POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

The Economic Consequences of Shareholder Breadth: Evidence on Operating Performance and Market Valuation

Tesi di Laurea Magistrale in
Management Engineering - Ingegneria Gestionale

Author: **Tommaso Rossi**

Student ID: 10711335
Advisor: Benedetta Montanaro
Co-advisor: Andriy Bodnaruk
Academic Year: 2024-2025

To my Mum and Dad,

With gratitude for everything you have given me and all you have made possible.

Abstract

This thesis investigates whether investor recognition, proxied by shareholder breadth, is systematically associated with corporate performance and market valuation. In incomplete information settings, limited investor participation can generate pricing distortions and segmentation effects. Expanding the shareholder base may reduce participation frictions, improve risk sharing and enhance information aggregation. The central research question is therefore whether increases in shareholder breadth are reflected in higher operating profitability and higher market valuation within firms over time.

To address this question, a firm-year panel dataset is constructed combining accounting data and ownership information. Shareholder breadth is operationalized as a participation-based measure derived from the number of shareholders and outcomes are measured through return on assets (ROA) and Tobin's Q. The empirical strategy integrates non-parametric diagnostics and firm- and year-fixed-effects regressions. Visual analysis, rank-based correlations and permutation tests provide distribution-robust evidence on unconditional associations, while parametric models with fixed effects and clustered standard errors identify within-firm relationships, accounting for time-invariant heterogeneity and common shocks. Non-linear specifications, including quadratic terms and spline functions, are employed to explore potential convexities in the recognition–valuation link.

The results indicate that shareholder breadth is only weakly related to operating profitability, whereas it exhibits a more consistent association with market valuation once firm fundamentals such as size and leverage are controlled for. Evidence of mild

convexity suggests that marginal recognition effects may intensify at higher levels of participation. Overall, the findings support the interpretation that investor recognition operates primarily through pricing channels rather than through direct improvements in accounting performance.

By implementing a transparent and symmetric empirical framework across performance and valuation outcomes, this thesis contributes to the literature on investor recognition and participation-based asset pricing. It provides structured evidence on how changes in the shareholder base are reflected in firm-level outcomes and offers a replicable methodology for studying recognition dynamics in panel data settings.

Key-words: Investor recognition, Shareholder breadth, Tobin's Q, Return on assets, Fixed-effects models, Participation frictions.

Abstract in italiano

Questa tesi analizza se il riconoscimento da parte degli investitori, misurato tramite la breadth azionaria, sia sistematicamente associato alla performance aziendale e alla valutazione di mercato. In contesti caratterizzati da informazione incompleta, una partecipazione limitata degli investitori può generare distorsioni nei prezzi e fenomeni di segmentazione. L'ampliamento della base azionaria può ridurre le frizioni di partecipazione, migliorare la condivisione del rischio e rafforzare i processi di aggregazione dell'informazione. La domanda di ricerca centrale è quindi se gli incrementi della breadth azionaria si riflettano in una maggiore redditività operativa e in una più elevata valutazione di mercato all'interno delle imprese nel tempo.

Per rispondere a tale quesito, viene costruito un dataset panel a livello impresa-anno che combina dati contabili e informazioni sulla proprietà. La breadth azionaria è operazionalizzata come misura di partecipazione derivata dal numero di azionisti, mentre gli outcome sono misurati tramite il return on assets e il Tobin's Q. La strategia empirica integra analisi non parametriche e modelli con effetti fissi di impresa e anno. L'analisi grafica, le correlazioni basate sui ranghi e i permutation test forniscono evidenza robusta alle distribuzioni sulle associazioni incondizionate, mentre i modelli parametrici con effetti fissi ed errori standard clusterizzati identificano relazioni intra-impresa, controllando per eterogeneità non osservata e shock comuni nel tempo. Specificazioni non lineari, incluse componenti quadratiche e funzioni spline, sono utilizzate per esplorare eventuali convessità nel legame tra riconoscimento e valutazione.

I risultati indicano che la breadth azionaria è solo debolmente correlata alla redditività operativa, mentre mostra un’associazione più coerente con la valutazione di mercato una volta controllate variabili fondamentali quali dimensione e leva finanziaria. L’evidenza di una lieve convessità suggerisce che gli effetti marginali del riconoscimento possano intensificarsi a livelli più elevati di partecipazione. Nel complesso, i risultati supportano l’interpretazione secondo cui il riconoscimento degli investitori opera principalmente attraverso canali di pricing piuttosto che tramite miglioramenti diretti della performance contabile.

