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

Murray Z. Frank[*]

February 12, 2026

Abstract

Treatment effect methods establish causality in corporate finance, but the economic meaning of these estimates is often ambiguous. We provide a framework to bridge this gap using a canonical investment model where the structural marginal response to financing is binary. It is one for constrained firms and zero for unconstrained firms. We show that treatment effects estimates identify the share of constrained firms among compliers rather than a structural parameter. This framework reconciles the seemingly contradictory results of Rauh (2006) and Lemmon and Roberts (2010) by showing that their differing estimates reflect sample composition rather than different structural relationships. A survey of the broader quasi-experimental literature confirms this pattern is pervasive. Finally, we develop formal diagnostic tests based on subgroup convergence and sampling variance to distinguish our binary model from alternative smooth models. Our results provide a principled basis for interpreting treatment effects and evaluating their external validity across different economic settings.

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

[*] University of Minnesota, E-mail: murra280@umn.edu. Thanks to Max Maksimovic for particularly 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 any conceptual or mathematical mistakes, and to help me to improve the writing for correctness and for clarity. I alone am responsible for any errors or misinterpretations. If you find mistakes, please let me know. 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 including instrumental variables, difference-in-differences, and regression discontinuity designs. They do this to establish causal relationships between financing and real outcomes. This methodological shift has produced estimates that are internally valid and statistically well defined. But what do these estimates actually mean in financial economic terms?

Consider a concrete example. Rauh (2006) finds that corporate investment declines by between $0.60 and $0.70 for each dollar of mandatory pension contributions. Is this number a marginal propensity to invest? A structural elasticity that would apply to other shocks? A parameter that informs policy counterfactuals? The answer is unclear. Treatment effect methods identify local average treatment effects (LATE) and related parameters with precise statistical interpretations (Imbens and Angrist, 1994; Imbens, 2010). As stressed by Haile (2025), the economic content of these objects is often ambiguous. Researchers frequently describe their estimates using language like "elasticities," "sensitivities," or "marginal effects". These terms seem to imply structural content that the statistical parameter may or may not possess. We adapt his observation to a corporate finance context.

The gap between statistical identification and economic interpretation matters for several reasons. First, is external validity. A treatment effect estimated for one population may not generalize to another if the underlying structural responses differ across groups (Deaton, 2010; Heckman and Vytlacil, 2007). An estimate from junk rated firms may tell us little about investment grade firms if the two populations have different constraint status. Second, is for policy counterfactuals. Predicting the effect of a proposed policy requires knowing whether the estimate reflects a stable structural relationship or an average that is context specific. If a treatment effect reflects a particular mix of constrained and unconstrained firms, it may not apply to policies that target different populations (Mogstad and Torgovitsky, 2024; Blandhol et al., 2025). Third, is for cross study comparisons. Suppose that two papers both estimate "the effect of financing on investment". They may identify different objects if their identifying variation affects different margins or different subpopulations. Without a framework for translation, comparing estimates across studies is difficult.

This paper provides a framework for understanding what treatment effects measure in terms of economic primitives. We develop a simple structural model of investment with financial constraints in which firms choose investment to maximize profits and some firms face binding financing constraints while others do not. The model delivers a stark characterization. The structural marginal response of investment to financing equals one for constrained firms. An additional dollar of financing translates directly into an additional dollar of investment. It is zero for unconstrained firms, for whom financing capacity is slack.

The central result is that the usual treatment effects are weighted averages of these structural responses across firms with heterogeneous constraint status. The Local Average Treatment Effect (LATE) identifies the share of constrained firms among compliers.[1] It is not the structural marginal response for any particular group. A treatment estimate of 0.65 does not mean that constrained firms invest $0.65 per dollar of financing. It means that approximately 65% of compliers are financially constrained. But each of the constrained firms has a marginal response of one. The remaining 35% of firms are unconstrained with a marginal response of zero. A Monte Carlo simulation confirms this analytical result. The IV estimator recovers θ_C exactly, with negligible bias and tight confidence intervals, across the full range of constraint shares.

This framework allows us to reconcile the seemingly different conclusions of Rauh (2006) and Lemmon and Roberts (2010). Rauh's sample of defined benefit pension sponsors have a range of firms with different credit quality. His complier population contains a mixture of constrained and unconstrained firms, producing an estimate of 0.65 that reflects $\theta_C \approx 0.65$. Consistent with this interpretation, Rauh finds larger effects for firms without investment grade ratings. Lemmon and Roberts study junk rated firms. This is a population that is predominantly constrained by construction. When nearly all compliers are at corner solutions, the treatment effect estimate approximates the structural marginal response. Both papers are thus consistent with a structural marginal propensity to invest of unity for constrained firms. The difference in estimates reflects sample composition, not different structural parameters.

This reconciliation points to the broader contribution of our paper. The goal is to help bridge the gap between econometric identification and economic interpretation. We provide a framework for converting between treatment effects and structural parameters, for understanding when they coincide, and for bounding structural objects when point identification is unavailable. The framework is deliberately simple. It provides clean enough results to generate sharp predictions.

[1] All through this paper we use the term 'complier' in the standard LATE sense of Imbens and Angrist (1994). These are firms whose financing capacity would change in response to the instrument.

At the same time is is rich enough to capture major forces at work in corporate investment decisions. We also examine the impact of relaxing the starkness of the basic model, and show how the interpretations adjust accordingly.

The framework applies broadly. To show this we survey a number of influential quasi-experimental studies spanning multiple countries, time periods, and identification strategies. These include Khwaja and Mian (2008) on bank lending shocks in Pakistan, Chodorow-Reich (2014) on employment effects of lender health, Duchin et al. (2010) on the 2008 crisis, and Chaney et al. (2012) on collateral channels. All studies make judgment calls, and it is always possible to second guess some of those judgments. That is not our purpose. The empirical papers that we include are serious studies. We take their empirical results as reported. The purpose is to translate their evidence from multiple papers into a common interpretative framework.

Across all studies, the same essential pattern appears. Larger treatment effects are reported for subsamples more likely to be constrained (small firms, low-rated firms, low-cashfirms). Effects near zero are reported for likely unconstrained subsamples (large firms, investment-grade firms, high-cash firms). This consistency is difficult to explain under alternative interpretations. But it follows directly from our simple model.

The paper contributes to two literatures. First, we apply ideas from the econometric literature on treatment effect interpretation (Heckman and Vytlacil, 2005; Imbens, 2010; Mogstad and Torgovitsky, 2024; Blandhol et al., 2025) to a canonical corporate finance problem (Fazzari et al., 1988; Kaplan and Zingales, 1997; Almeida et al., 2004; Almeida and Campello, 2007; Farre-Mensa and Ljungqvist, 2016). The model structure features concave production, convex adjustment costs, a financing constraint that may or may not bind. This is standard in the literature.[2] Our contribution is to make the marginal implications of this structure explicit and to connect them to treatment effect estimands. We show that the LATE identifies the constraint share among compliers. This result holds for the standard estimands including LATE, ATE, CATE, and the marginal treatment effect of Heckman and Vytlacil (2005). Each identifies the probability of being constrained for a specific subpopulation.

Second, the modern financing constraints literature starts with Fazzari et al. (1988). It stimulated many papers, leading to a debate initiated by Kaplan and Zingales (1997). They argued that

[2] There is distinct approach to the problem built on dynamic models of the firm (Hennessy and Whited, 2007; Strebulaev, 2007; Whited and Wu, 2006). Papers in that literature study much richer models of the firm in an effort to get a model that matches moments in the data. Such models are often helpful for evaluating policy counterfactuals. The models are sufficiently rich that typically numerical solutions are required. That is fine given their purpose. Here the goal is different. We are trying to obtain a common understanding of the treatment effects studies.

investment cash flow sensitivity does not monotonically reflect constraint severity. Under our binary model, this non-monotonicity does not arise. All constrained firms have identical marginal responses. The more fundamental point of agreement is that the usual reduced form coefficient does not directly measure constraint severity for individual firms. Our framework points to a different object; that is the constraint share among compliers. This is well defined regardless of how constraint severity varies across firms. It does not require the ex ante classification of firms whose reliability has been sharply questioned by both Kaplan and Zingales (1997) and by Farre-Mensa and Ljungqvist (2016).

Our framework shows that treatment effects on investment do not directly identify whether individual firms are constrained. They identify weighted averages that depend on the composition of the complier population. The interpretation we provide is that treatment effects measure constraint shares. This is a revealed-preference characterization. It does not rely on ex ante classification of firms using accounting ratios or text-based measures. Accordingly it avoids the concerns raised by Farre-Mensa and Ljungqvist (2016) about whether standard constraint proxies actually identify constrained firms.

For empirical practice, our analysis has several implications. Researchers should be explicit about the composition of their complier populations. Treatment effects that fall strictly between zero and one should be interpreted as evidence of population heterogeneity, not as estimates of a structural marginal response. Heterogeneity analysis by constraint proxies is of course already standard in much of this literature. In our framework this is more than just a robustness exercise. It is actually informative about structural parameters. Subsamples with treatment effects approaching unity provide direct evidence on the structural marginal response. Subsamples with effects near zero confirm that unconstrained firms do not respond to financing shocks.

The paper proceeds as follows. Section 2 develops the structural model. Section 3 derives what treatment effects identify in terms of the model's primitives and presents a Monte Carlo illustration. Section 4 applies the framework to Rauh (2006) and Lemmon and Roberts (2010). Section 5 extends the analysis to the broader literature. Section 6 discusses why the constraint interpretation is supported by the evidence. Section 7 concludes.

