, all estimated using the same identification strategy, the same country, and the same shock. Table 3 reports the implied constraint shares for each size group.

The pattern maps directly into the binary model's prediction. For the top three size deciles (the largest firms), the effect on total borrowing is approximately zero. These are firms with access to multiple lenders and sufficient financing capacity to absorb a bank-specific liquidity shock. Under the binary model, the zero effect reflects $\hat{\theta}_{C,g} \approx 0$. Essentially none of these large compliers are

constrained.

Under the smooth model, these firms should still have a positive response. The bank specific shock raises their effective cost of borrowing, and for firms with positive marginal products of investment this cost increase should reduce investment at the margin. That ought to be a small but detectable amount given the precision of the estimates. One might object that the smooth model's predicted response for large firms is simply too small to distinguish from zero. This is possible, and the test has limited power at the unconstrained end if the smooth model's marginal response for these firms is, say, 0.02 rather than 0.10. But the point estimates for the top deciles are not merely insignificant; they are economically negligible. That pattern is more naturally explained by an exact zero. Frms are at interior solutions where financing is slack. It is not a positive but vanishingly small marginal response that just happens to be for firms across a range of industries and investment opportunities.

For the bottom two size deciles (the smallest firms), effects are in the range of 0.6–0.8. Under the binary model, this shows a constraint share of 60–80% in these deciles. A reasonable concern is that these estimates seem well below unity. If the binary model is correct, why do we not observe treatment effects near one even for the smallest firms? There are forces that may help mitigate the concern.

Firm size is a noisy proxy for constraint status. Even within the bottom decile, firms differ by dimensions that affect financing capacity. Think of cash holdings, leverage, the number of banking relationships, and the quality of collateral. A decile defined solely by size will contain both firms that are genuinely at corner solutions and firms that are small but financially comfortable. A constraint share of 60–80% in the smallest deciles is entirely consistent with this within decile heterogeneity. Finer conditioning might be done by for example, interacting size deciles with cash holdings or leverage. That would likely push estimates closer to one for the most constrained subgroups and closer to zero for the least constrained. To actually do this requires access to the underlying data.

The relevant comparison is not whether the estimates reach one. It is whether they move toward one as conditioning tightens. Moving from the full sample estimate of 0.6 to decile level estimates that range from approximately zero to 0.6–0.8 is exactly the dispersion the binary model predicts. The smooth model also predicts larger effects for smaller firms. But it does not predict that the extreme subgroups should approach the boundary values of zero and one. Under the smooth model, the smallest firms should have high but interior marginal responses such as per-

haps 0.5 or 0.6. That would reflect their position along a continuous cost curve. There would be no particular reason to approach unity. The fact is that the estimates for the smallest firms are as high as 0.8 in some deciles. The estimates for the largest firms are indistinguishable from zero. That is more consistent with an underlying distribution concentrated at the binary bounds than with a smooth interior distribution.

The intermediate deciles trace out a monotonically declining path from near zero to the 0.6–0.8 range. Under the binary model, this monotonicity is due to smoothly increasing constraint shares as firm size decreases. Under the smooth model, it is due to smoothly increasing marginal responses. The two models are difficult to distinguish in the middle of the distribution. In that range both predict intermediate values, but for different structural reasons. They are most sharply distinguished at the extremes. The estimates approach or reach the lower bound of zero convincingly for large firms. They move substantially toward the upper bound of one for the smallest firms. This is the pattern the binary model predicts and the smooth model does not.

6.2.3 Cross-Study Evidence

The convergence pattern in Khwaja and Mian is replicated across the broader set of studies, albeit with coarser partitions. The available evidence on the extreme subgroups is summarized here.

At the unconstrained end, several studies report effects indistinguishable from zero for their least-constrained subsample. Chaney et al. (2012) estimate a coefficient of 0.01 with a standard error of 0.03 for firms with bond market access ($t = 0.33$). Chodorow-Reich (2014) finds effects on employment at large, transparent firms that are indistinguishable from zero. Campello et al. (2010) report that unconstrained CFOs planned capital spending cuts of only 0.4%, compared to 9% for constrained CFOs. In each case, the least-constrained subsample shows an effect that is economically and statistically negligible. Under the smooth model, these firms have valuable investment opportunities and should have at least a small positive response to financing shocks.

At the constrained end, Lemmon and Roberts (2010) report a near-unity estimate ($\hat{\tau} \approx 0.97$) for a sample consisting entirely of below-investment-grade firms. Rauh's lowest-rated subsample shows larger effects than his full sample, moving toward one. Almeida et al. (2012) find the largest effects for unrated and small firms. In each case, the most constrained subsample produces estimates approaching the upper binary bound.

The direction of heterogeneity is also informative as a consistency check. In every study surveyed in Table 4, the more constrained subsample has a larger treatment effect than the less con-

strained subsample. The probability of this pattern arising by chance across nine independent studies, if the direction were equally likely to go either way, is $2^{-9} < 0.002$. Both the binary and smooth models predict this ordering, so it does not distinguish between them. But it strongly supports both models against unstructured alternatives where financing constraints play no systematic role.

6.2.4 What a Definitive Test Would Require

The ideal test of convergence to binary bounds would take a single study's detailed data, progressively refine the partition by constraint-relevant observables, and track whether subgroup estimates converge to $\{0, 1\}$ or stabilize at interior values. Khwaja and Mian's decile analysis approximates this exercise, and the results favor convergence. But ten deciles may not be fine enough to be decisive, particularly if constraint status varies within deciles along dimensions other than size.

A fully satisfying implementation would proceed as follows. Start with a coarse partition (e.g., above versus below median size) and estimate treatment effects for each subgroup. Then refine the partition by adding a second dimension (e.g., cash holdings or leverage) and re-estimate. Continue refining, adding credit ratings, industry controls, and other determinants of constraint status. At each step, record whether the extreme subgroup estimates move toward the binary bounds or stabilize. Under the binary model, the finest subgroups (small, low-rated, low-cash, high-leverage firms) should produce estimates near one. The coarsest unconstrained subgroup (large, investment-grade, high-cash firms) should produce estimates near zero. Under the smooth model, even the finest subgroups should produce interior estimates reflecting the smooth variation in marginal responses.

This exercise requires access to the underlying data, estimation of treatment effects at fine partition levels, and sufficient sample sizes within each cell to produce precise estimates. These requirements are not met by published tables and figures, which typically report at most two or three subsample splits. We therefore regard the convergence test as a promising avenue for future work rather than a test that can be implemented definitively from existing publications.

The evidence currently available is nonetheless informative. The qualitative pattern across studies is exactly what the binary model predicts. The effects are near zero for unconstrained subsamples, effects approaching unity for constrained subsamples, and monotonic variation in between. The smooth model can accommodate larger effects for more constrained firms, but it

cannot easily explain why the extreme subsamples approach the binary bounds of zero and one rather than stabilizing at interior values. This is the pattern most difficult to reconcile with the smooth alternative.

6.2.5 About Variance Based Diagnostics

A natural alternative approach would decompose the variance of treatment effects into within-group and between-group components and check whether the components match the Bernoulli benchmark $\bar{\theta}_C(1-\bar{\theta}_C)$. As we show in Appendix D, this approach has no power. When the within-group variance is computed from subgroup means using the Bernoulli formula $\hat{\tau}_g(1-\hat{\tau}_g)$, the resulting decomposition is an algebraic identity that holds for any distribution of individual effects. The overstatement of within-group variance under the smooth model is exactly offset by the corresponding overstatement of total variance. The decomposition cannot distinguish the two models.