Attraverso l’adozione di un framework empirico trasparente e simmetrico applicato sia alla performance sia alla valutazione, questa tesi contribuisce alla letteratura sul riconoscimento degli investitori e sull’asset pricing basato sulla partecipazione. Essa fornisce evidenza strutturata su come le variazioni nella base azionaria si riflettano negli outcome a livello di impresa e propone una metodologia replicabile per lo studio delle dinamiche di riconoscimento in contesti panel.

Parole chiave: Investor recognition, Shareholder breadth, Tobin’s Q, Return on assets, Modelli a effetti fissi, Frizioni di partecipazione

Contents

1 Introduction

1.1. Topic and scope of the Thesis

This thesis studies the role of investor recognition in equity markets and its association with firm-level outcomes. Investor recognition refers to the extent to which a firm is known, followed and held by market participants. In settings where information acquisition is costly and participation is incomplete, not all firms are equally visible to investors and differences in recognition can translate into differences in ownership structure, pricing and real outcomes.

The analysis focuses on shareholder breadth, defined as the size of the shareholder base, as an observable proxy for investor recognition. Shareholder breadth captures how widely a firm's equity is held across investors and reflects the degree of market participation in that firm. A broader shareholder base implies that more investors are aware of, willing to hold and able to trade the firm's stock, while a narrow base reflects limited recognition and participation.

The thesis examines how changes in shareholder breadth are related to two distinct but complementary dimensions of firm outcomes. The first is operating performance, measured through accounting-based indicators such as return on assets. The second is market valuation, measured through Tobin's Q. Studying these two endpoints jointly allows the analysis to distinguish between realized profitability and market expectations and to assess whether recognition is associated primarily with operating efficiency, with valuation or with both.

The scope of the thesis is empirical and focuses on firm-year variation in a large panel of publicly listed firms. The objective is not to test a single structural model, but rather to document robust empirical regularities that are consistent with recognition-based theories of asset pricing and corporate finance, while carefully addressing measurement, distributional and identification challenges.

1.2. Academic and practical relevance

From an academic perspective, investor recognition occupies a central position in theories of incomplete information and limited participation. Classical asset pricing models assume that all investors are aware of all securities and can costlessly hold any asset. In practice, investors face attention constraints, information acquisition costs and institutional frictions that limit participation. As a result, some firms are effectively invisible to a large fraction of the investor population.

Recognition-based theories predict that limited participation can affect equilibrium prices, expected returns and risk sharing. Firms with low recognition may face a higher cost of capital because their risk is borne by a smaller set of investors. Conversely, increases in recognition may improve risk sharing, liquidity and pricing efficiency.

These mechanisms suggest that recognition should matter not only for returns, but also for valuation levels and potentially for real firm outcomes through financing conditions and managerial incentives.

Despite its theoretical importance, investor recognition is difficult to observe directly. Much of the empirical literature relies on indirect proxies such as analyst coverage, media exposure, advertising or inclusion in indices. Shareholder breadth offers a complementary perspective, as it directly reflects realized investor participation rather than attention or information supply alone.

From a practical standpoint, the topic is relevant for corporate managers, investor relations professionals and policymakers. Firms actively invest resources in visibility-enhancing activities, such as disclosures, roadshows and marketing, with the expectation that broader investor participation will translate into tangible benefits. Understanding whether and how shareholder breadth is associated with performance and valuation outcomes provides empirical grounding for these practices.

Moreover, the distinction between operating performance and market valuation is crucial for practitioners. A broader shareholder base may raise valuation through demand-side effects even in the absence of immediate improvements in profitability or it may coincide with performance improvements if recognition facilitates access to capital or improves governance. Disentangling these channels is therefore of both academic and practical interest.

1.3. Related literature and research gap identification

A substantial body of literature examines the implications of investor recognition and limited participation for asset prices. Early theoretical contributions show that when investors are unaware of some securities, equilibrium prices reflect incomplete risk sharing and segmented demand. Subsequent empirical work documents that firms with lower recognition tend to earn higher average returns, consistent with a recognition premium.

Empirical proxies for recognition vary widely. Analyst coverage has been used to capture information production, media coverage to proxy for attention and advertising expenditure to measure visibility. Ownership-based measures, such as the number of shareholders or institutional ownership breadth, have been used to capture participation more directly. These studies generally focus on returns, liquidity or cost of capital.

Much less attention has been paid to the relationship between investor recognition and operating performance. When performance is considered, it is often treated as a control rather than as an outcome of interest. Similarly, valuation is frequently analyzed in isolation, without explicit comparison to accounting-based performance measures.