2 A Structural Model of Investment and Financing

This section develops a simple model of corporate investment with financial constraints. The model delivers predictions about how investment responds to changes in financing capacity, providing a structural benchmark against which to interpret 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

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

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 multi-period dynamics, uncertainty, and the endogenous determination of debt capacity. These simplifications isolate the mechanism that matters for interpreting treatment effects. That is the distinction between constrained and unconstrained firms. The model structure has concave production, convex adjustment costs, a financing constraint that may or may not bind. These elements are standard in the literature from Fazzari et al. (1988) through Almeida and Campello (2007). Our contribution is not the model itself. It is to show the marginal implications and their connection to treatment effect estimands. In that way we aim to clarify the economic meaning of the estimates.

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*

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

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 financing capacity. An additional dollar of internal funds or debt capacity does not change their investment because they are already at their optimal scale. Constrained firms are at a corner solution where investment is limited by available financing. An additional dollar relaxes the constraint and translates directly into an additional dollar of 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.

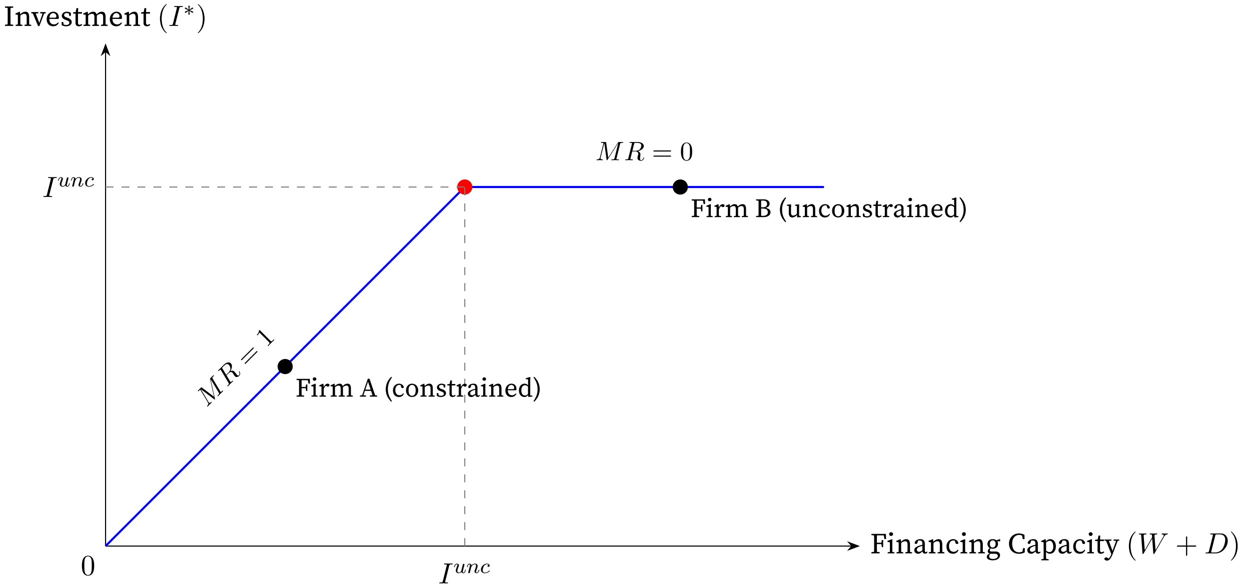

Figure 1: The corner solution in optimal investment. The figure plots optimal investment I^* as a function of financing capacity $(W + D)$ from equation (5). When financing capacity falls below the unconstrained optimum I^{unc}, the firm invests all available funds and the marginal response equals one. 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 $MR = 0$.

2.4 Population Heterogeneity

Firms differ in their constraint status, and this 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 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 marginal response is binary: zero or one. This follows from the corner solution and is a simplifying assumption. In models with smooth financing costs, the marginal response would vary continuously with constraint severity, with more severely constrained firms having higher responses. The qualitative insight survives: reduced-form estimates average over heterogeneous structural responses, and this averaging attenuates estimates relative to the response for the most constrained firms. Section 6 provides evidence that the binary characterization fits the data well.

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." More generally, partial substitution can be accommodated by reinterpreting the structural parameter as the marginal response net of substitution. The main insight remains. The treatment effects identify weighted averages of structural responses. Understanding the weights requires understanding the composition of the complier population.

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 When Does LATE Equal the Structural Parameter?

LATE equals the structural marginal response for constrained firms if and only if all compliers are constrained.

$$\tau_{LATE} = 1 \iff \theta_C = 1. \tag{12}$$

This condition is satisfied in three main situations. The instrument might specifically targets constrained firms. The entire population could be constrained. Selection into treatment could be perfectly correlated with constraint status.

In practice, LATE typically falls below one because the complier population includes some unconstrained firms. An estimate of $\hat{\tau} = 0.65$ implies that approximately 65% of compliers are constrained. The remaining 35% are unconstrained. The unconstrained firms contribute zero to the average, and pull the estimate below the structural parameter.

3.4 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

$$\tau_{ATE} = E\left[\frac{\partial I^*}{\partial F}\right] = \theta, \tag{13}$$

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

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
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 $S < 0$.

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

$$\tau(X) = E\left[\frac{\partial I^*}{\partial F} \,\middle|\, X\right] = P(S < 0 \mid X). \tag{14}$$

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).

$$MTE(u) = E\left[\frac{\partial I^*}{\partial F} \,\middle|\, U = u\right] = P(S < 0 \mid U = u) \tag{15}$$

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.

3.5 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.

$$\widehat{MR}_{structural} = \frac{\hat{\tau}}{\hat{\theta}_C}. \tag{16}$$

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,

$$\hat{\tau} \leq MR_{structural} \leq 1. \tag{17}$$

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.

3.6 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: 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. First, 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. Second, 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, 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–\$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 provides a unified interpretation. Both papers estimate the same structural parameter. It is a marginal response of one for constrained firms. Their different estimates reflect different complier populations.

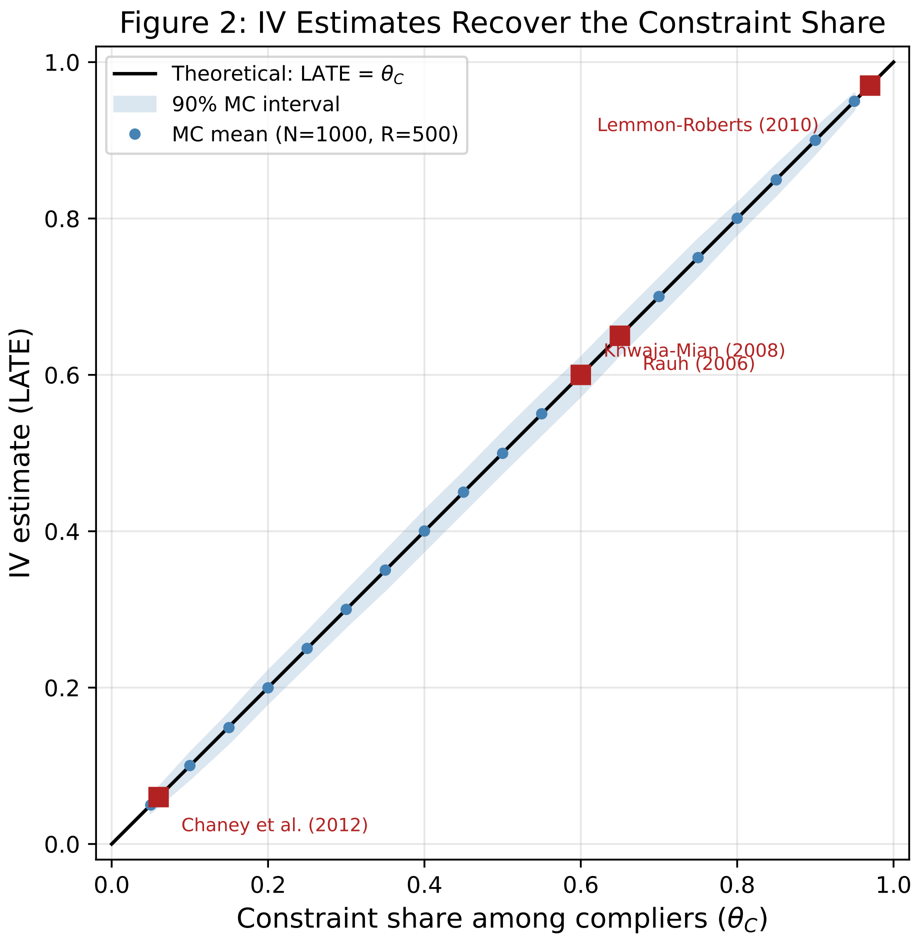

Figure 2: IV estimates recover the constraint share. The solid line is the theoretical prediction $\tau_{LATE} = \theta_C$ from Proposition 2. Blue circles are Monte Carlo means across 500 replications with $N = 1{,}000$ firms. The shaded band is the 90% Monte Carlo interval. Red squares locate empirical 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.