A variance-based test with genuine power would compare reported subgroup standard errors against the Bernoulli prediction $\sqrt{\hat{\tau}_g(1-\hat{\tau}_g)/N_g}$. Under the binary model, reported standard errors should be consistent with this formula. Under the smooth model, they should be systematically smaller because the within-group dispersion of individual effects is less than the Bernoulli benchmark (Appendix D, equation 61). In practice, this test is difficult to implement because reported standard errors in instrumental variables and difference-in-differences settings incorporate estimation uncertainty from the first stage, clustering adjustments, and other sources of variation beyond the sampling variance of individual treatment effects. The test is conservative. Inflated standard errors cannot reject the binary model, but standard errors systematically and substantially below the Bernoulli benchmark would provide evidence against it. The few cases where both standard errors and sample sizes are available for the same subgroup specification yield variance ratios that are consistent with, but not decisively supportive of, the Bernoulli prediction.

Table 5 collects the predictions that distinguish the two models.

The convergence-to-bounds prediction is the most informative diagnostic because it is where the two models diverge most sharply. The cleanest available evidence comes from the within-study variation in Khwaja and Mian (2008), where ten size deciles estimated under a single identification strategy show a pattern that matches the binary model's prediction. There are near-zero effects at the top and effects approaching the upper bound at the bottom, with smooth variation

Table 5: Distinguishing Predictions: Binary versus Smooth Model

Prediction	Binary model	Smooth model
Subgroup estimates under fine conditioning	Converge to $\{0, 1\}$	Stabilize at interior values
Effect for unconstrained subsamples	Equals zero	Positive (small)
Effect for constrained subsamples	Approaches one	High but < 1
Direction of heterogeneity across subsamples	Larger for more constrained	Larger for more constrained

Notes: The first three rows distinguish the two models; the fourth does not. Under the binary model, individual treatment effects are $\{0, 1\}$, so subgroup means converge to the bounds with fine conditioning. Under the smooth model, individual effects vary continuously, producing interior subgroup means even under fine conditioning. Both models predict larger effects for more constrained subsamples. The available evidence favors the binary model on the first three predictions (see Table 3 and the cross-study evidence discussed in the text) and is consistent with both models on the fourth.

in between. The cross study evidence reinforces this pattern. Across nine independent studies spanning multiple countries, decades, and identification strategies, the finest available subsample estimates cluster near the binary bounds rather than spreading continuously across the unit interval.

A definitive formal test requires data to construct arbitrarily fine partitions and track convergence directly. The qualitative evidence assembled here is consistent with the binary model and difficult to reconcile with the smooth alternative. But we recognize that "difficult to reconcile" is not the same as "formally rejected," and we view detailed data based convergence tests as a productive direction for future research.

6.3 Large Shocks and Endogenous Constraint Status

The baseline model treats constraint status as fixed. A firm is either constrained or unconstrained, and the financing shock does not change which regime applies. This is appropriate when shocks are small relative to firms' financing slack, but may not hold when large shocks push previously unconstrained firms across the constraint boundary. This subsection extends the framework to accommodate regime switching and shows that the binary model is the correct local approximation for small shocks, with bounded and interpretable deviations for large shocks.

6.3.1 Three Types of Compliers

Consider a complier with initial constraint slack $S \equiv W + D - I^{unc}$ who is exposed to a financing shock of magnitude $\Delta > 0$. Three cases arise, depending on the firm's position relative to the constraint boundary.

Always-constrained ($S < 0$). The firm was constrained before the shock and remains constrained afterward. Investment falls dollar-for-dollar:

$$\Delta I = -\Delta, \qquad MR = 1. \tag{22}$$

Always-unconstrained ($S \geq \Delta$). The firm's slack exceeds the shock. It remains at its unconstrained optimum and investment does not change:

$$\Delta I = 0, \qquad MR = 0. \tag{23}$$

Switchers ($0 \leq S < \Delta$). The firm was unconstrained before the shock but is pushed across the constraint boundary. The first S dollars of the shock are absorbed by slack; the remainder reduces investment:

$$\Delta I = -(\Delta - S), \qquad MR = \frac{\Delta - S}{\Delta} \in (0, 1). \tag{24}$$

The marginal response is strictly between zero and one, reflecting partial absorption by the firm's pre-shock financing buffer.

6.3.2 Modified LATE Decomposition

Let $\theta_{AC} \equiv P(S < 0 \mid C)$, $\theta_{SW} \equiv P(0 \leq S < \Delta \mid C)$, and $\theta_{AU} \equiv P(S \geq \Delta \mid C)$ denote the shares of always-constrained, switcher, and always-unconstrained compliers, respectively, where C denotes the complier population. The LATE becomes

$$\begin{aligned} \tau_{LATE} &= \theta_{AC} \cdot 1 + \theta_{SW} \cdot E\left[\frac{\Delta - S}{\Delta} \;\middle|\; 0 \leq S < \Delta,\, C \right] + \theta_{AU} \cdot 0 \\ &= \theta_{AC} + \theta_{SW} \cdot \bar{\lambda}_{SW} \end{aligned} \tag{25}$$

where $\bar{\lambda}_{SW} \equiv E[(\Delta - S)/\Delta \mid 0 \leq S < \Delta, C] \in (0, 1)$ is the average marginal response among switchers. Switchers contribute a positive but fractional response, lying between the always-

constrained response of one and the always-unconstrained response of zero.

Proposition 3 (LATE Under Large Shocks)

(a) ***Local approximation.*** *Suppose the distribution of S conditional on being a complier has a bounded density $f_S(\cdot)$ in a neighborhood of zero. Then as $\Delta \to 0$,*

$$\theta_{SW} = P(0 \leq S < \Delta \mid C) = f_S(0) \cdot \Delta + o(\Delta) \to 0 \tag{26}$$

and

$$\tau_{LATE} \to \theta_{AC} = P(S < 0 \mid C) = \theta_C. \tag{27}$$

The binary decomposition is exact in the limit of small shocks.

(b) ***Bounds for finite shocks.*** *For any $\Delta > 0$, since $\bar{\lambda}_{SW} \in (0, 1)$,*

$$\theta_{AC} < \tau_{LATE} < \theta_{AC} + \theta_{SW}. \tag{28}$$

The LATE lies strictly between the pre-shock constraint share θ_{AC} and the post-shock constraint share $\theta_{AC} + \theta_{SW}$.

Proof. Part (a): $\theta_{SW} = \int_0^\Delta f_S(s)\, ds = f_S(0) \cdot \Delta + o(\Delta) \to 0$ as $\Delta \to 0$ by the bounded density assumption. The switcher contribution $\theta_{SW} \cdot \bar{\lambda}_{SW} \leq \theta_{SW} \to 0$, so $\tau_{LATE} \to \theta_{AC}$. In the limit, $\theta_{AC} = P(S < 0 \mid C) = \theta_C$ because there are no switchers. Part (b): $\theta_{SW} \cdot \bar{\lambda}_{SW} > 0$ because $\theta_{SW} > 0$ and $\bar{\lambda}_{SW} > 0$, giving the lower bound. $\theta_{SW} \cdot \bar{\lambda}_{SW} < \theta_{SW}$ because $\bar{\lambda}_{SW} < 1$, giving the upper bound. □

Part (a) establishes that the binary model is not a knife-edge assumption. It is the correct first-order approximation whenever the financing shock is small relative to the distribution of slack among compliers. Part (b) characterizes the direction and magnitude of the approximation error for large shocks. The binary model, which interprets τ_{LATE} as the constraint share, overstates θ_{AC} (the pre-shock share) but understates $\theta_{AC} + \theta_{SW}$ (the post-shock share). The approximation error is bounded by θ_{SW}, which is the fraction of compliers near the constraint boundary.