Two gaps therefore motivate this thesis. First, there is limited evidence on how recognition, proxied by shareholder breadth, is associated with both operating performance and valuation within a unified empirical framework. Second, existing studies often impose linear relationships, even though recognition effects may be nonlinear, for example if marginal recognition becomes more valuable once a firm reaches a critical mass of investors.

This thesis addresses these gaps by analysing performance and valuation side by side, using identical timing conventions, comparable cleaning rules and parallel empirical designs. It further allows for nonlinear relationships by considering quadratic and spline-based specifications and complements parametric models with non-parametric evidence.

1.4. Research Gap and Research Questions

A substantial literature links investor recognition and information frictions to pricing, liquidity and the cost of capital. Classic incomplete-information models predict that when a firm is recognized by a wider clientele, idiosyncratic risk is shared across more balance sheets and required returns fall (Merton, 1987). Empirical work anchors this view using proxies such as analyst coverage, index inclusions, advertising intensity and norms that shape investor clientele, documenting lower expected returns or higher valuation where recognition is greater (e.g., Chen, Hong, & Stein, 2002; Lehavy & Sloan, 2008; Hong & Kacperczyk, 2009; Lou, 2014). Liquidity-based channels provide a complementary rationale: when participation deepens, price impact tends to decline and expected returns fall, consistent with illiquidity premia (Amihud, 2002; Acharya & Pedersen, 2005). At the same time, corporate outcomes are not solely market-based. If financing conditions and investor scrutiny feed back into real decisions, recognition could co-move with operating performance as well.

Despite these advances, two gaps remain. First, recognition is rarely measured directly through the dispersion of the shareholder base itself. Much of the empirical evidence relies on proxies, sell-side coverage, index membership, name "fluency," or product-market advertising, that correlate with investor attention but are not the investor base. Shareholder counts, as reported by Compustat, offer a transparent, firm-year measure of how many holders are actually on the register. Yet these data are noisy and incomplete and the literature has not settled on a standard, replicable pipeline to clean, transform and deploy the series in panel designs. Second, the shape of the recognition–outcome relation is often assumed rather than tested. Theory implies diminishing marginal benefits: at low recognition, extra breadth should matter more than at already high recognition. Flexible empirical tests that compare linear, quadratic and spline forms within a unified fixed-effects framework remain scarce, particularly when the same breadth construct is related both to operating profitability and to valuation on harmonized samples.

This thesis addresses both gaps. It builds a long-horizon panel of North American firms (2000–2024), constructs breadth from shareholder counts using conservative, auditable treatments of missingness and scale and studies two outcome domains in parallel: operating performance, via return on assets (ROA) and market valuation, via Tobin's Q. The empirical design proceeds from distribution-free corroboration (rank correlations and permutation tests) to firm- and year-fixed-effects regressions that allow for linear and flexible (quadratic and B-spline) associations, with transparent, economically motivated controls for size and liabilities. The focus is on within-firm variation over time, so that persistent firm heterogeneity and common shocks are absorbed and the shape of the relation is read from the data rather than imposed.

Guided by this motivation, the research is organized around three questions:

- RQ1 – Performance: Is a broader shareholder base associated with higher operating performance (ROA) within firms over time?
- RQ2 – Valuation: Is a broader shareholder base associated with higher valuation (Tobin's Q) within firms over time?
- RQ3 – Shape: Is the relationship between breadth and each outcome non-linear, specifically, are marginal associations economically larger at low breadth and weaker at high breadth?

These questions map into testable hypotheses that hold firm and time effects constant:

- H1) breadth is positively associated with ROA
- H2) breadth is positively associated with Tobin's Q
- H3) the association is concave, with diminishing marginal effects as breadth rises.

By placing the same breadth construct against both outcomes in a single panel framework and by explicitly comparing linear and flexible functional forms, the study seeks to clarify where along the breadth distribution economically meaningful co-movement resides and whether it differs for operating versus market-based endpoints.

1.5. Contributions of this Research

1.5.1. Contributions to Academics

This thesis advances the recognition–outcomes literature along three fronts.

First, it reframes "investor recognition" as a directly observed participation margin by constructing shareholder-base breadth from Compustat holder counts through a transparent, auditable pipeline. Rather than relying on indirect attention proxies such as analyst coverage, index inclusions or advertising, breadth is measured as the dispersion of actual equity holders and then mapped, carefully and reproducibly, into estimators suited for panel inference. This measurement shift matters conceptually because it aligns the empirical object with theory that links clientele size and information reach to required returns (e.g., Merton, 1987; Easley and O'Hara, 2005; Lambert, Leuz and Verrecchia, 2007) and it matters empirically because it yields a recognition variable that can be replicated and stress-tested across time and specifications.