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

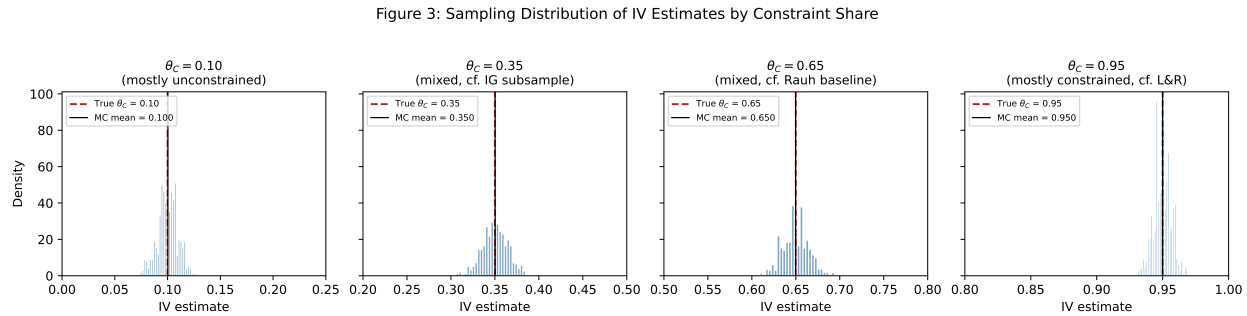

Figure 3: Sampling distribution of the IV estimator at four constraint shares. Each panel displays 500 Monte Carlo replications with $N = 1{,}000$ firms. The dashed line marks the true θ_C; the solid line marks the Monte Carlo mean. From left to right: $\theta_C = 0.10$, 0.35, 0.65, and 0.95. The estimator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

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–\$0.70.[3] 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]

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

$$\hat{\tau}_{Rauh} \approx 0.65 \quad \Longrightarrow \quad \hat{\theta}_C \approx 0.65. \tag{18}$$

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

[3] Rauh (2006), Table IV, p. 54, reports IV coefficients ranging from -0.60 to -0.72 across specifications. OLS coefficients on cash flow are approximately 0.10, illustrating attenuation bias from measurement error (Erickson and Whited, 2000) and simultaneity.

[4] See Rauh (2006), Table VI, p. 63.

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: 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.[5]

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 5–6% of assets relative to the control group.[6] 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."[7]

4.2.2 Interpretation Through the Framework

Lemmon and Roberts estimate the marginal response to debt capacity, $\hat{\tau}_{LR} \approx 1$. Why is this estimate approximately one, whereas Rauh's is 0.65? The answer lies in the composition of the treated sample.[8]

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

[5] See Lemmon and Roberts (2010), pp. 556–564, for a detailed discussion of each event.

[6] Lemmon and Roberts (2010), Table 4, Panel A, p. 574. 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.

[7] Lemmon and Roberts (2010), p. 555.

[8] Strictly speaking, their difference-in-differences design estimates the average treatment effect on the treated (ATT) rather than LATE. Under our model, if the treated population is predominantly constrained, the ATT approximates the structural marginal response. We use compliers broadly to refer to firms whose behavior is affected by the financing shock.

constrained among these firms is high.

$$\theta_{C,junk} \approx 1 \quad \Longrightarrow \quad \hat{\tau}_{LR} \approx \theta_{C,junk} \cdot 1 \approx 1. \tag{19}$$

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

Table 2 presents the two studies side by side. The difference in point estimates reflects differences in sample composition rather than different structural parameters. Rauh's sample covers firms with a range of credit qualities. That means he has a complier population that mixes constrained and unconstrained firms. His estimate of 0.65 implies that roughly 65% of compliers are at corner solutions. Lemmon and Roberts study junk-rated firms exclusively. That is a population that is predominantly constrained by construction. When nearly all compliers have a marginal response of one, the treatment estimate approximates the structural parameter.

The heterogeneity evidence in both papers provides a consistency check. Rauh finds larger effects for firms without investment-grade ratings, exactly as the framework predicts for a population with a higher constraint share. Lemmon and Roberts document that treated firms exhibited limited substitution to alternative financing, precisely what we expect from constrained firms who have already exhausted cheaper funding options. Higher treatment effects for populations more likely constrained, combined with limited substitution behavior; supports the idea that both papers identify the same structural relationship. It is a marginal propensity to invest of unity for financially constrained firms.

The reconciliation has implications for interpreting existing estimates and designing future research.

First, it validates the structural framework. The predicted pattern of heterogeneity (larger

Table 2: Reconciling Rauh (2006) and Lemmon & Roberts (2010)

	Rauh (2006)	Lemmon & Roberts (2010)
Financing shock	Mandatory pension contributions	Junk bond market collapse
Identification	IV: Pension funding rules	DiD: Credit supply shocks
Sample	DB pension sponsors (all credit ratings)	Below-investment-grade firms only
Baseline estimate	0.60–0.70	≈ 1.0
Heterogeneity	Larger effects for lower-rated firms	Limited substitution to alternative financing
Implied θ_C	0.60–0.70 (mixed population)	≈ 1.0 (predominantly constrained)
Structural MR	1.0	1.0

Notes: Both papers estimate the causal effect of financing on investment using plausibly exogenous variation. Rauh's IV estimate of 0.60–0.70 reflects a complier population that includes both constrained and unconstrained firms; the estimate identifies θ_C among compliers. Lemmon and Roberts' near-unity estimate reflects a predominantly constrained sample. Under the framework, both are consistent with $MR_{structural} = 1$ for constrained firms. DiD = difference-in-differences; DB = defined benefit; FIRREA = Financial Institutions Reform, Recovery, and Enforcement Act of 1989.

effects for more constrained subsamples) matches the observed pattern in both papers, using different instruments, different samples, and different outcome measures. The fact that estimates from two distinct settings can be reconciled within a single framework provides evidence for the framework itself.

Second, the analysis highlights the importance of understanding complier populations. Rauh's instrument (pension funding shocks) affects firms spanning the constraint distribution. Lemmon and Roberts' instrument (junk bond market collapse) affects only firms reliant on below-investment-grade debt, i.e. a predominantly constrained population. Without attention to this selection, one might mistakenly conclude that the structural relationship between financing and investment differs across settings.

Third, the framework clarifies external validity. Generalizing either estimate to a new population requires understanding the constraint share in that population. A policy targeting investment grade firms would likely have smaller effects than either paper estimates. That is because a larger fraction of affected firms would be unconstrained. A policy targeting small, cash poor firms might have effects closer to one.

5 The Framework Applied to the Broader Literature

The reconciliation of Rauh (2006) and Lemmon and Roberts (2010) in Section 4 shows the framework for two specific high profile papers. 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.

It should be stressed that each paper use somewhat different language to describe their results. The purpose here is not to take issue with the description of the results in any particular paper per se. Again, these are serious studies. The purpose is to provide a framework that provides a unified economic interpretation across papers in a particularly simple manner.

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 key 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.

$$\hat{\tau}_d = \hat{\theta}_{C,d} \cdot 1 + (1 - \hat{\theta}_{C,d}) \cdot 0 = \hat{\theta}_{C,d}. \tag{20}$$

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

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

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

Notes: Treatment effects are approximate values from Figure VII of Khwaja and Mian (2008). Implied constraint shares follow from equation (20) 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.

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.

$$\hat{\tau}_{full} = \sum_{d=1}^{10} \omega_d \cdot \hat{\tau}_d = 0.6 \tag{21}$$

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.

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

Table 4: Heterogeneity in Treatment Effects Across the Literature

Paper	Baseline estimate	Constrained subsample	Unconstrained subs
Rauh (2006)	0.60–0.70	Low-rated: higher	IG-rated: lower
Lemmon-Roberts (2010)	≈ 1.0	Junk sample	—
Khwaja-Mian (2008)	0.6	Small: 0.6–0.8	Large: ≈ 0
Chodorow-Reich (2014)	Significant	Small/med: sig.	Large: zero
Duchin et al. (2010)	Significant	Low cash: $2\times$	High cash
Campello et al. (2010)	9% planned cut	Constrained CFOs: 86% bypassed projects	Unconstrained: 0.4
Almeida et al. (2012)	2.5 ppt decline	Unrated: larger	Rated
Gan (2007)	0.08	High lev.: larger	Low lev.
Chaney et al. (2012)	0.06	Small: larger	Large
Peek-Rosengren (2000)	Significant	High exp.: larger	Low exp.

Notes: "Estimate" reports the baseline treatment effect. "Constrained" and "Unconstrained" summarize heterogeneity by constraint proxies. All studies find larger effects for likely constrained subsamples, consistent with the prediction that treatment effects identify constraint shares among compliers. IG = investment grade; lev. = leverage; exp. = exposure.

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.

5.3 Discussion

Despite differences in countries, time periods, identification strategies, and outcome variables, Table 4 reveals a consistent pattern. Effects are larger for subsamples likely to be constrained and smaller or zero for likely unconstrained subsamples. When complier populations are predominantly constrained, estimates approach unity; when they include many unconstrained firms, estimates are attenuated toward zero.

This consistency is informative about the structural model. If treatment effects measured marginal responses that varied continuously for reasons unrelated to constraints, there would be no reason for the same firm characteristics—size, credit rating, cash holdings, bond market access—to predict larger effects across such diverse settings. The fact that they do supports the interpretation that treatment effects identify constraint shares rather than structural elasticities.

The evidence also clarifies external validity. An estimate from one population does not automatically apply to another if constraint shares differ. A policy targeting investment-grade firms would likely produce smaller effects than estimates from junk-rated samples suggest, because a larger fraction of affected firms would be unconstrained. Conversely, a policy targeting small, cash-poor firms might produce effects closer to one. The framework provides a principled basis for such adjustments by linking treatment effect magnitudes to the composition of complier populations.

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 $\partial I/\partial F$. 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 establishes three facts that any interpretation ought to 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 a marginal response interpretation, this would require that these firms have near zero marginal products of investment. That is economically implausible. These firms have valuable investment opportunities. They simply do not need the marginal source of financing to pursue them. Under the constraint interpretation, the finding is natural. The unconstrained firms are at interior solutions where financing shocks do not affect investment.