6.3.3 Robustness of the Empirical Applications

The two main applications in Section 4 are robust to the large-shock concern for different reasons.

Rauh (2006): Small shocks. Mandatory pension contributions are a relatively modest fraction of total financing capacity for most firms in Rauh's sample. The shock Δ is small relative to the support of S, so the switcher population θ_{SW} is thin. The binary approximation $\tau_{LATE} \approx \theta_C$ is accurate because few firms are pushed across the constraint boundary by the pension funding shock. The heterogeneity by credit rating has larger effects for lower-rated firms. This is consistent with variation in θ_{AC} across subgroups rather than with a substantial switcher population.

Lemmon and Roberts (2010): Deeply constrained firms. The junk bond market collapse is a large shock, but the treated population consists of below-investment-grade firms with S well below zero. A large negative shock applied to already constrained firms does not create switchers. It deepens existing constraints, and the marginal response remains one. The relevant concern would arise if the shock were applied to a population of firms clustered near $S = 0$, but this is precisely the population that Lemmon and Roberts exclude by focusing on junk-rated borrowers. The near-unity estimate is robust because the vast majority of treated firms are always constrained, not switchers.

More generally, the large shock concern is most severe for studies that apply substantial financing shocks to populations spanning the constraint boundary. These are firms with moderate slack that could plausibly be pushed into constrained territory. In such settings, the binary model overstates the pre-shock constraint share. The true LATE reflects a mixture of always constrained firms (with $MR = 1$), switchers (with $MR \in (0, 1)$), and always unconstrained firms (with $MR = 0$).

6.3.4 A Testable Prediction

The extended framework generates a prediction that the baseline binary model does not. If two instruments of different magnitudes $\Delta_1 < \Delta_2$ are applied to the same population, the larger shock should produce a weakly larger LATE:

$$\tau(\Delta_2) \geq \tau(\Delta_1) \tag{29}$$

because the larger shock converts a wider band of unconstrained firms into switchers. This is not a change in the structural response; it reflects the mechanical expansion of the switcher population.

Under the binary model with fixed constraint status, the LATE is instrument-invariant: $\tau(\Delta_1) = \tau(\Delta_2) = \theta_C$. Detecting a systematic positive relationship between instrument magnitude and es-

timated treatment effects, conditional on the same complier population, would indicate the presence of switchers and the empirical relevance of endogenous constraint status. Conversely, finding no such relationship supports the binary approximation.

Testing this prediction requires either a single setting with continuous variation in shock magnitude or multiple instruments of known different strengths applied to overlapping populations. These conditions are rarely met in existing studies, but the prediction provides a framework for future research designs. In particular, studies with continuous instruments could estimate the LATE as a function of Δ and test whether it is approximately constant (supporting the binary model) or increasing (indicating switchers). An example might be variation in the size of collateral value changes.

6.3.5 Connection to the Distribution of Constraint Slack

If one could vary Δ continuously and estimate $\tau(\Delta)$ at each value, the resulting function would trace out information about the cumulative distribution of slack among compliers. From equation (25), $\tau(\Delta)$ is increasing in Δ, and its behavior as Δ grows reveals the density of firms at successive distances from the constraint boundary. At $\Delta = 0$, one identifies $P(S < 0 \mid C)$, the point mass below the boundary. As Δ increases, one integrates progressively more of the density above zero, accumulating the constraint shares that would obtain under successively larger shocks. In the limit as $\Delta \to \infty$, $\tau(\Delta) \to 1$: a sufficiently large shock constrains all firms.

This observation connects directly to the marginal treatment effect framework of Heckman and Vytlacil (2005). The MTE curve in our model reflects the probability of being constrained at a given quantile of unobserved resistance to the instrument. Or equivalently, the density of constraint slack at the corresponding margin. Varying Δ traces out the MTE curve, revealing the structural distribution of financial slack among compliers. In practice, the discrete nature of most natural experiments in corporate finance limits this exercise. But the connection clarifies what additional variation would be needed to move beyond the binary approximation, and confirms that the binary model's constraint share θ_C is the well defined limiting case. It is the value of the MTE distribution function evaluated at the point of zero slack.

6.4 Robustness

The constraint interpretation does not require the binary model to hold exactly. Suppose instead that constrained firms have marginal response MR_C and unconstrained firms have MR_U, where $MR_C > MR_U$ but neither is necessarily one or zero. The treatment effect becomes

$$\hat{\tau} = \theta_C \cdot MR_C + (1 - \theta_C) \cdot MR_U \tag{30}$$

and the qualitative predictions survive. Effects are larger for more constrained subsamples, and aggregate estimates reflect composition-weighted averages of the two structural responses. The empirical finding that effects approach zero for unconstrained firms and unity for constrained firms pins down $MR_U \approx 0$ and $MR_C \approx 1$.

6.5 Limitations

Several limitations should be noted. Our model is static and abstracts from precautionary savings, multi-period dynamics, and endogenous debt capacity. Richer models with smooth financing costs would generate marginal responses that vary continuously with constraint severity rather than discretely.

The binary characterization of constraint status is a simplification. Firms near the boundary where $S \approx 0$ may have intermediate responses, and large financing shocks could shift constraint status itself.

Our interpretation of specific estimates as constraint shares depends on the assumption that constrained firms cannot substitute across financing sources. Lemmon and Roberts (2010) find limited substitution in their setting, but this need not hold generally. When substitution is possible, the structural marginal response for constrained firms is less than one, and the treatment effect identifies a weighted average of responses rather than a constraint share directly.

We take constraint status as exogenous when characterizing marginal responses. The IV assumptions require only that the financing shock is exogenous, not that constraint status itself is exogenous. But if the instrument shifts firms across the constraint boundary, the LATE reflects responses averaged over firms whose constraint status may be changing, complicating the clean decomposition in Proposition 2.

These limitations suggest directions for future work. Estimating the continuous marginal treatment effect function $MTE(u)$ would provide a richer characterization than the binary model al-

lows. Combining credible identification with systematic heterogeneity analysis is already standard practice in many of the papers we survey. It remains the most productive strategy for recovering structural parameters from treatment effects estimates. Explicitly formalizing the underlying economic model helps crystallize the meaning of our estimates.

7 Conclusion

The central message is simple. Treatment effect estimates of financing on investment recover constraint shares among compliers, not structural marginal responses. In the canonical model, constrained firms have a marginal response of one and unconstrained firms have a marginal response of zero. An intermediate estimate therefore reflects a mixture of the two types. Cross study differences in magnitudes are due to variation in complier composition. Once this mapping is recognized, the apparently contradictory findings of Rauh (2006), Bakke and Whited (2012), and Lemmon and Roberts (2010) become coherent, and the broader pattern across the quasi-experimental literature falls into place.