Second, it delivers a unified identification design that studies operating performance and valuation on harmonized samples with the same breadth construct, the same treatment of outliers and zeros and the same two-way fixed-effects architecture. By moving from distribution-free diagnostics to linear and flexible (quadratic and spline) forms within a single framework, the analysis tests, rather than assumes, the shape of the relation between breadth and outcomes. This directly addresses a common gap in prior work: theory implies diminishing marginal effects of recognition, but evidence on where concavity sets in has often been indirect or proxy-driven. Reading the shape from the data, while holding firm and year effects constant, helps disentangle broad participation from time-invariant firm quality and economy-wide shocks and clarifies why associations should be stronger for valuation, where recognition and liquidity premia are expected to bite (Amihud, 2002; Acharya and Pedersen, 2005, than for contemporaneous profitability.

Third, it contributes a coherent, replicable empirical blueprint. The construction of breadth variants (levels and log maps that are robust to zeros and boundary masses), the disciplined use of identical row sets across visuals and regressions and the side-by-side treatment of ROA and Tobin's Q create a research template that other scholars can adopt to revisit recognition mechanisms with public data. In doing so, the thesis helps bridge two strands of evidence, participation/attention effects on expected returns (e.g., Chen, Hong and Stein, 2002; Lehavy and Sloan, 2008; Hong and Kacperczyk, 2009; Lou, 2014) and information/liquidity channels in asset pricing, within a single, transparent panel design that can be extended to alternative markets, horizons and outcomes.

1.5.2. Contributions to Practitioners

For investor-relations teams, corporate finance functions and market participants, the thesis offers a practical lens on a frequently discussed but loosely measured construct. The breadth measures quantify how widely a firm's equity is held and translate that dispersion into variables that are stable enough for monitoring and benchmarking across time. Because the pipeline preserves an audit trail (including flags for filled values and clear treatment of zeros and boundary masses), firms can track breadth as a KPI alongside coverage, turnover and liquidity.

For capital-markets practitioners, the results inform expectations about where changes in breadth matter economically. If marginal associations concentrate at low breadth and flatten at high breadth, initiatives that expand the investor base, disclosure improvements, IR outreach, index candidacy, branding choices that raise salience, are likely to deliver the largest valuation payoffs for firms that are initially under-recognized, while incremental gains may be modest for already widely held issuers. Conversely, if conditional links to contemporaneous ROA are small once scale and liabilities are held constant, breadth should not be overstated as a driver of operating profitability in the short run; it is better viewed as a recognition/liquidity state variable whose primary impact is expected on pricing rather than on immediate accounting outcomes.

For policy and standards audiences, the study illustrates how public fundamentals can be leveraged to build recognition metrics without relying on proprietary holdings data. The approach is implementable with standard disclosures, facilitating cross-firm comparability and historical back-testing. More broadly, the findings help bridge the vocabulary of investor recognition and the operational language of financial reporting: breadth can be tracked in the same dashboards as assets, liabilities and valuation multiples, sharpening discussions about attention, clientele and cost of capital in settings where granular microstructure data are not available.

1.6. Data construction and empirical approach

To address the research questions, the thesis constructs a firm-year panel dataset combining accounting information, market data and ownership measures. Shareholder breadth is measured using the number of shareholders, transformed where appropriate to account for skewness and to exclude mechanically defined cases.

Operating performance is measured using return on assets, defined as net income divided by total assets, with attention paid to the distributional properties of accounting data and the presence of extreme values. Valuation is measured using Tobin's Q, constructed as the ratio of market capitalization to total assets, after reconciling alternative market value sources.

The empirical strategy combines two complementary approaches. Non-parametric analyses rely on visual diagnostics, rank-based correlations and permutation tests to assess monotonic associations without imposing functional-form assumptions. Parametric analyses rely on firm- and year-fixed-effects regressions, which exploit within-firm variation over time while controlling for unobserved heterogeneity and common shocks.

To ensure comparability, performance and valuation analyses adopt parallel design choices, including similar trimming or winsorization rules, consistent timing and identical fixed-effects structures. Nonlinearities are explored using quadratic terms and spline functions, allowing the data to inform the shape of the recognition–outcome relationship.

1.7. Summary of main findings

The empirical results yield several consistent patterns. For operating performance, both non-parametric and parametric evidence indicate a positive association between shareholder breadth and profitability within firms, although the magnitude of the effect is modest and the relationship exhibits some curvature.