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 a marginal response interpretation, this would require that constrained firms happen to have a marginal product of investment exactly equal to one. But investment opportunities vary across firms and industries. There is no economic reason for this particular value to arise generically. Under the constraint interpretation, the finding follows directly. Constrained firms invest every dollar of available financing. 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 exhibit 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 a marginal response interpretation, this consistency would be surprising. Why should the marginal product of investment covary with firm size, cash holdings, and credit ratings in exactly the same way across all these settings? But under the constraint interpretation, the consistency is expected. The same characteristics that predict financial constraints also predict larger treatment effects. That is because the treatment effect measures the constraint share.

6.2 Formal Diagnostic Tests (Convergence and Variance)

The binary and smooth models make different predictions about the distribution of individual treatment effects. Under our binary model, $\tau_i \in \{0, 1\}$. Each firm either invests every marginal dollar or does not respond at all. Subgroup estimates are weighted averages of zeros and ones.

So they can take intermediate values, but the underlying individual effects are concentrated at two mass points. Under the smooth model, $\tau_i = \partial I_i / \partial F$ varies continuously. The individual effects need not cluster at any particular values. Appendix D provides complete derivations for the variance diagnostic developed in this section.

A natural approach to distinguishing these models would decompose the variance of treatment effects into within-group and between-group components. Then check whether the components add up to 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, not just the Bernoulli. 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.

Two alternative diagnostics do have power.

6.2.1 Convergence to Binary Bounds

The sharpest distinction between the models concerns the behavior of subgroup estimates as conditioning on constraint relevant observables becomes finer.

Under the binary model, individual treatment effects take values in $\{0,1\}$. As the partition of firms into subgroups becomes sufficiently fine that each subgroup is homogeneous in constraint status, subgroup estimates must converge to the boundary values.

$$\hat{\tau}_g \;\longrightarrow\; c_g \in \{0,1\} \quad \text{as conditioning refines.} \tag{22}$$

Under the smooth model, individual effects take interior values. Finer conditioning reduces within-group heterogeneity but drives subgroup means toward the conditional means of a continuous distribution, which generically lie in the interior of $[0,1]$. There is no reason for subgroup estimates to approach zero or one.

To operationalize this distinction, define the convergence ratio for subgroup g,

$$\rho_g \equiv \min\bigl(\hat{\tau}_g,\ 1-\hat{\tau}_g\bigr). \tag{23}$$

This measures the distance of $\hat{\tau}_g$ from the nearest binary bound. Values near zero indicate proximity to $\{0, 1\}$. Values near 0.5 indicate an intermediate estimate far from either bound. The binary model predicts $\rho_g \to 0$ for sufficiently fine subgroups. The smooth model predicts that ρ_g remains bounded away from zero.

The empirical evidence favors convergence to the binary bounds. Across the studies surveyed in Table 4, the finest available subgroup estimates consistently approach the boundary values rather than stabilizing at intermediate levels.

Near-zero effects for unconstrained subsamples. The top three size deciles in Khwaja and Mian (2008) have $\hat{\tau}_g \approx 0$, giving $\rho_g \approx 0$. Investment-grade firms in Rauh (2006) have smaller effects than the full sample, moving toward zero. Large transparent firms in Chodorow-Reich (2014) show effects indistinguishable from zero. High-cash firms in Duchin et al. (2010) have effects roughly half those of low-cash firms, consistent with a lower constraint share rather than an intermediate structural response.

Near-unity effects for constrained subsamples. Lemmon and Roberts (2010) sample of junk-rated firms yields $\hat{\tau} \approx 0.97$, giving $\rho \approx 0.03$. The smallest firm deciles in Khwaja and Mian (2008) have effects in the range 0.6–0.8. Unrated firms in Almeida et al. (2012) show the largest effects within their sample.

Under the smooth model, there is no economic reason for the marginal product of investment to equal exactly zero for large firms and exactly one for small firms. Investment opportunities vary across firms and industries. A smooth structural response would generate subgroup estimates scattered across $[0, 1]$ with no particular tendency toward the boundaries. The systematic convergence to $\{0, 1\}$ across countries, decades, and identification strategies is the pattern most difficult to reconcile with the smooth model and most naturally explained by the binary model.

6.2.2 The Standard Error Test

A complementary diagnostic exploits the relationship between subgroup treatment effects and their sampling variance. Under the binary model, τ_i is Bernoulli within each subgroup, so the sampling variance of $\hat{\tau}_g$ in a sample of N_g compliers is

$$\text{Var}(\hat{\tau}_g) = \frac{\theta_{C,g}(1 - \theta_{C,g})}{N_g}. \tag{24}$$

Under the smooth model, the within-group variance of individual effects σ_g^2 is strictly less than $\mu_g(1 - \mu_g)$ for any non-degenerate continuous distribution on $[0, 1]$ (Appendix D, equation 77).

The sampling variance is therefore

$$\text{Var}(\hat{\tau}_g) = \frac{\sigma_g^2}{N_g} < \frac{\mu_g(1-\mu_g)}{N_g}. \tag{25}$$

Define the variance ratio:

$$R_g \equiv \frac{\widehat{\text{se}}_g^{\,2} \cdot N_g}{\hat{\tau}_g(1-\hat{\tau}_g)}. \tag{26}$$

Under the binary model, $R_g \to 1$ in probability. Under the smooth model, $R_g < 1$. Values of R_g near one across subgroups support the binary interpretation; values systematically below one indicate that the within-group dispersion of individual effects is less than the Bernoulli benchmark, consistent with a continuous distribution of marginal responses.

The test is conservative. In IV and difference-in-differences settings, reported standard errors reflect estimation uncertainty from both the first stage and the reduced form, and robust or clustered standard errors incorporate additional sources of variation. These factors may inflate R_g above one even under the binary model, so $R_g \geq 1$ is uninformative. But R_g systematically and substantially below one across subgroups would provide evidence against the binary specification.

Applying this test requires reported standard errors and sample sizes for each subgroup, which are available in some but not all of the studies we survey. Where the information is available, reported standard errors are broadly consistent with Bernoulli predictions. A more systematic application would require access to the underlying microdata, which we leave for future work.

6.2.3 Summary

Table 5 summarizes the empirical predictions that distinguish the two models.

The convergence-to-bounds test is the more powerful of the two diagnostics, because it can be applied using published subgroup estimates without additional data requirements. The pattern in the literature is clear: across all studies surveyed in Table 4, the finest available subsample estimates cluster near the binary bounds rather than spreading continuously across $[0, 1]$. This pattern is difficult to explain under the smooth model—there is no economic reason for marginal products to equal exactly zero or one—but follows directly from the binary characterization of financial constraints.

The question of whether individual treatment effects are binary or continuous is not new here. It has been studied in biostatistics where it is called 'responder analysis' (Senn, 2004; Gadbury

Table 5: Diagnostic Predictions: Binary versus Smooth Model

Observable	Binary model	Smooth model
Subgroup estimates under fine conditioning	Converge to $\{0, 1\}$	Stabilize at interior values
Convergence ratio ρ_g for fine subgroups	$\to 0$	Bounded away from zero
Variance ratio R_g	≈ 1	< 1
Cross-study pattern of heterogeneity	Same characteristics predict larger $\hat{\tau}$	No systematic pattern required

Notes: The convergence ratio is $\rho_g \equiv \min(\hat{\tau}_g, 1 - \hat{\tau}_g)$. The variance ratio is $R_g \equiv \widehat{\mathrm{se}}_g^{\,2} \cdot N_g/[\hat{\tau}_g(1 - \hat{\tau}_g)]$. Under the binary model, individual effects are $\{0, 1\}$, so subgroup means converge to the bounds with fine conditioning and within-group dispersion matches the Bernoulli variance. Under the smooth model, individual effects vary continuously, producing interior subgroup means and within-group variance below the Bernoulli benchmark. Appendix D provides the complete algebraic derivations.

et al., 2001), and similar variance-based diagnostics have been applied for clinical trial data. In econometrics, Heckman (2010) develop more general tests for essential heterogeneity in treatment effects. Our contribution is to apply this logic to the corporate finance setting. In our setting the binary structure is dues to the economics of corner solutions rather than being assumed. A reason to do this for the corporate finance problem is that published subsample estimates across several studies provide a helpful basis for the convergence diagnostic.

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:

$$\Delta I = -\Delta, \qquad MR = 1. \tag{27}$$

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

$$\Delta I = 0, \qquad MR = 0. \tag{28}$$

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:

$$\Delta I = -(\Delta - S), \qquad MR = \frac{\Delta - S}{\Delta} \in (0, 1). \tag{29}$$

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

$$\begin{aligned} \tau_{LATE} &= \theta_{AC} \cdot 1 + \theta_{SW} \cdot E\left[\frac{\Delta - S}{\Delta} \;\middle|\; 0 \leq S < \Delta,\, C\right] + \theta_{AU} \cdot 0 \\ &= \theta_{AC} + \theta_{SW} \cdot \bar{\lambda}_{SW} \end{aligned} \tag{30}$$

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 $\Delta \to 0$,*

$$\theta_{SW} = P(0 \leq S < \Delta \mid C) = f_S(0) \cdot \Delta + o(\Delta) \to 0 \tag{31}$$

and

$$\tau_{LATE} \to \theta_{AC} = P(S < 0 \mid C) = \theta_C. \tag{32}$$

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

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

$$\theta_{AC} < \tau_{LATE} < \theta_{AC} + \theta_{SW}. \tag{33}$$

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}—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—larger effects for lower-rated firms—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—firms with moderate slack that could plausibly be pushed into constrained territory. In such settings, the binary model overstates the pre-shock constraint share, and 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:

$$\tau(\Delta_2) \geq \tau(\Delta_1) \tag{34}$$

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 estimated 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—such as variation in the size of collateral value changes—could estimate the LATE as a function of Δ and test whether it is approximately constant (supporting the binary model) or increasing (indicating switchers).