The analysis suggests practical recommendations. A treatment effect strictly between zero and one should be read as evidence of population heterogeneity. It is not an estimate of a structural parameter.

Heterogeneity analysis by constraint proxies is not merely a robustness exercise. It is the primary tool for recovering structural content. Subsamples with effects near one identify the structural marginal response. Subsamples with effects near zero confirm that unconstrained firms do not respond.

Generalizing an estimate to a new population requires explicit accounting for constraint shares. A policy targeting investment grade firms would likely produce effects near zero. A policy targeting small, cash poor firms would produce effects closer to one. The underlying structural relationship is identical in both cases.

Several limitations deserve emphasis. The model is static and treats constraint status as given. We show that the binary specification remains the correct local approximation for small shocks and that precautionary savings are empirically small. They are roughly three cents per dollar in the Lemmon and Roberts (2010). Richer models with smooth financing costs would instead generate continuously varying marginal responses. The interpretation also relies on limited substitution across financing sources. Lemmon and Roberts (2010) document that this condition holds

in their setting. But there is no guarantee that it will hold generally in other contexts. When substitution is material the treatment effect identifies a weighted average of (now-smaller) responses rather than a pure constraint share.

The framework extends naturally to other real outcomes that respond to financing shocks through binding constraints. These include cash holdings (Almeida et al., 2004), employment (Chodorow-Reich, 2014), and collateral driven investment (Chaney et al., 2012). Productive directions for future work include estimating constraint shares directly, implementing the convergence to bounds diagnostic with sufficiently fine partitions, and integrating the static framework with dynamic models that endogenize constraint status.

References

Almeida, H. and M. Campello (2007). Financial constraints, asset tangibility, and corporate investment. *Review of Financial Studies 21*(5), 1429–1460.

Almeida, H., M. Campello, B. Laranjeira, and S. Weisbenner (2012). Corporate debt maturity and the real effects of the 2007 credit crisis. *Critical Finance Review 1*(1), 3–58.

Almeida, H., M. Campello, and M. S. Weisbach (2004). The cash flow sensitivity of cash. *Journal of Finance 59*(4), 1777–1804.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Bakke, T.-E. and T. M. Whited (2012). Threshold events and identification: A study of cash shortfalls. *The Journal of Finance 67*(3), 1083–1111.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2022). When is tsls actually late? Technical report, https://a-torgovitsky.github.io/tslslate.pdf.

Campello, M., J. R. Graham, and C. R. Harvey (2010). The real effects of financial constraints: Evidence from a financial crisis. *Journal of Financial Economics 97*(3), 470–487.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011). Estimating marginal returns to education. *American Economic Review 101*(6), 2754–2781.

Chaney, T., D. Sraer, and D. Thesmar (2012). The collateral channel: How real estate shocks affect corporate investment. *American Economic Review 102*(6), 2381–2409.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2025). Fisher–schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. *Econometrica 93*(4), 1121–1164.

Chodorow-Reich, G. (2014). The employment effects of credit market disruptions: Firm-level evidence from the 2008–9 financial crisis. *The Quarterly Journal of Economics 129*(1), 1–59.

Duchin, R., O. Ozbas, and B. A. Sensoy (2010). Costly external finance, corporate investment, and the subprime mortgage credit crisis. *Journal of Financial Economics 97*(3), 418–435.

Erickson, T. and T. M. Whited (2000). Measurement error and the relationship between investment and q. *Journal of political economy 108*(5), 1027–1057.

Farre-Mensa, J. and A. Ljungqvist (2016). Do measures of financial constraints measure financial constraints? *The review of financial studies 29*(2), 271–308.

Fazzari, S. M., R. G. Hubbard, and B. C. Petersen (1988). Financing constraints and corporate investment. *Brookings Papers on Economic Activity 1988*(1), 141–206.

Gan, J. (2007). Collateral, debt capacity, and corporate investment: Evidence from a natural experiment. *Journal of Financial Economics 85*(3), 709–734.

Goldsmith-Pinkham, P. (2024). Tracking the credibility revolution across fields. Technical report, arXiv preprint arXiv:2405.20604.

Haile, P. A. (2025). Models, measurement, and the language of empirical economics. Technical report, https://sites.google.com/view/philhaile/home/teaching.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of Econometrics 6*, 4779–4874.

Hennessy, C. A. and T. M. Whited (2007). How costly is external financing? evidence from a structural estimation. *Journal of Finance 62*(4), 1705–1745.

Hoberg, G. and V. Maksimovic (2015). Redefining financial constraints: A text-based analysis. *The Review of Financial Studies 28*(5), 1312–1352.

Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature 48*(2), 399–423.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Kaplan, S. N. and L. Zingales (1997). Do investment-cash flow sensitivities provide useful measures of financing constraints? *The Quarterly Journal of Economics 112*(1), 169–215.

Khwaja, A. I. and A. Mian (2008). Tracing the impact of bank liquidity shocks: Evidence from an emerging market. *American Economic Review 98*(4), 1413–1442.

Lemmon, M. and M. R. Roberts (2010). The response of corporate financing and investment to changes in the supply of credit. *Journal of Financial and Quantitative Analysis 45*(3), 555–587.

Mogstad, M. and A. Torgovitsky (2024). Instrumental variables with unobserved heterogeneity in treatment effects. In *Handbook of Labor Economics*, Volume 5, pp. 1–114. Elsevier.

Peek, J. and E. S. Rosengren (2000). Collateral damage: Effects of the Japanese bank crisis on real activity in the United States. *American Economic Review 90*(1), 30–45.

Rauh, J. D. (2006). Investment and financing constraints: Evidence from the funding of corporate pension plans. *The Journal of Finance 61*(1), 33–71.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers 3*(2), 135–146.

Strebulaev, I. A. (2007). Do tests of capital structure theory mean what they say? *The Journal of Finance 62*(4), 1747–1787.

Whited, T. M. and G. Wu (2006). Financial constraints risk. *Review of Financial Studies 19*(2), 531–559.

A Appendix: Connection to the Roy Model and Marginal Treatment Effects

This appendix formalizes the connection between our structural model of financial constraints and the econometric literature on heterogeneous treatment effects stemming from Roy (1951) and Heckman and Vytlacil (2005). We show that our framework is a special case of the marginal treatment effect (MTE) framework, where the corner solution in optimal investment generates a step-function MTE.

A.1 The MTE Framework

In the generalized Roy model of Heckman and Vytlacil (2005), treatment selection is represented as $D_i = \mathbf{1}[\mu(Z_i) \geq U_i]$, where Z_i is an instrument, $\mu(\cdot)$ is the propensity score, and $U_i \in [0,1]$ indexes unobserved resistance to treatment. The marginal treatment effect,

$$MTE(u) = E[Y_{1i} - Y_{0i} \mid U_i = u], \tag{31}$$

represents the average treatment effect for individuals at the u-th quantile of resistance. The central result of Heckman and Vytlacil (2005) is that LATE can be expressed as a weighted average of the MTE:

$$LATE = \int_0^1 MTE(u) \cdot \omega(u)\, du \tag{32}$$

where the weights $\omega(u)$ depend on which individuals are induced to change treatment status by the instrument. Different instruments identify different weighted averages because they move different complier populations.