For valuation, the results are more sensitive to model specification. Breadth-only models yield weak or negative associations, reflecting the strong correlation between recognition, firm size and capital structure. Once size and leverage are controlled for, the association between recognition and valuation becomes positive on average, with evidence of convexity suggesting that recognition effects strengthen at higher levels of breadth.

Across both outcomes, non-parametric tests confirm that the detected associations are unlikely to arise from random rank correlations, while fixed-effects models show that the results are not driven by time-invariant firm characteristics. Taken together, the findings support the view that investor recognition is economically relevant for both performance and valuation, but that its effects depend on the outcome considered and on proper control for confounding factors.

1.8. Structure of the thesis

The remainder of this thesis is structured as follows.

Chapter 1 introduces the topic and clarifies why investor recognition is a first-order concept in equity markets where awareness and participation are costly. It motivates the analysis from both academic and practical perspectives, reviews the central intuition behind recognition-based pricing mechanisms, formulates the research questions and provides an overview of how the dataset is assembled and how the empirical strategy is designed to address identification and comparability concerns.

Chapter 2 presents a focused review of the theoretical and empirical literature on investor recognition, incomplete information and ownership breadth. The chapter starts from foundational participation-based frameworks and then synthesizes the empirical evidence on shareholder base measures, attention and visibility proxies and return-based tests that interpret recognition as a priced friction. Special attention is given to the distinction between recognition, information production and monitoring and to the channels through which breadth may affect outcomes, including risk sharing, liquidity, financing conditions and managerial incentives. The chapter closes by identifying the specific gap addressed by this thesis, namely the joint analysis of operating performance and valuation in a common design and the limited evidence on nonlinearities in the recognition–outcome mapping.

Chapter 3 describes the data sources, sample construction and variable definitions. It details how the panel is assembled at the firm-year level, how key identifiers and fiscal-year timing are aligned and which filters are applied to ensure internally consistent ratios and economically meaningful support. The chapter then explains the construction of the core variables used throughout the thesis. Recognition is measured via shareholder breadth in both levels and logarithmic form, accounting performance is measured via return on assets and valuation is measured via Tobin's Q built from market capitalization and total assets. Particular emphasis is placed on the logic behind the cleaning rules, the treatment of extreme values and the reason why comparable timing and transformations are required to make performance and valuation results interpretable side by side.

Chapter 4 outlines the empirical strategy and provides the roadmap for the results sections. The chapter explains why the analysis combines non-parametric evidence and parametric fixed-effects models. The non-parametric component is motivated as a low-assumption diagnostic tool that focuses on rank-based association and does not rely on functional-form choices, while the fixed-effects framework is motivated as the primary design to study within-firm movements in recognition and outcomes, controlling for persistent heterogeneity and common time shocks. The chapter also discusses the rationale for exploring nonlinear specifications, including quadratic terms and spline-based approximations and clarifies how inference is conducted using firm-clustered standard errors. The objective is to make clear, before presenting results, which empirical patterns should be expected under the recognition hypothesis and which patterns would call for caution or alternative interpretations.

Chapter 5 reports the empirical evidence on operating performance. It begins with visual diagnostics and non-parametric tests that examine whether higher breadth is associated with higher profitability in a way that is detectable without parametric structure. It then introduces firm- and year-fixed-effects regressions, first in breadth-only form to establish baseline within-firm co-movement and then in richer specifications with controls to separate recognition from scale and financing effects that co-move with performance. The chapter further compares linear and nonlinear parameterizations to assess whether the marginal association between recognition and profitability changes along the distribution of shareholder breadth and it concludes by consolidating the non-parametric and parametric evidence into a coherent interpretation of what the data suggest for the performance channel.

Chapter 6 presents the valuation analysis and mirrors the structure of Chapter 5 to ensure direct comparability. It motivates Tobin's Q as the market-based endpoint most naturally linked to recognition in incomplete-information settings, then documents the distributional challenges of Q and the need for year-wise tail harmonization. The chapter reports non-parametric association patterns using both raw Q and transformed variants and then estimates fixed-effects models for log-transformed Q under alternative functional forms. A key focus is placed on reconciling breadth-only and controlled designs, because valuation is strongly intertwined with size and capital structure. The chapter therefore emphasizes interpretation of coefficients in the presence of controls and discusses how nonlinearity may signal diminishing or increasing marginal recognition effects on valuation.

Finally, Chapter 7 concludes by summarizing the research questions and the principal empirical findings and by relating them back to the literature on investor recognition and participation frictions. It revisits the thesis' contributions in light of the results, discusses the main limitations that arise from measurement constraints and identification challenges in observational panel data and outlines avenues for future research. These extensions may include alternative recognition proxies, richer ownership composition measures, cross-sectional heterogeneity in recognition effects and designs that more directly connect recognition shifts to exogenous variation in information and participation.