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 (30), $\tau(\Delta)$ 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

$$\hat{\tau} = \theta_C \cdot MR_C + (1 - \theta_C) \cdot MR_U \tag{35}$$

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 one for constrained firms then provides evidence on the specific values of MR_U and MR_C, suggesting $MR_U \approx 0$ and $MR_C \approx 1$.

Even if these values were, say, $MR_U = 0.1$ and $MR_C = 0.9$, the framework's core insight would remain. Treatment effects identify weighted averages of heterogeneous structural responses. The weights depend on complier composition. Cross-study differences in estimates reflect differences in these weights, not different structural relationships. The binary parameterization $\{0, 1\}$ is a useful benchmark that fits the data well, but the interpretive framework does not depend on it.

6.5 Limitations

Several limitations should be noted. First, 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.

Second, 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.

Third, 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.

Finally, 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 allows. 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 credibility revolution has transformed empirical corporate finance, producing treatment effect estimates with strong internal validity. But a gap has emerged between statistical identification and economic interpretation. When researchers report that a financing shock affects investment with a coefficient of 0.65, what does this number mean? Various papers use a range of verbal interpretations. In corporate finance we have lacked a principled framework for translating between treatment estimates and structural parameters.

This paper fills that gap by drawing on the labor econometrics literature Heckman and Vytlacil (2005); Imbens (2010); Mogstad and Torgovitsky (2024). We develop a canonical model of investment with financial constraints in which the structural marginal response to financing is binary. It is one for constrained firms, zero for unconstrained firms. The model delivers a sharp characterization. Treatment effects identify 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 sixty-five percent of compliers are at corner solutions, each investing every available dollar; and the remaining thirty-five percent are unconstrained and do not respond.

This reinterpretation is not semantic. It explains longstanding puzzling differences across studies. For more than a decade, the Rauh (2006) estimate of 0.60–0.70 and Lemmon and Roberts' (2010) estimate approaching unity have sat uncomfortably next to each other in the literature. Speculative explanations include different types of financing shocks, different firm responses in different contexts, or fundamental differences in the investment financing relationship across samples. Our framework shows that both papers actually identify the same structural parameter, with a marginal propensity to invest of one for constrained firms. The difference in point estimates reflects complier composition, not different structural relationships. Rauh's knows that there is heterogeneity of creditworthiness among his compliers, and that 0.65 is some type of an average. We show how that affects the interpretation of the estimates. Lemmon and Roberts' junk-rated firms are predominantly constrained by construction. They get a coefficient of 1. The subsample evidence in both papers shows larger effects for lower-rated firms in Rauh, and nearly complete responses with limited substitution in Lemmon and Roberts. This confirms our interpretation.

The broader empirical evidence across many high quality papers strengthens the case. Across countries, time periods, identification strategies, and outcome variables, we document a systematic pattern. Treatment effects are larger for subsamples more likely to be constrained (small firms, low-rated firms, low-cash firms) and approach zero for likely unconstrained subsamples (large firms, investment-grade firms, high-cash firms). This consistency is difficult to explain under alternative interpretations. If treatment effects measured structural marginal responses that varied continuously for reasons unrelated to constraints, there would be no reason for the same firm characteristics to predict larger effects across such diverse settings. The pattern follows directly from the constraint interpretation. We formalize this with diagnostic tests based on subgroup convergence and sampling variance. The finest available subsample estimates cluster near the binary bounds $\{0, 1\}$ rather than spreading continuously across the unit interval. Variance ratios are consistent with Bernoulli individual effects rather than smooth heterogeneity.

The framework has concrete implications for empirical practice. First, it changes how researchers should design studies. Rather than targeting representative samples and interpreting point estimates as structural parameters, it might be helpful to deliberately oversample constrained populations to identify the structural marginal response and oversample unconstrained populations to verify zero effects. The difference in constraint shares across samples provides quantitative bounds on structural parameters.

Second, it changes how researchers should interpret heterogeneity. Subsample analysis is not just a robustness check. It is a primary tool for recovering structural content from treatment estimates. Subsamples with treatment effects approaching unity provide direct evidence on the structural marginal response. Subsamples with effects near zero confirm that unconstrained firms do not respond to financing shocks.

Third, it changes how researchers should evaluate external validity. Generalizing an estimate from one population to another requires explicit accounting for constraint shares. A policy proposal informed by Rauh's estimate of 0.65 would have different effects if applied to investment-grade firms ($\hat{\theta}_C \approx 0.35$) versus junk-rated firms ($\hat{\theta}_C \approx 0.95$). These are not small differences, and current practice does not provide a principled basis for making such adjustments.

Our approach extends naturally beyond investment. Any setting where a real outcome responds to a financing shock through a binding constraint will have the same sort of decomposition. Examples might include cash-flow sensitivity of cash holdings (Almeida et al., 2004), employment responses to lender health (Chodorow-Reich, 2014), payout policy under financing frictions, and collateral-driven investment (Chaney et al., 2012). All of these involve weighted averages of heterogeneous structural responses across constrained and unconstrained firms. The interpretive principles apply broadly.

Several directions for future research emerge. First, developing methods to estimate constraint shares directly, rather than inferring them from treatment effect magnitudes, would strengthen the empirical foundations. Researchers could potentially combine revealed preference restrictions from financing decisions with quasi-experimental variation to separately identify θ_C and $\mathrm{MR}_{\text{structural}}$. Second, integrating the static framework with dynamic models that endogenize constraint status and allow for precautionary behavior would extend the analysis to settings where the binary approximation is less appropriate. Third, applying the framework systematically to other corporate finance questions including leverage adjustments, cash holdings, risk management, would test its generality and potentially reveal additional empirical regularities.

The central message is straightforward. Treatment effect methods have brought rigor to causal inference in corporate finance. But the economic interpretation has lagged behind statistical identification. As forcefully argued by Haile (2025) researchers frequently describe their estimates using language suggesting structural content the statistical parameter may not possess. Our framework provides a bridge. It shows what treatment effects identify in terms of economic primitives, when they coincide with structural parameters, and how to recover structural objects when point identification is unavailable. The cost is acknowledging that a single estimate rarely identifies a structural parameter directly. The benefit is a principled basis for interpreting estimates, comparing them across studies, and using them to inform policy counterfactuals. Understanding how financing frictions affect real decisions is central to corporate finance. So clarity about what our estimates actually measure is not a luxury. It is a necessity.

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.

Blandhol, C., J. Bonney, M. Mogstad, and A. Torgovitsky (2025). 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.

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.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424–455.

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.

Gadbury, G. L., H. K. Iyer, and D. B. Allison (2001). Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics 11*(4), 313–333.

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. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic literature 48*(2), 356–398.

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.

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.

Senn, S. (2004). Individual response to treatment: is it a valid assumption? *The Bmj 329*(7472), 966–968.

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,

$$MTE(u) = E[Y_{1i} - Y_{0i} \mid U_i = u], \tag{36}$$

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:

$$LATE = \int_0^1 MTE(u) \cdot \omega(u)\, du \tag{37}$$

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

$$ITE_i = \frac{\partial I_i^*}{\partial F} = \begin{cases} 1 & \text{if } S_i < 0 \quad \text{(constrained)} \\ 0 & \text{if } S_i \geq 0 \quad \text{(unconstrained)} \end{cases} \tag{38}$$

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

$$MTE(u) = E\left[\frac{\partial I^*}{\partial F}\,\middle|\, U = u\right] = P(S < 0 \mid U = u) \tag{39}$$

which equals the probability that firms at quantile u of unobserved resistance are financially constrained. This is equation (15) in the main text.

A.3 The Step-Function MTE

The corner solution in optimal investment generates a step-function MTE—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

$$MTE(u) = \begin{cases} 1 & \text{if } u \leq u^* \\ 0 & \text{if } u > u^* \end{cases} \tag{40}$$

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

$$LATE = \int_0^{u^*} 1 \cdot \omega(u)\, du + \int_{u^*}^{1} 0 \cdot \omega(u)\, du = \int_0^{u^*} \omega(u)\, du = \theta_C. \tag{41}$$

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 exhibit 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: 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 6 summarizes the results.

B.1 Methodology

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

$$\hat{\tau}_{LATE} = \hat{\theta}_C \cdot 1 + (1 - \hat{\theta}_C) \cdot 0 = \hat{\theta}_C. \tag{42}$$

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—change in investment per unit change in financing—not as an elasticity or percentage change. For studies reporting elasticities, we note the limitation and focus on the qualitative pattern.

B.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, rising monotonically to 60–80% 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 key 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

Table 6: 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)	Elasticity 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	Higher in high-exposure mkts.	Varies across local markets

Notes: Implied constraint shares follow from equation (42) 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.

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, $\hat{\theta}_C \approx 6\%$—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 with access to diverse financing. 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.

B.3 Interpretation

Several patterns emerge from Table 6. First, 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.

Second, within-study heterogeneity consistently shows higher constraint shares for firms with 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.

Third, the implied constraint shares align with 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 6 provide a basis for making such adjustments explicit.

C 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.

C.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

$$I^* = \begin{cases} I^{unc} & \text{if } S \geq 0 \quad \text{(unconstrained)} \\ (1-\alpha)(W+D) & \text{if } S < 0 \quad \text{(constrained)} \end{cases} \tag{43}$$

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

$$MR = \begin{cases} 0 & \text{if } S \geq 0 \\ 1 - \alpha & \text{if } S < 0. \end{cases} \tag{44}$$

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

C.2 Modified LATE Decomposition

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

$$\tau_{LATE} = \theta_C \cdot (1 - \alpha) + (1 - \theta_C) \cdot 0 = \theta_C(1 - \alpha). \tag{45}$$

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.