A.2 Correspondence to Our Model

Our model maps directly into this framework. Define treatment as receiving an additional dollar of financing capacity. The individual treatment effect is

$$ITE_i = \frac{\partial I_i^*}{\partial F} = \begin{cases} 1 & \text{if } S_i < 0 \quad \text{(constrained)} \\ 0 & \text{if } S_i \geq 0 \quad \text{(unconstrained)} \end{cases} \tag{33}$$

where $S_i = W_i + D_i - I_i^{unc}$ is constraint slack. The marginal treatment effect is

$$MTE(u) = E\left[\frac{\partial I^*}{\partial F}\,\middle|\, U = u\right] = P(S < 0 \mid U = u) \tag{34}$$

which equals the probability that firms at quantile u of unobserved resistance are financially constrained. This is equation (14) in the main text.

A.3 The Step-Function MTE

The corner solution in optimal investment generates a step-function MTE. This is a clean special case of the general framework. Because constrained firms have $\partial I^*/\partial F = 1$ and unconstrained firms have $\partial I^*/\partial F = 0$, there exists a threshold u^* such that

$$MTE(u) = \begin{cases} 1 & \text{if } u \leq u^* \\ 0 & \text{if } u > u^* \end{cases} \tag{35}$$

where u^* is determined by the population distribution of constraint slack S. The LATE decomposition then simplifies to

$$LATE = \int_0^{u^*} 1 \cdot \omega(u)\, du + \int_{u^*}^1 0 \cdot \omega(u)\, du = \int_0^{u^*} \omega(u)\, du = \theta_C. \tag{36}$$

The integral of complier weights below the threshold equals the probability of being constrained among compliers. This is Proposition 2.

This binary structure eliminates the continuous variation in treatment effects that typically characterizes the MTE. It arises directly from the corner solution. When the financing constraint binds, additional financing translates dollar-for-dollar into investment; when the constraint is slack, it has zero effect. The discontinuity in the investment function at $S = 0$ maps into a discontinuity in the MTE at u^*.

A.4 Selection on Constraint Status

In the standard Roy model, individuals have positive selection on gains. Those with high treatment effects are more likely to select into treatment. Our model generates an analogous pattern. Firms with strong investment opportunities relative to financing capacity ($S < 0$) have both higher treatment effects ($\partial I^*/\partial F = 1$) and lower resistance to treatment. They are more likely to be compliers

because financing shocks are more consequential for them.

Unlike the standard Roy model, however, firms do not choose to be constrained. Constraint status is an outcome of the interaction between investment opportunities and available financing. The selection mechanism is indirect. The instrument affects firms with strong investment opportunities (low u), who are also more likely constrained. The MTE is weakly decreasing in u, consistent with the negative selection pattern described by Heckman and Vytlacil (2005).

A.5 Continuous Treatment Effects

The binary treatment effect is a simplifying assumption. With smooth costs of external finance, the marginal response would vary continuously with constraint severity, generating a smooth, decreasing MTE with $MTE(u) \approx 1$ for low u (severely constrained) and $MTE(u) \approx 0$ for high u (unconstrained). The qualitative insight survives. The LATE remains a weighted average that understates the response for the most constrained firms and overstates it for the least constrained.

Estimating the continuous MTE requires either continuous variation in treatment intensity or a rich set of instruments affecting different margins of the u distribution (Carneiro et al., 2011). These conditions are rarely met in corporate finance, where most natural experiments involve discrete policy changes. The binary approximation thus reflects both theoretical parsimony and practical data limitations.

B Appendix: Precautionary Savings and the Identification of Constraint Shares

The baseline model assumes that constrained firms invest all available financing, yielding a marginal response of exactly one. If constrained firms instead retain a fraction of marginal financing as precautionary cash, the marginal response falls below one and a single treatment effect estimate no longer identifies the constraint share. This appendix extends the framework to accommodate precautionary savings, shows how cross-study variation restores identification, and demonstrates that the data discipline the precautionary parameter to be small in the settings we study.

B.1 Modified Model

Suppose constrained firms allocate a fraction $\alpha \in [0, 1)$ of each marginal dollar of financing to precautionary cash reserves rather than investment. Optimal investment becomes

$$I^* = \begin{cases} I^{unc} & \text{if } S \geq 0 \quad \text{(unconstrained)} \\ (1-\alpha)(W+D) & \text{if } S < 0 \quad \text{(constrained)} \end{cases} \tag{37}$$

where α captures the shadow value of liquidity for constrained firms facing future financing uncertainty. The structural marginal response is now

$$MR = \begin{cases} 0 & \text{if } S \geq 0 \\ 1-\alpha & \text{if } S < 0. \end{cases} \tag{38}$$

The baseline model is nested as the special case $\alpha = 0$.

B.2 Modified LATE Decomposition

The LATE for a population with constraint share θ_C among compliers becomes

$$\tau_{LATE} = \theta_C \cdot (1-\alpha) + (1-\theta_C) \cdot 0 = \theta_C(1-\alpha). \tag{39}$$

A single treatment effect estimate no longer pins down θ_C. For example, $\hat{\tau} = 0.65$ is consistent with $\theta_C = 0.65$ and $\alpha = 0$ (baseline interpretation), or $\theta_C = 0.81$ and $\alpha = 0.20$, or $\theta_C = 1$ and $\alpha = 0.35$. The constraint share and the precautionary parameter are not separately identified from a single estimate.

B.3 Identification from Multiple Populations

Cross-study or cross-subsample variation restores identification. Suppose we observe treatment effects from two populations $j = 1, 2$ with potentially different constraint shares but a common precautionary parameter:

$$\hat{\tau}_j = \theta_{C,j}(1-\alpha), \quad j = 1, 2. \tag{40}$$

The assumption that α is common across populations requires that precautionary behavior depends on the nature of financing frictions rather than on the specific sample. This is appropriate

when comparing subsamples within a given institutional environment, though it may be less defensible across settings with very different financing structures.

Proposition 4 (Identification with Precautionary Savings)

(a) ***Relative constraint shares.*** *The ratio of treatment effects identifies relative constraint shares without knowledge of α:*

$$\frac{\hat{\tau}_1}{\hat{\tau}_2} = \frac{\theta_{C,1}}{\theta_{C,2}}. \tag{41}$$

(b) ***Anchor population.*** *If one population is known to be entirely constrained ($\theta_{C,2} = 1$), its treatment effect identifies the precautionary parameter directly:*

$$\hat{\alpha} = 1 - \hat{\tau}_2, \tag{42}$$

and the constraint share in any other population is recovered as

$$\hat{\theta}_{C,1} = \frac{\hat{\tau}_1}{\hat{\tau}_2}. \tag{43}$$

Proof. Part (a) follows immediately from dividing equation (40) for $j = 1$ by the same equation for $j = 2$. The $(1-\alpha)$ terms cancel. Part (b) substitutes $\theta_{C,2} = 1$ into equation (40) to obtain $\hat{\tau}_2 = 1-\alpha$, which gives (42). Substituting into the expression for population 1 yields $\hat{\tau}_1 = \theta_{C,1} \cdot \hat{\tau}_2$, from which (43) follows. □

B.4 Application: Lemmon and Roberts as Anchor

Lemmon and Roberts (2010) provide a natural anchor population. Their treated firms are exclusively below-investment-grade, and the authors document "almost no substitution" to alternative financing sources and no increase in cash holdings among treated firms. If we take $\theta_{C,junk} \approx 1$, their near-unity estimate identifies precautionary savings:

$$\hat{\tau}_{LR} \approx 0.97 \quad \Longrightarrow \quad \hat{\alpha} \approx 1 - 0.97 = 0.03. \tag{44}$$