2 Literature Review

2.1. Investor Recognition, Information Frictions and Asset Pricing

A central premise of this thesis is that breadth of the shareholder base is an observable footprint of investor recognition. In incomplete-information settings, recognition determines a firm's effective investor clientele. When recognition is limited, fewer investors hold or even consider holding the stock, idiosyncratic risk is shared across a smaller set of balance sheets and investors demand a higher expected return. Conversely, broader recognition enlarges the clientele, improves risk sharing and lowers the required return; with cash flows held constant, valuation rises (Merton, 1987). This logic provides a direct theoretical link from breadth to Tobin's Q (H2). It also motivates a focus on within-firm variation: changes in recognition for the same firm over time should map into changes in pricing, conditional on time-invariant firm characteristics and common shocks. At the same time, the recognition mechanism is likely to be non-linear: marginal benefits of expanding the investor base should be stronger at low initial breadth and attenuate at high breadth, which justifies testing for concavity (H3).

Recognition interacts with the information environment. When information is unevenly distributed (i.e., a large share is private) uninformed investors face an "information risk" premium that lifts expected returns (Easley and O'Hara, 2005).

Improving the publicness and precision of information reduces that premium and can therefore lower the cost of capital. Relatedly, disclosure quality can affect pricing both directly, by altering investors' assessments of the covariance structure of cash flows with systematic factors and indirectly, by influencing firms' real decisions that feed back into risk and cash flows (Lambert, Leuz and Verrecchia, 2007). These frameworks support interpreting breadth as a reduced-form proxy for how widely a firm is followed and how public its information set effectively is. They also motivate the non-linearity tests: informational improvements (or attention gains) should matter most when a firm starts from a low base of recognition, consistent with concave associations between breadth and outcomes (H3). While these models speak most directly to pricing (H2), they also suggest a weaker pathway to operating performance (H1) if financing conditions and market scrutiny influence resource allocation and managerial discipline.

Recognition plausibly improves secondary-market conditions. Price impact (illiquidity) is positively related to expected returns (Amihud, 2002) and a formal liquidity-adjusted asset pricing framework shows that both expected illiquidity and systematic liquidity risk command premia (Acharya and Pedersen, 2005). If broader breadth thickens order books and reduces price impact, the cost of capital falls and Q rises. In this thesis, liquidity is treated as an auxiliary channel rather than the primary object of study: the tests are structured around breadth itself (measured directly) and its association with Q and ROA under firm and year fixed effects. This positioning avoids over-interpreting liquidity mechanics while preserving a coherent rationale for why breadth and valuation move together.

2.2. Participation, Breadth and Market-Based Evidence

Empirical work on investor attention and market participation provides external validity for breadth as an economically meaningful state variable. Early "neglected firm" studies show that firms receiving less attention earn higher subsequent returns, consistent with a recognition premium (Arbel, 1982; Arbel, Carvell and Strebel, 1983). More directly, breadth of ownership, how many investors hold a stock, predicts future returns: when breadth narrows, subsequent returns tend to be higher, especially when short-sale constraints limit arbitrage, which is consistent with recognition/liquidity frictions being priced (Chen, Hong and Stein, 2002). Broader constructs of recognition, such as analyst coverage, index inclusions or disclosure, also correlate with lower expected returns (Lehavy and Sloan, 2008). Investor norms and clientele constraints further restrict who can hold certain stocks; these limits affect pricing and by implication, valuation (Hong and Kacperczyk, 2009). Together, this evidence supports the view that participation on the investor margin is priced, which underpins the thesis expectation that broader breadth is associated with higher Q (H2) and potentially with modest gains in ROA (H1).

Exogenous or quasi-exogenous shocks to recognition help isolate the role of visibility. Corporate name "fluency" provides a salient example: firms with more pronounceable, fluent names exhibit broader ownership, higher turnover, lower price impact and higher valuation multiples; when firms change names to more fluent variants, breadth, liquidity and valuation subsequently rise (Green and Jame, 2013). Product-market advertising operates similarly by amplifying investor attention; increased advertising raises salience among investors and is associated with valuation effects consistent with a recognition channel (Lou, 2014). These studies do not speak to the precise mechanism for every firm-year in this sample, but they reinforce the interpretation of breadth as an outcome-relevant dimension of recognition and provide plausibility for non-linear effects (H3), with larger gains when recognition is initially low.