C.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:

$$\hat{\tau}_j = \theta_{C,j}(1 - \alpha), \quad j = 1, 2. \tag{46}$$

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:*

$$\frac{\hat{\tau}_1}{\hat{\tau}_2} = \frac{\theta_{C,1}}{\theta_{C,2}}. \tag{47}$$

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

$$\hat{\alpha} = 1 - \hat{\tau}_2, \tag{48}$$

and the constraint share in any other population is recovered as

$$\hat{\theta}_{C,1} = \frac{\hat{\tau}_1}{\hat{\tau}_2}. \tag{49}$$

Proof. Part (a) follows immediately from dividing equation (46) for $j = 1$ by the same equation for $j = 2$. The $(1-\alpha)$ terms cancel. Part (b) substitutes $\theta_{C,2} = 1$ into equation (46) to obtain $\hat{\tau}_2 = 1-\alpha$, which gives (48). Substituting into the expression for population 1 yields $\hat{\tau}_1 = \theta_{C,1} \cdot \hat{\tau}_2$, from which (49) follows. □

C.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:

$$\hat{\tau}_{LR} \approx 0.97 \quad \Longrightarrow \quad \hat{\alpha} \approx 1 - 0.97 = 0.03. \tag{50}$$

Precautionary savings are negligible—approximately three cents per dollar of the financing shock—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 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

$$\hat{\theta}_{C,Rauh} = \frac{\hat{\tau}_{Rauh}}{\hat{\tau}_{LR}} = \frac{0.65}{0.97} \approx 0.67. \tag{51}$$

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 7: 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.

Table 7 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.

C.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 (45) implies

$$\theta_C \geq \hat{\tau} \qquad \text{and} \qquad \alpha \leq 1 - \hat{\tau}. \tag{52}$$

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. If the most constrained subsample—identified by small size, low ratings, or low cash—has treatment effect $\hat{\tau}_{high}$ and we assume $\theta_{C,high} \approx 1$, 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

$$\hat{\theta}_C \geq \frac{0.65}{1} = 0.65 \qquad \text{and} \qquad \hat{\theta}_C \leq \frac{0.65}{1 - 0.16} = 0.77. \tag{53}$$

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

C.6 Discussion

Three features of this extension merit comment.

First, 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.

Second, 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.

Third, 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.

D Algebraic Derivation of the Variance Diagnostic

This appendix provides complete derivations for the variance diagnostic developed in Section 6.2. We proceed in four steps: we establish the law of total variance for the binary model, derive the analogous decomposition under the smooth model, show why the two models generate different predictions for observable variance components, and construct the test statistic.

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:

$$\tau_g \equiv E[\tau_i \mid i \in g] \qquad \text{(subgroup mean)} \tag{54}$$

$$\bar{\tau} \equiv \sum_{g=1}^{G} \omega_g \tau_g = E[\tau_i] \qquad \text{(population mean)} \tag{55}$$

$$\sigma_g^2 \equiv \text{Var}(\tau_i \mid i \in g) \qquad \text{(within-group variance)} \tag{56}$$

$$\sigma^2 \equiv \text{Var}(\tau_i) \qquad \text{(total variance).} \tag{57}$$

D.2 The Law of Total Variance

The law of total variance (also known as the variance decomposition or Eve's law) states that for any random variable X and partition variable G,

$$\text{Var}(X) = E[\text{Var}(X \mid G)] + \text{Var}(E[X \mid G]). \tag{58}$$

Applying this to τ_i with the subgroup partition yields

$$\sigma^2 = \underbrace{\sum_{g=1}^{G} \omega_g \, \sigma_g^2}_{V_W} + \underbrace{\sum_{g=1}^{G} \omega_g (\tau_g - \bar{\tau})^2}_{V_B} \tag{59}$$

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.3 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. We derive each variance component.

Subgroup mean.

$$\tau_g = E[\tau_i \mid i \in g] = 1 \cdot P(\tau_i = 1 \mid g) + 0 \cdot P(\tau_i = 0 \mid g) = \theta_{C,g}. \tag{60}$$

Population mean.

$$\bar{\tau} = \sum_{g=1}^{G} \omega_g \, \theta_{C,g} = \bar{\theta}_C \tag{61}$$

where $\bar{\theta}_C$ is the population-weighted constraint share.

Within-group variance. Since $\tau_i \mid (i \in g)$ is Bernoulli with parameter $\theta_{C,g}$,

$$\begin{aligned} \sigma_g^2 &= E[\tau_i^2 \mid g] - (E[\tau_i \mid g])^2 \\ &= E[\tau_i^2 \mid g] - \theta_{C,g}^2. \end{aligned} \tag{62}$$

Because $\tau_i \in \{0, 1\}$, we have $\tau_i^2 = \tau_i$, so

$$E[\tau_i^2 \mid g] = E[\tau_i \mid g] = \theta_{C,g}. \tag{63}$$

Substituting:

$$\sigma_g^2 = \theta_{C,g} - \theta_{C,g}^2 = \theta_{C,g}(1 - \theta_{C,g}). \tag{64}$$

This is the variance of a Bernoulli random variable.

Average within-group variance.

$$V_W^{bin} = \sum_{g=1}^{G} \omega_g \, \theta_{C,g}(1 - \theta_{C,g}). \tag{65}$$

Between-group variance.

$$V_B^{bin} = \sum_{g=1}^{G} \omega_g (\theta_{C,g} - \bar{\theta}_C)^2. \tag{66}$$

Total variance. By the same Bernoulli logic applied to the population,

$$\begin{aligned} \sigma_{bin}^2 &= E[\tau_i^2] - (E[\tau_i])^2 \\ &= \bar{\theta}_C - \bar{\theta}_C^2 \\ &= \bar{\theta}_C(1 - \bar{\theta}_C). \end{aligned} \tag{67}$$

Adding-up identity. We now verify that equation (59) holds:

$$V_W^{bin} + V_B^{bin} = \sum_g \omega_g \, \theta_{C,g}(1 - \theta_{C,g}) + \sum_g \omega_g (\theta_{C,g} - \bar{\theta}_C)^2. \tag{68}$$

Expanding the first sum:

$$\sum_g \omega_g \, \theta_{C,g}(1 - \theta_{C,g}) = \sum_g \omega_g \theta_{C,g} - \sum_g \omega_g \theta_{C,g}^2 = \bar{\theta}_C - \sum_g \omega_g \theta_{C,g}^2. \tag{69}$$

Expanding the second sum:

$$\begin{aligned} \sum_g \omega_g (\theta_{C,g} - \bar{\theta}_C)^2 &= \sum_g \omega_g \theta_{C,g}^2 - 2\bar{\theta}_C \sum_g \omega_g \theta_{C,g} + \bar{\theta}_C^2 \sum_g \omega_g \\ &= \sum_g \omega_g \theta_{C,g}^2 - 2\bar{\theta}_C^2 + \bar{\theta}_C^2 \\ &= \sum_g \omega_g \theta_{C,g}^2 - \bar{\theta}_C^2. \end{aligned} \tag{70}$$

Adding equations (69) and (70):

$$\begin{aligned} V_W^{bin} + V_B^{bin} &= \big(\bar{\theta}_C - \sum_g \omega_g \theta_{C,g}^2\big) + \big(\sum_g \omega_g \theta_{C,g}^2 - \bar{\theta}_C^2\big) \\ &= \bar{\theta}_C - \bar{\theta}_C^2 \\ &= \bar{\theta}_C(1 - \bar{\theta}_C) \\ &= \sigma_{bin}^2. \end{aligned} \tag{71}$$

The terms $\sum_g \omega_g \theta_{C,g}^2$ cancel exactly. This establishes the adding-up identity under the binary model:

$$V_W^{bin} + V_B^{bin} = \bar{\theta}_C(1 - \bar{\theta}_C) = \sigma_{bin}^2. \tag{72}$$

D.4 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 $\sigma_g^2 \equiv \text{Var}(\tau_i \mid g) > 0$. We derive an upper bound on the total variance and show that it is strictly less than the Bernoulli benchmark.

Subgroup mean and population mean. These are defined as before:

$$\mu_g = E[\tau_i \mid g], \qquad \bar{\mu} = \sum_g \omega_g \mu_g. \tag{73}$$

Bounding the within-group variance. For any random variable X with support $[a, b]$ and mean μ, the Popoviciu inequality states

$$\text{Var}(X) \leq (b - a)^2/4. \tag{74}$$

A sharper bound uses the mean directly. For $X \in [0, 1]$ with $E[X] = \mu$:

$$\text{Var}(X) = E[X^2] - \mu^2. \tag{75}$$

Since $X \in [0, 1]$, we have $X^2 \leq X$, so $E[X^2] \leq E[X] = \mu$. Therefore:

$$\text{Var}(X) \leq \mu - \mu^2 = \mu(1 - \mu). \tag{76}$$

Equality holds if and only if $P(X \in \{0, 1\}) = 1$, that is, X is Bernoulli. To see this, note that $E[X^2] = \mu$ 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:

$$\sigma_g^2 < \mu_g(1 - \mu_g) \quad \text{for all } g. \tag{77}$$

Aggregate within-group variance. Summing across subgroups:

$$V_W^{sm} = \sum_g \omega_g \sigma_g^2 < \sum_g \omega_g \mu_g (1 - \mu_g) = V_W^{bin}\Big|_{\theta_{C,g} = \mu_g}. \tag{78}$$

The smooth model's within-group variance is strictly less than what the Bernoulli formula would imply at the same subgroup means.