Precautionary savings are negligible as they are approximately three cents per dollar of the financing shock. This is consistent with Lemmon and Roberts' direct evidence that treated firms did not accumulate cash. In a severe credit contraction affecting firms already at the edge of their

Table 6: Constraint Shares with and without Precautionary Savings

Study	$\hat{\tau}$	$\hat{\theta}_C$ (baseline, $\alpha = 0$)	$\hat{\theta}_C$ (adjusted, $\hat{\alpha} = 0.03$)
Lemmon-Roberts (2010)	0.97	0.97	1.00 (anchor)
Rauh (2006)	0.65	0.65	0.67
Khwaja-Mian (2008)	0.60	0.60	0.62
Chaney et al. (2012)	0.06	0.06	0.06

Notes: Baseline constraint shares assume $MR_C = 1$ ($\alpha = 0$), so $\hat{\theta}_C = \hat{\tau}$. Adjusted shares use $\hat{\alpha} = 0.03$ from the Lemmon and Roberts anchor, so $\hat{\theta}_C = \hat{\tau}/(1 - \hat{\alpha}) = \hat{\tau}/0.97$. The adjustment is small because precautionary savings are empirically negligible in these settings.

financing capacity, the shadow value of future liquidity is dominated by the immediate need for current investment.

Using this anchor, Rauh's constraint share is

$$\hat{\theta}_{C,Rauh} = \frac{\hat{\tau}_{Rauh}}{\hat{\tau}_{LR}} = \frac{0.65}{0.97} \approx 0.67. \tag{45}$$

This is virtually identical to the baseline estimate of $\hat{\theta}_C = 0.65$ obtained under $\alpha = 0$. The reason is straightforward. When precautionary savings are small, the correction is small. The baseline framework is a good approximation.

Table 6 reports the adjusted constraint shares for the studies analyzed in the main text. In every case, the adjustment is modest: constraint shares increase by approximately three percentage points relative to the baseline estimates. The qualitative conclusions are unchanged.

B.5 Bounds When No Anchor Population Is Available

When no population can be assumed entirely constrained, partial identification is still possible. Since $\alpha \geq 0$ and $\theta_C \leq 1$, equation (39) implies

$$\theta_C \geq \hat{\tau} \qquad \text{and} \qquad \alpha \leq 1 - \hat{\tau}. \tag{46}$$

The treatment effect is a lower bound on the constraint share and an upper bound on the precautionary rate. For Rauh's estimate, $\theta_C \geq 0.65$ and $\alpha \leq 0.35$.

These bounds tighten with subsample estimates. Suppose that the most constrained subsample (small size, low ratings, or low cash) has treatment effect $\hat{\tau}_{high}$. Assume that $\theta_{C,high} \approx 1$. Then $\hat{\alpha} \leq 1 - \hat{\tau}_{high}$. For instance, if the lowest rated subsample in Rauh's data has $\hat{\tau}_{low\text{-}rated} = 0.84$, then

$\alpha \leq 0.16$, which tightens the full sample constraint share to

$$\hat{\theta}_C \geq \frac{0.65}{1} = 0.65 \qquad \text{and} \qquad \hat{\theta}_C \leq \frac{0.65}{1 - 0.16} = 0.77. \tag{47}$$

The constraint share lies between 65% and 77%. Additional subsample estimates narrow the range further.

B.6 Discussion

Three features of this extension are of note. The precautionary parameter α may itself vary across settings. During a severe credit contraction such as the junk bond collapse studied by Lemmon and Roberts, firms face acute financing needs and cannot afford to stockpile cash; α should be near zero. In less severe episodes, firms with ongoing financing uncertainty may retain more precautionary liquidity, implying a larger α. The assumption of a common α is most defensible when comparing subsamples within a single study, where firms face similar macroeconomic conditions and institutional environments.

The extension preserves the paper's central insight. Whether $\alpha = 0$ or $\alpha = 0.03$ or $\alpha = 0.10$, treatment effects remain weighted averages of heterogeneous structural responses, and the weights depend on complier composition. The interpretation shifts slightly. It goes from "the estimate equals the constraint share" to "the estimate equals the constraint share times the net marginal response". But the qualitative implications for empirical practice are unchanged. Researchers should still pay careful attention to complier composition, interpret intermediate estimates as reflecting heterogeneity, and use subsample analysis to recover structural parameters.

The empirical evidence disciplines α to be small. The near unity estimates in the most constrained populations (Lemmon and Roberts at 0.97, the smallest firms in Khwaja and Mian at 0.6–0.8) leave little room for precautionary savings. If α were larger, say 0.20 or higher, we would never observe treatment effects approaching one, even in predominantly constrained samples. The fact that near unity effects appear repeatedly provides direct evidence that the baseline binary model, with $MR_C = 1$ and $\alpha = 0$, is a good approximation for the settings studied in this paper.

C Internet Appendix: Quantifying Constraint Shares Across the Literature

The framework developed in Sections 2 and 3 provides a precise interpretation of treatment effect estimates: $\tau_{LATE} = \theta_C$, where θ_C is the share of financially constrained firms among compliers. This appendix applies this interpretation to the papers surveyed in Section 5, calculating the implied fraction of constrained versus unconstrained firms in each study's complier population. Table 7 summarizes the results.

C.1 Methodology

Under the structural model, the treatment effect identifies the probability that a complier is financially constrained.

$$\hat{\tau}_{LATE} = \hat{\theta}_C \cdot 1 + (1 - \hat{\theta}_C) \cdot 0 = \hat{\theta}_C. \tag{48}$$

Given an estimated treatment effect $\hat{\tau}$, we directly infer $\hat{\theta}_C = \hat{\tau}$ and $1 - \hat{\theta}_C = 1 - \hat{\tau}$. This calculation requires that the treatment effect be expressed as a marginal response not as an elasticity or percentage change. It should be a change in investment per unit change in financing. For studies reporting elasticities, we note the limitation and focus on the qualitative pattern.

C.2 Study-by-Study Calculations

Consider Khwaja and Mian (2008). The baseline estimate of 0.6 implies that 60% of compliers are constrained. The heterogeneity by firm size provides refined estimates. $\hat{\theta}_C \approx 0\%$ for the top three size deciles. It rises monotonically to 60 to 80% for the smallest deciles (see Table 3 in the main text).

Consider Chodorow-Reich (2014). Effects are significant for small and medium firms but indistinguishable from zero for the largest and most transparent firms. The qualitative pattern implies $\hat{\theta}_C \approx 0\%$ for large firms and substantially positive for small and medium firms, consistent with the binary prediction.

Consider Duchin et al. (2010). Firms in the bottom tercile of cash holdings experienced investment declines roughly twice as large as those in the top tercile. While the paper does not report treatment effects as marginal responses, the pattern indicates that $\hat{\theta}_C$ is substantially higher for low-cash firms. If the bottom tercile has $\hat{\theta}_C \approx 0.6$–$0.7$, the top tercile may have $\hat{\theta}_C \approx 0.3$–$0.35$.

These values are illustrative. The point is that constraint shares vary systematically with pre-crisis financial health.