2.3. Breadth, Ownership Structure and Corporate Outcomes

Breadth (how many investors hold the stock) is conceptually distinct from ownership concentration (how shares are distributed among top holders). The ownership–performance literature highlights that simple cross-sectional relationships often vanish once endogeneity is addressed: ownership structure and firm characteristics are jointly determined and robust structural links with performance are elusive (Demsetz and Villalonga, 2001). This distinction justifies focusing on breadth as a participation margin and using a within-firm, fixed-effects design in the empirical sections. It also clarifies interpretation: results here speak to the dispersion of the investor base, not to control rights or blockholder incentives, reducing conflation between breadth and governance concentration.

The composition and width of the shareholder base can correlate with corporate policies. Firms with narrow bases tend to exhibit different payout behavior, consistent with recognition and clientele frictions shaping financial policies (Bodnaruk and Östberg, 2013). While policy outcomes are not the primary endpoints of this thesis, this evidence provides an intuitive bridge to operating performance: if financing frictions relax as breadth expands, some firms may adjust investment or payout in ways that, over time, register in accounting outcomes such as ROA. The empirical expectation is therefore a clearer association with Q (pricing) and a more modest, possibly heterogeneous link with ROA (operations), in line with H1–H2

2.4. Synthesis and Implications for the Empirical Design

The literature jointly implies three testable implications that guide the empirical strategy. First, broader recognition, proxied here by shareholder-base breadth, should be associated with a lower cost of capital and hence, higher Tobin's Q within firms over time (H2). Second, if relaxed financing conditions and stronger market scrutiny shape decisions at the margin, breadth may be positively associated with ROA, though the effect is a priori weaker than for Q (H1). Third, mechanisms grounded in recognition and information predict diminishing marginal effects: gains from expanding breadth should be larger at low initial recognition and smaller at high recognition, warranting flexible specifications (quadratic and spline) to test for concavity (H3). These implications motivate the use of firm and year fixed effects, identical sample construction across non-parametric and regression analyses and careful attention to functional form in the empirical chapters that follow.

3 Data and Variable Construction

3.1. Introduction to the dataset

The empirical work draws on Compustat North America annual fundamentals over fiscal years 2000–2024. The unit of observation is the firm–year (gvkey, fyear). The raw extract contains 315,113 observations and 50 variables spanning accounting items (assets, liabilities, equity, revenues, cash, net income), market fields (price, shares, reported market value) and the shareholder-count series cshr that proxies breadth of the shareholder base. The overarching design is to place breadth, performance and valuation on a single, comparable accounting template and to retain exactly one observation per firm–year, so that all subsequent transformations and tests operate on a coherent panel.

3.2. Row cleaning

The objective of row cleaning is to place every observation on a single, comparable accounting template and to ensure a unique firm–year record. Compustat exposes multiple statement variants for the same issuer and fiscal year; in the raw extract many firm–years appear more than once because both the industrial (INDL) and financial-services (FS) templates can be present. Retaining both would generate duplicate firm–years and mix fundamentally different balance-sheet structures, undermining comparability for ROA and Tobin's Q. For this reason, only industrial statements are kept (indfmt = INDL). The choice is both conceptual and practical: industrial statements align assets, liabilities and income-statement categories with the definitions used for performance and valuation, while the FS template reflects sector-specific accounting for financial institutions that is not commensurate with industrial firms. Empirically, this restriction removes the large block of duplicate rows that arise when the same gvkey–fyear is reported under both templates.

In addition to the template restriction, three standard Compustat flags are imposed to harmonize reporting: consolidated accounts (consol = C) to avoid mixing parent-only and consolidated statements; standard format (datafmt = STD) to exclude special filing formats; and domestic population source (popsrc = D) to keep the North American scope of the study well defined. These filters eliminate residual heterogeneity in statement construction while preserving the coverage relevant for the hypotheses.

After aligning templates and flags, remaining duplicates at the firm–year level reflect multiple filings for the same fiscal year (for instance, amendments). Within each (gvkey, fyear) group, the observation with the most recent statement date (datadate) is retained, which privileges the latest available accounting revision and yields exactly one record per firm–year.

Finally, fiscal year 2024 is excluded from the working panel for the main analyses because key variables show an abrupt rise in missingness that is not present in prior years. For example, in 2024 the share of missing observations reaches approximately 62.7% for the shareholder-count series, around 51.5% for earnings per share and roughly 43% for core balance-sheet items such as total assets, equity and revenues.

Removing 2024 avoids mechanically shrinking effective sample sizes in the fixed-effects specifications and prevents undue leverage from year-specific data gaps.

Applying these steps reduces the dataset from 315,113 observations and 50 variables to 273,999 observations on the same 50 variables, with a single, harmonized industrial statement per firm–year and no remaining duplicates at the (gvkey, fyear) key.