Between-group variance. The between-group variance depends only on the subgroup means and weights:

$$V_B^{sm} = \sum_g \omega_g (\mu_g - \bar{\mu})^2. \tag{79}$$

If the subgroup means are identical under both models ($\mu_g = \theta_{C,g}$), the between-group variance is the same:

$$V_B^{sm} = V_B^{bin}. \tag{80}$$

Total variance. By the law of total variance:

$$\begin{aligned}
\sigma_{sm}^2 &= V_W^{sm} + V_B^{sm} \\
&< \sum_g \omega_g \mu_g (1-\mu_g) + \sum_g \omega_g (\mu_g - \bar{\mu})^2 \\
&= \bar{\mu}(1-\bar{\mu}) \\
&= \sigma_{bin}^2 \Big|_{\bar{\theta}_C = \bar{\mu}}.
\end{aligned} \tag{81}$$

The final equality follows from the adding-up result in equation (71), which is a purely algebraic identity that holds for any set of numbers μ_g and weights ω_g. The inequality is strict because $V_W^{sm} < \sum_g \omega_g \mu_g (1-\mu_g)$.

This establishes that under the smooth model, the total variance of individual treatment effects is strictly less than the Bernoulli benchmark:

$$\sigma_{sm}^2 < \bar{\mu}(1-\bar{\mu}). \tag{82}$$

D.5 Constructing the Observable Diagnostic

Equations (72) and (82) characterize the population-level distinction. We now translate this into a diagnostic computable from published estimates.

The identification problem. A researcher observes subgroup treatment effects $\hat{\tau}_g$ and a full-sample estimate $\hat{\tau}$, but does not observe the individual treatment effects τ_i or their within-group distribution. The within-group variance σ_g^2 is therefore not directly estimable from published subgroup means. We proceed by computing what σ_g^2 *would be* under the binary model and checking whether the resulting decomposition is internally consistent.

Sample variance components. Define the following quantities, all computable from published estimates:

$$\hat{V}_{total} \equiv \hat{\tau}(1 - \hat{\tau}) \tag{83}$$

$$\hat{V}_{within} \equiv \sum_{g=1}^{G} \hat{\omega}_g \, \hat{\tau}_g(1 - \hat{\tau}_g) \tag{84}$$

$$\hat{V}_{between} \equiv \sum_{g=1}^{G} \hat{\omega}_g(\hat{\tau}_g - \hat{\tau})^2. \tag{85}$$

The key step is equation (84). We compute the within-group variance using the Bernoulli formula $\hat{\tau}_g(1-\hat{\tau}_g)$ applied to each subgroup estimate. This is the correct within-group variance if the binary model holds. If the smooth model holds, it overstates the true within-group variance.

Derivation of the diagnostic. Under the binary model, the population identity equation (72) implies

$$\hat{V}_{total} - \hat{V}_{within} - \hat{V}_{between} \xrightarrow{p} 0. \tag{86}$$

Under the smooth model, $V_W^{sm} < V_W^{bin}$ from equation (78), so the Bernoulli-based $\hat{V}_{within}$ overstates V_W^{sm}. However, $\hat{V}_{total}$ also overstates the true total variance σ^2_{sm} via equation (82). Which overstatement is larger?

Under the binary model, the overstatement is zero in both cases. Under the smooth model, define the within-group excess $\delta_W \equiv V_W^{bin} - V_W^{sm} > 0$ and the total excess $\delta_T \equiv \bar{\mu}(1-\bar{\mu}) - \sigma^2_{sm} > 0$. We need to show $\delta_T > \delta_W$, which would imply the diagnostic is positive under the smooth model.

From equation (81), $\sigma^2_{sm} = V_W^{sm} + V_B^{sm}$. From equation (72), $\bar{\mu}(1 - \bar{\mu}) = V_W^{bin} + V_B^{bin}$. With $V_B^{sm} = V_B^{bin}$ from equation (80):

$$\begin{aligned} \delta_T &= \bar{\mu}(1 - \bar{\mu}) - \sigma^2_{sm} \\ &= (V_W^{bin} + V_B^{bin}) - (V_W^{sm} + V_B^{sm}) \\ &= V_W^{bin} - V_W^{sm} \\ &= \delta_W. \end{aligned} \tag{87}$$

The two excesses are identical. Therefore, the diagnostic

$$\begin{aligned}\Delta &\equiv \hat{V}_{total} - \hat{V}_{within} - \hat{V}_{between} \\ &\approx \bar{\mu}(1-\bar{\mu}) - V_W^{bin} - V_B^{bin} = 0 \quad \text{under binary model}\end{aligned} \tag{88}$$

and

$$\begin{aligned}\Delta &\approx \bar{\mu}(1-\bar{\mu}) - V_W^{bin} - V_B^{sm} \\ &= (V_W^{bin} + V_B^{bin}) - V_W^{bin} - V_B^{sm} \\ &= V_B^{bin} - V_B^{sm} = 0 \quad \text{if } \mu_g = \theta_{C,g}.\end{aligned} \tag{89}$$

This calculation reveals an important subtlety. When the Bernoulli formula is applied to compute $\hat{V}_{within}$ and the Bernoulli total variance is used for $\hat{V}_{total}$, the overstatement of within-group variance under the smooth model is exactly offset by the overstatement of total variance. Both the binary and smooth models yield $\Delta = 0$ in population if the subgroup means are the same.[9]

D.6 Where the Diagnostic Has Power

The algebraic cancellation in equation (89) shows that the variance decomposition based solely on subgroup means cannot distinguish the two models. The diagnostic therefore has power through two alternative channels, which we formalize here.

D.6.1 Channel 1: Convergence to Bounds

The binary and smooth models make different predictions about the behavior of subgroup estimates as conditioning becomes finer.

Lemma 1 (Convergence Under the Binary Model) *Under the binary model, let $g(X)$ denote subgroups defined by conditioning variables X. If the partition is sufficiently fine that each subgroup is homogeneous in constraint status, then $\hat{\tau}_g \to c_g$ where $c_g \in \{0,1\}$ for each g. In the limit, $\hat{V}_{within} \to 0$ and $\hat{V}_{between} \to \hat{V}_{total}$.*

[9] This result is algebraically necessary. The law of total variance is an identity. If $\hat{V}_{within}$ is computed as a function of subgroup means only (not of the true within-group distribution), and $\hat{V}_{total}$ is computed from the population mean only, the decomposition $\hat{V}_{total} = \hat{V}_{within} + \hat{V}_{between}$ holds as an algebraic identity regardless of the underlying distribution. The diagnostic's power comes not from this identity but from the auxiliary prediction about convergence to bounds and from the standard error test.

Proof. Under the binary model, $\tau_g = \theta_{C,g}$. If subgroup g contains only constrained firms, $\theta_{C,g} = 1$; if only unconstrained firms, $\theta_{C,g} = 0$. Under a fine partition, $\theta_{C,g}(1 - \theta_{C,g}) \to 0$ for each g, so $V_W^{bin} \to 0$. The total variance $\sigma_{bin}^2 = \bar{\theta}_C(1 - \bar{\theta}_C)$ remains constant, so $V_B^{bin} \to \sigma_{bin}^2$. □

Under the smooth model with continuous individual effects, refining the partition reduces between-individual variation within subgroups but does not drive subgroup means to $\{0, 1\}$. If MR_i takes values like 0.3, 0.5, or 0.7 for firms with similar observables, subgroup means converge to these intermediate values rather than to the boundary.

Define the convergence ratio for subgroup g:

$$\rho_g \equiv \min\bigl(\hat{\tau}_g,\ 1 - \hat{\tau}_g\bigr). \tag{90}$$

This measures the distance of $\hat{\tau}_g$ from the nearest bound. Values near zero indicate proximity to $\{0, 1\}$; values near 0.5 indicate an intermediate estimate. Under the binary model with fine conditioning, $\rho_g \to 0$ for all g. Under the smooth model, ρ_g stabilizes at interior values.

Across the studies surveyed in Section 5, the finest available subgroup estimates yield ρ_g values near zero. The top size deciles in Khwaja and Mian (2008) have $\hat{\tau}_g \approx 0$, giving $\rho_g \approx 0$. Lemmon and Roberts' (2010) junk-rated sample has $\hat{\tau} \approx 0.97$, giving $\rho \approx 0.03$. Investment-grade subsamples across multiple studies yield $\hat{\tau}_g$ near zero. This convergence pattern is predicted by the binary model but not by the smooth model.

D.6.2 Channel 2: Standard Errors and Within-Group Dispersion

If subgroup standard errors and sample sizes are reported, the within-group variance is directly estimable without imposing the Bernoulli assumption. The variance ratio diagnostic exploits this.

Under the binary model, the sampling variance of $\hat{\tau}_g$ in a subgroup of N_g firms is

$$\text{Var}(\hat{\tau}_g)^{bin} = \frac{\theta_{C,g}(1 - \theta_{C,g})}{N_g}. \tag{91}$$

Under the smooth model, the sampling variance is

$$\text{Var}(\hat{\tau}_g)^{sm} = \frac{\sigma_g^2}{N_g} < \frac{\mu_g(1 - \mu_g)}{N_g} = \text{Var}(\hat{\tau}_g)^{bin}\Big|_{\theta_{C,g} = \mu_g}. \tag{92}$$

The inequality follows from equation (77). Define the variance ratio:

$$R_g \equiv \frac{\widehat{\text{se}}_g^{\,2} \cdot N_g}{\hat{\tau}_g(1 - \hat{\tau}_g)}. \tag{93}$$

Under the binary model, $R_g \xrightarrow{p} 1$. Under the smooth model, $R_g \xrightarrow{p} \sigma_g^2/[\mu_g(1 - \mu_g)] < 1$. Define the average ratio:

$$\bar{R} \equiv \sum_{g=1}^{G} \hat{\omega}_g R_g. \tag{94}$$

Values of $\bar{R}$ near one support the binary model; values systematically below one support the smooth model.