Consider Campello et al. (2010). Constrained U.S. firms planned to cut capital spending by approximately 9%, compared to 0.4% for unconstrained firms. Among U.S. CFOs who self-identify as constrained, 86% report bypassing attractive investment projects. This provides a benchmark from a different measurement approach. Among firms that view themselves as constrained, the vast majority behave as if constraints bind.

Consider Almeida et al. (2012). Firms with long-term debt maturing right after the crisis onset cut their investment-to-capital ratio by 2.5 percentage points more than matched controls, with larger effects for smaller and unrated firms. The magnitude suggests a substantial fraction of treated firms were constrained. For unrated and small firms in the most affected subsamples, $\hat{\theta}_C$ likely approaches unity.

Consider Gan (2007). A 10% decrease in land value reduces the investment rate by approximately 0.8%, yielding an elasticity of 0.08. Because this is an elasticity rather than a marginal response, direct conversion to $\hat{\theta}_C$ requires additional structure. The finding that effects are larger for high-leverage and low-liquidity firms indicates higher constraint shares in those subsamples.

Consider Chaney et al. (2012). U.S. corporations invest \$0.06 per dollar of real estate collateral. Interpreting this as a marginal response, $\hat{\theta}_C \approx 6\%$. That means only a small fraction of publicly traded firms are constrained with respect to real estate collateral. This low share reflects the sample composition. Large firms have access to diverse financing sources. Effects are concentrated among smaller firms without bond market access, where $\hat{\theta}_C$ is likely considerably higher.

Consider Peek and Rosengren (2000). Markets with greater Japanese bank penetration experienced larger declines in commercial real estate construction. The qualitative finding suggests variation in constraint shares across local markets, with $\hat{\theta}_C$ higher where Japanese banks represented a larger share of credit supply.

C.3 Interpretation

Several patterns emerge from Table 7. Constraint shares vary widely across studies, from 6% in Chaney et al. (2012) to approaching 100% in the most constrained subsamples. This variation reflects differences in sample composition, not different structural parameters. All studies are consistent with a marginal response of one for constrained firms and zero for unconstrained firms.

The within-study heterogeneity consistently shows higher constraint shares for firms with

Table 7: Implied Constraint Shares Among Compliers

Study	Estimate	Implied $\hat{\theta}_C$	Heterogeneity
Khwaja-Mian (2008)	0.60	60%	0% (large) to 60–80% (small)
Chodorow-Reich (2014)	Significant	Substantial for small/med.	$\approx 0\%$ for large
Duchin et al. (2010)	$2\times$ ratio	60–70% (low cash)	30–35% (high cash)
Campello et al. (2010)	9% planned cut	High (survey-based)	86% of constrained CFOs bypassed projects
Almeida et al. (2012)	2.5 ppt decline	Substantial (inferred)	Approaching 100% for unrated
Gan (2007)	Elast. 0.08	Modest (aggregate)	Higher for high lev., low liq.
Chaney et al. (2012)	0.06	6%	Concentrated in small firms
Peek-Rosengren (2000)	Significant	High-exposure mkts.	Varies across local markets

Notes: Implied constraint shares follow from equation (48) under the assumption that $MR_C = 1$ and $MR_U = 0$. Where the original estimate is an elasticity or percentage change rather than a marginal response, direct calculation of $\hat{\theta}_C$ is not possible and we report qualitative patterns. Constraint shares for Rauh (2006) and Lemmon and Roberts (2010) are reported in Table 2 in the main text.

characteristics associated with financing frictions such as small size, low credit ratings, high leverage, low cash holdings, and lack of bond market access. If treatment effects measured marginal responses that varied for reasons unrelated to constraints, there would be no reason for the same firm characteristics to predict larger effects across diverse settings.

The implied constraint shares match economic intuition. Studies focusing on junk-rated firms or firms requiring refinancing during severe credit disruptions find constraint shares approaching unity. Studies using broad samples that include large, highly rated firms find substantial fractions of unconstrained compliers.

These calculations reinforce the main message of the paper. Researchers estimating financing effects on real outcomes should not interpret their coefficient as a structural marginal response. It measures the share of firms for whom financing binds. Mechanically applying an estimate from one population to another produces misleading predictions whenever constraint shares differ. The implied constraint shares in Table 7 provide a basis for making such adjustments explicit.

D Internet Appendix: Why the Variance Decomposition Cannot Distinguish the Models

A natural approach to testing whether individual treatment effects are binary or continuous would decompose the observed variance of subgroup treatment effects into within-group and between-

group components, and then check whether the components are consistent with the Bernoulli benchmark $\bar{\theta}_C(1-\bar{\theta}_C)$. This appendix shows that this approach has no power. When the within-group variance is computed from subgroup means using the Bernoulli formula, the resulting decomposition is an algebraic identity that holds for any distribution of individual effects. We derive this result in detail so that the source of the cancellation is transparent.

D.1 Setup and Notation

Consider a population of firms indexed by i, partitioned into G subgroups indexed by $g = 1, \ldots, G$. Let $\omega_g > 0$ denote the population weight of subgroup g, with $\sum_{g=1}^{G} \omega_g = 1$. Each firm has an individual treatment effect $\tau_i \equiv \partial I_i^* / \partial F$. Define:

$$\tau_g \equiv E[\tau_i \mid i \in g] \qquad \text{(subgroup mean)} \tag{49}$$

$$\bar{\tau} \equiv \sum_{g=1}^{G} \omega_g \tau_g = E[\tau_i] \qquad \text{(population mean)} \tag{50}$$

$$\sigma_g^2 \equiv \text{Var}(\tau_i \mid i \in g) \qquad \text{(within-group variance)} \tag{51}$$

$$\sigma^2 \equiv \text{Var}(\tau_i) \qquad \text{(total variance).} \tag{52}$$

The law of total variance gives the standard decomposition:

$$\sigma^2 = \underbrace{\sum_{g=1}^{G} \omega_g \, \sigma_g^2}_{V_W} + \underbrace{\sum_{g=1}^{G} \omega_g (\tau_g - \bar{\tau})^2}_{V_B} \tag{53}$$

where V_W is the average within-group variance and V_B is the between-group variance. This identity holds for any distribution of τ_i, under both the binary and smooth models. The two models differ in the values of σ^2, σ_g^2, and V_W.

D.2 Variance Components Under the Binary Model

Under the binary model, $\tau_i \in \{0, 1\}$ with $P(\tau_i = 1 \mid i \in g) = \theta_{C,g}$, where $\theta_{C,g}$ is the constraint share in subgroup g. The subgroup mean is $\tau_g = \theta_{C,g}$ and the population mean is $\bar{\tau} = \sum_g \omega_g \theta_{C,g} \equiv \bar{\theta}_C$. Since τ_i is Bernoulli within each subgroup, the within-group variance is

$$\sigma_g^2 = \theta_{C,g}(1 - \theta_{C,g}). \tag{54}$$

To verify this, note that $\tau_i \in \{0,1\}$ implies $\tau_i^2 = \tau_i$, so $E[\tau_i^2 \mid g] = E[\tau_i \mid g] = \theta_{C,g}$ and $\sigma_g^2 = \theta_{C,g} - \theta_{C,g}^2$. The average within-group variance and between-group variance are

$$V_W^{bin} = \sum_{g=1}^{G} \omega_g \, \theta_{C,g}(1 - \theta_{C,g}) \tag{55}$$

$$V_B^{bin} = \sum_{g=1}^{G} \omega_g (\theta_{C,g} - \bar{\theta}_C)^2. \tag{56}$$

The total variance is also Bernoulli at the population level: $\sigma_{bin}^2 = \bar{\theta}_C(1 - \bar{\theta}_C)$. The adding-up identity $V_W^{bin} + V_B^{bin} = \sigma_{bin}^2$ can be verified by expanding the two sums. Expanding V_W^{bin}:

$$\sum_g \omega_g \, \theta_{C,g}(1 - \theta_{C,g}) = \bar{\theta}_C - \sum_g \omega_g \theta_{C,g}^2. \tag{57}$$

Expanding V_B^{bin}:

$$\sum_g \omega_g (\theta_{C,g} - \bar{\theta}_C)^2 = \sum_g \omega_g \theta_{C,g}^2 - \bar{\theta}_C^2. \tag{58}$$

Adding:

$$V_W^{bin} + V_B^{bin} = \bar{\theta}_C - \bar{\theta}_C^2 = \bar{\theta}_C(1 - \bar{\theta}_C) = \sigma_{bin}^2. \tag{59}$$

The terms $\sum_g \omega_g \theta_{C,g}^2$ cancel exactly.

D.3 Variance Components Under the Smooth Model

Under the smooth model, τ_i is continuously distributed on $[0,1]$ with subgroup means $\mu_g \equiv E[\tau_i \mid g]$ and within-group variances $\sigma_g^2 \equiv \text{Var}(\tau_i \mid g) > 0$. The bound is that for any random variable $X \in [0,1]$ with $E[X] = \mu$,

$$\text{Var}(X) \leq \mu(1 - \mu) \tag{60}$$

with equality if and only if $P(X \in \{0,1\}) = 1$. The proof is immediate. $X \in [0,1]$ implies $X^2 \leq X$, so $E[X^2] \leq E[X] = \mu$ and $\text{Var}(X) = E[X^2] - \mu^2 \leq \mu - \mu^2$. Equality requires $X^2 = X$ almost surely, which holds only when $X \in \{0,1\}$.

Under the smooth model, τ_i takes values in the interior of $[0,1]$ with positive probability, so the inequality is strict:

$$\sigma_g^2 < \mu_g(1 - \mu_g) \quad \text{for all } g. \tag{61}$$

The smooth model's within-group variance is strictly less than the Bernoulli formula evaluated at

the same subgroup mean. Summing across subgroups:

$$V_W^{sm} = \sum_g \omega_g \sigma_g^2 < \sum_g \omega_g \mu_g (1 - \mu_g). \tag{62}$$

The between-group variance depends only on the subgroup means and weights, so if the two models produce the same subgroup means ($\mu_g = \theta_{C,g}$), the between-group variance is identical: $V_B^{sm} = V_B^{bin}$.

D.4 The Cancellation Result

The preceding sections establish that the smooth model has strictly less within-group variance than the binary model and strictly less total variance. A diagnostic that compared observed within-group variance against the Bernoulli benchmark would have power if the reduction in within-group variance were larger, or smaller, than the reduction in total variance. This subsection shows that the two reductions are identical, so the naive decomposition has no power.

Define the within-group excess as $\delta_W \equiv V_W^{bin} - V_W^{sm}$ and the total excess as $\delta_T \equiv \bar{\mu}(1-\bar{\mu}) - \sigma_{sm}^2$. Both are strictly positive under the smooth model. From equation (53), $\sigma_{sm}^2 = V_W^{sm} + V_B^{sm}$. From equation (59), $\bar{\mu}(1-\bar{\mu}) = V_W^{bin} + V_B^{bin}$. Since $V_B^{sm} = V_B^{bin}$ when subgroup means coincide:

$$\begin{aligned} \delta_T &= \left(V_W^{bin} + V_B^{bin}\right) - \left(V_W^{sm} + V_B^{sm}\right) \\ &= V_W^{bin} - V_W^{sm} \\ &= \delta_W. \end{aligned} \tag{63}$$

The overstatement of within-group variance is exactly equal to the overstatement of total variance. As a consequence, the diagnostic

$$\Delta \equiv \hat{V}_{total} - \hat{V}_{within} - \hat{V}_{between} \tag{64}$$

where $\hat{V}_{total} = \hat{\tau}(1-\hat{\tau})$, $\hat{V}_{within} = \sum_g \hat{\omega}_g \hat{\tau}_g(1-\hat{\tau}_g)$, and $\hat{V}_{between} = \sum_g \hat{\omega}_g(\hat{\tau}_g - \hat{\tau})^2$, converges to zero under both models. The Bernoulli-based within-group formula overstates the true within-group variance under the smooth model, but the Bernoulli-based total variance formula overstates the true total variance by exactly the same amount. The decomposition $\hat{V}_{total} = \hat{V}_{within} + \hat{V}_{between}$ is an algebraic identity that holds for any set of numbers $\hat{\tau}_g$ and weights $\hat{\omega}_g$, regardless of the distribution

that generated them.

This result is important because it rules out a class of tests that might initially appear promising. Any diagnostic that computes within-group, between-group, and total variance from subgroup means alone (ie. without access to the within-group distribution of individual effects) will find that the three components add up by construction. The diagnostic cannot distinguish a world where the subgroup means arise from averaging zeros and ones (binary) from a world where they arise from averaging continuously distributed values (smooth). The discrimination must come from elsewhere.

D.5 Where Discrimination Is Possible

Two sources of information can in principle distinguish the models, though both require data beyond published subgroup means.

The behavior of subgroup estimates under progressively finer conditioning differs across models. Under the binary model, subgroup means converge to $\{0, 1\}$ as subgroups become homogeneous in constraint status. Under the smooth model, subgroup means converge to interior conditional expectations. This prediction is testable with data that permits arbitrary partitioning of the sample. Section 6.2 evaluates this prediction using the coarsest available partitions from published studies.

If standard errors and sample sizes are reported for the same subgroup specification, one can estimate the within-group variance directly rather than imputing it from the Bernoulli formula. Under the binary model, the sampling variance of $\hat{\tau}_g$ in a subgroup of N_g firms is

$$\text{Var}(\hat{\tau}_g)^{bin} = \frac{\theta_{C,g}(1 - \theta_{C,g})}{N_g}. \tag{65}$$

Under the smooth model, the sampling variance is

$$\text{Var}(\hat{\tau}_g)^{sm} = \frac{\sigma_g^2}{N_g} < \frac{\mu_g(1 - \mu_g)}{N_g} \tag{66}$$

where the inequality follows from equation (61). The variance ratio

$$R_g \equiv \frac{\widehat{\text{se}}_g^{\,2} \cdot N_g}{\hat{\tau}_g(1 - \hat{\tau}_g)} \tag{67}$$

should converge to one under the binary model and to a value strictly below one under the smooth

model. In practice, reported standard errors in instrumental variables and difference-in-differences settings incorporate first-stage estimation uncertainty, clustering adjustments, and other sources of variation, all of which may inflate R_g above one even under the binary model. The test is therefore conservative: $R_g \geq 1$ is uninformative, but R_g systematically below one across subgroups would provide evidence against the binary specification. The few cases where the required information is available from published papers yield variance ratios consistent with the Bernoulli prediction. But that support is not decisive due the large amount of variation. This approach might be more helpful in subsequent work that has the needed data.