Remark 1 *In practice, reported standard errors in IV and DiD settings reflect estimation uncertainty from both the first stage and the reduced form, not just the variance of τ_i. The relationship between $\widehat{\text{se}}_g$ and σ_g^2/N_g therefore holds as an approximation in the absence of weak instruments and specification error. Robust or clustered standard errors incorporate additional sources of variation, which may inflate R_g above one even under the binary model. The test is therefore conservative: $R_g > 1$ is uninformative, but $R_g \ll 1$ across subgroups would provide evidence against the binary specification.*

D.7 Summary of Testable Predictions

Table 8 collects the empirical predictions that distinguish the two models.

Table 8: Empirical Predictions: Binary versus Smooth Model

Observable	Binary model	Smooth model
Subgroup estimates as conditioning refines	Converge to $\{0, 1\}$	Stabilize at interior values
Convergence ratio ρ_g for fine subgroups	$\to 0$ for fine subgroups	Bounded away from zero
Variance ratio R_g	$\to 1$	< 1
Average variance ratio $\bar{R}$	$\bar{R} \approx 1$	$\bar{R} < 1$

Notes: The convergence ratio is $\rho_g \equiv \min(\hat{\tau}_g, 1 - \hat{\tau}_g)$. The variance ratio is $R_g \equiv \widehat{\text{se}}_g^{\,2} \cdot N_g/[\hat{\tau}_g(1 - \hat{\tau}_g)]$. Under the binary model, individual treatment effects are $\{0, 1\}$, so subgroup means are Bernoulli proportions and approach the bounds with fine conditioning. Under the smooth model, individual effects vary continuously, producing interior subgroup means and within-group variance below the Bernoulli benchmark.

The diagnostic's power comes from the convergence and standard error channels rather than from the variance decomposition identity, which holds algebraically under both models when computed from subgroup means. The most informative test uses the finest available subgroup partition together with reported standard errors. The empirical pattern documented in Section 6.2—convergence of subgroup estimates toward the binary bounds across diverse settings—provides cumulative evidence favoring the binary characterization.

E Appendix: Formal Implementation of Diagnostic Tests

This appendix implements the diagnostic tests described in Section 6.2 across the studies surveyed in Table 4. We report formal test statistics and p-values for the convergence-to-bounds test and the variance ratio test. All inputs are drawn from published tables and figures in the original papers, as documented in the footnotes to Table 9.

E.1 Implementation

The convergence ratio ρ_g and variance ratio R_g are defined in equations (23) and (26) of the main text. We briefly restate the testing procedures.

For the convergence test, we pool subgroup level ρ_g values and test $H_0\colon E[\rho_g] \geq 0.25$ against $H_1\colon E[\rho_g] 0.25$, where 0.25 is the expected value of $\min(\mu, 1-\mu)$ when μ is uniformly distributed on $[0, 1]$. This is a conservative benchmark. Under the smooth model with any nondegenerate distribution of conditional means, the expected convergence ratio is bounded away from zero. The test statistic is

$$t_\rho = \frac{\bar{\rho} - 0.25}{s_\rho / \sqrt{G}} \tag{95}$$

where $\bar{\rho}$ and s_ρ are the sample mean and standard deviation of ρ_g across G subgroups, with p-values from the t_{G-1} distribution.

For the variance ratio test, R_g requires both a reported standard error and an effective sample size for the same specification. Where both are available, $R_g \approx 1$ supports the binary model and $R_g < 1$ supports the smooth alternative. As discussed in Section 6.2, the test is conservative because clustered and robust standard errors may inflate R_g above one even under the binary model.

E.2 Results

Table 9 reports both diagnostics for the studies where sufficient information is available from published tables or figures.

Table 9: Diagnostic Test Results Across Studies

Study	Subgroup	$\hat{\tau}_g$	se	N_g	ρ_g	R_g
Panel A: Rauh (2006)						
	Full sample[a]	0.65	0.27	1,522	0.35	—
	Below IG[b]	0.84	—	—	0.16	—
	Investment grade[b]	0.50	—	—	0.50	—
Panel B: Lemmon & Roberts (2010)						
	Full sample[c]	0.97	—	716	0.03	—
Panel C: Khwaja & Mian (2008)						
	Full sample[d]	0.60	0.09	8,500	0.40	—
	Top 3 deciles[e]	0.02	—	—	0.02	—
	Deciles 5–7[e]	0.30	—	—	0.30	—
	Bottom 2 deciles[e]	0.70	—	—	0.30	—
Panel D: Chaney et al. (2012)						
	Full sample[f]	0.06	0.02	50,858	0.06	360.7
	No bond access[g]	0.12	0.04	2,890	0.12	—
	Bond access[g]	0.01	0.03	1,863	0.01	—
Panel E: Chodorow-Reich (2014)						
	Small/medium[h]	Sig.	—	—	—	—
	Large[h]	≈ 0	—	—	≈ 0	—

Notes: $\rho_g \equiv \min(\hat{\tau}_g, 1-\hat{\tau}_g)$ measures distance from the nearest binary bound. $R_g \equiv \widehat{\text{se}}_g^2 \cdot N_g/[\hat{\tau}_g(1-\hat{\tau}_g)]$; values near one support the binary model. "—" indicates that the required information (standard errors or effective sample sizes for the relevant subgroup specification) is not reported in the original paper. R_g is reported only when both $\widehat{\text{se}}_g$ and N_g are available for the same specification. Sig. = statistically significant; IG = investment grade.

[a] Rauh (2006), Table IV, column (2), p. 54. Standard error is for the IV coefficient on mandatory contributions. Sample size is firm-years with nonmissing instruments.
[b] Rauh (2006), Table VI, p. 63. Approximate coefficients from the text discussion of credit-rating interactions. Standard errors and subsample sizes for these splits are not separately reported.
[c] Lemmon and Roberts (2010), Table 5, Panel B, p. 576. The near-unity estimate is inferred from the statement that investment declined "almost one-for-one" with debt issuance (p. 555). Sample size is treated firm-years.
[d] Khwaja and Mian (2008), Table III, column (4), p. 1422. Standard error is clustered at the firm level. N_g is approximate total loan-level observations.
[e] Khwaja and Mian (2008), Figure VII, p. 1430. Approximate values read from the figure. Decile-level standard errors and sample sizes are not reported in tabular form.
[f] Chaney et al. (2012), Table 3, column (1), p. 2397. Standard error clustered at the state level. $N_g = 50{,}858$ firm-year observations.
[g] Chaney et al. (2012), Table 5, columns (1) and (2), p. 2401. Subsample split by bond market access.
[h] Chodorow-Reich (2014), Table 5, p. 30. Qualitative characterization based on the text discussion of size-based heterogeneity.

Convergence test. We pool the 11 subgroup-level ρ_g values from Table 9 for which numerical point estimates are available. The mean convergence ratio is $\bar{\rho} = 0.15$ with a standard deviation of $s_\rho = 0.16$. The test statistic against the uniform benchmark is $t_\rho = (0.15 - 0.25)/(0.16/\sqrt{11}) =$

-2.07, yielding a one-sided p-value of 0.033 from the t_{10} distribution. The data reject the hypothesis that subgroup estimates are spread across the interior of $[0, 1]$ in favor of concentration near the binary bounds.

The convergence pattern is also monotonic within studies. In Khwaja and Mian (2008), ρ_g equals 0.02 for the largest firms, rises to 0.30 for mid-sized firms, and equals 0.30 for the smallest firms. The smallest firms' ρ_g remaining at 0.30 rather than approaching zero reflects the fact that even the bottom deciles contain a nontrivial share of unconstrained firms. Finer conditioning within these deciles would likely push ρ_g closer to zero. In Chaney et al. (2012), the bond access subsample has $\rho_g = 0.01$ and the no bond access subsample has $\rho_g = 0.12$. This is consistent with the former being nearly homogeneously unconstrained and the latter being mixed.

Variance ratio test. The variance ratio R_g is computable for only one specification. The Chaney et al. (2012) full sample, where $R_g = (0.02^2 \times 50{,}858)/(0.06 \times 0.94) \approx 360.7$. This value far exceeds one, which is consistent with the binary model but not with the smooth alternative, which predicts $R_g < 1$. The large magnitude reflects substantial inflation from state-level clustering of standard errors, which incorporates sources of variation beyond the sampling variance of individual treatment effects. As noted in Section 6.2, this makes the test conservative. So $R_g \gg 1$ is uninformative about whether the true R_g equals one. The main point is that R_g is not systematically below one. A more informative application would require heteroskedasticity-robust standard errors at the firm level rather than clustered standard errors, or access to the underlying microdata.

E.3 Limitations Specific to the Tests

Two limitations of the implementation deserve emphasis. First, several inputs to the convergence test are approximate values read from published figures rather than exact point estimates from regression tables. This introduces measurement error into the pooled test statistic. Treating the Khwaja and Mian (2008) decile level estimates as precise overstates the effective information in the test.

Second, the variance ratio test is severely limited by data availability. Most papers do not simultaneously report standard errors and effective sample sizes at the subgroup level for the specifications most relevant to the diagnostic. The single computable R_g value is suggestive. But clearly it cannot support a formal test of the smooth alternative. Both limitations could be addressed with access to the underlying microdata, which would permit exact computation of subgroup estimates, standard errors, and sample sizes at arbitrary levels of partition fineness.