3.3. Column cleaning

After resolving the row dimension, the dataset is pruned of fields that are either purely administrative or too sparsely populated to be usable in the present analysis. The guiding principle is to retain all quantitative items required to construct breadth, performance and valuation measures, while removing variables that do not enter the construction or estimation stages and would only add noise and memory footprint. Two groups of variables are dropped.

i. *Administrative and descriptive fields*

These variables are not used downstream and after the row filters described in Section 3.2, are either constant by construction or informationally redundant.

Specifically:

- *indfmt*: financial-statement presentation format; fixed after restricting the sample to INDL

- *consol*: consolidation flag; fixed after restricting to consolidated statements
- *datafmt*: data format code; fixed after restricting to STD
- *popsrc*: population source; fixed after restricting to D
- *costat*: data status indicator
- *conm*: company name; a textual identifier not used in estimation
- *curcd*: currency code; not required because all outcome variables are ratios and the market-value/book-value construction is internally currency-consistent within firm-years
- *busdesc*: free-text business description
- *datadate*: accounting date used for de-duplication in Section 3.2; once the latest filing is selected within each firm-year, this timestamp is no longer required for modeling

ii. *Variables with high missingness*

These variables exhibit very limited coverage, do not enter the study's core constructions and would compromise sample stability if retained.

Specifically:

- *xad*: advertising expense; available for only a minority of firm-years
- *xintd*: interest and related debt expense detail; coverage is low and interest components are not modeled
- *tie*: leverage/coverage metric with sparse reporting, not used in the main outcomes

- *ipodate*: IPO date; not required once the panel is formed and de-duplicated
- S&P supplemental "spc" variables (e.g., *spce*, *spced*, *spceeps*, *spcsrc*, *spcindcd*)

This trimming step reduces the column count from 50 to 33 while leaving intact all quantitative fields used in the constructions that follow. In particular, all items required for breadth (cshr), the accounting aggregates used to compute ROA (ni and at) and the market-value components used for Tobin's Q (mkvalt, csho, prcc_f, together with at) are retained, as are the auxiliary totals employed for size and leverage controls (at, lt, ceq, ch, revt). The resulting "slim" dataset contains 273,999 observations and 33 variables and constitutes the working base for the shareholder-breadth treatment, variable engineering and estimation in Sections 3.4–3.6.

3.4. Treatment of the shareholder-count series (CSHR)

This section documents how the Compustat shareholder-count series (cshr) is prepared for analysis. The aim is to obtain a panel that preserves within-firm time variation, avoids imputations over long gaps and carries forward an explicit audit trail for any filled values.

The analysis starts from the industrial-only panel described in Sections 3.2–3.3. Within this dataset, the key ownership variable, cshr, exhibits substantial missingness, with coverage varying both across firms and over time (on average, 52.61% of firm-year observations are missing). Because the empirical strategy exploits within-firm variation over time, incomplete time series pose a non-trivial threat to identification.

3.4.1. Shareholder-count screening

To address this issue, firms are screened based on two simple but economically meaningful time-series criteria, computed after sorting observations by fiscal year within each gvkey:

i. *Coverage*

Coverage is defined as the fraction of firm-year observations for which cshr is non-missing. Firms are required to have coverage of at least 60%, ensuring that a majority of the time series is observed and that ownership dynamics are anchored in actual data rather than imputed values.

ii. *Maximum consecutive missing span*

For each firm, we compute the longest run of consecutive fiscal years in which cshr is missing. Firms are retained only if this maximum gap does not exceed two consecutive years, ruling out series with extended interruptions.

These criteria serve complementary purposes. The coverage requirement guarantees sufficient overall information content within each firm's time series, while the restriction on consecutive gaps prevents situations in which filling missing observations would require long-range interpolation. Such interpolation would be particularly fragile in a fixed-effects framework, where identification relies on short- to medium-horizon within-firm changes.

Taken together, this filtering procedure strikes a deliberate balance between sample size and measurement quality: it preserves firms with sufficiently dense ownership histories to support credible estimation, while excluding those for which missingness would undermine the reliability of within-firm inference.

For the firms that pass the eligibility screen, missing values in cshr are handled using a deliberately conservative imputation strategy that fills only short internal gaps and explicitly avoids extrapolation across long interruptions.

3.4.2. Handling Missing Data – Imputation

The procedure distinguishes between internal gaps and boundary gaps within each firm's time series:

i. Short internal gaps

- Missing values occurring in runs of one or two consecutive fiscal years that are bracketed by observed values on both sides are filled by linear interpolation in fiscal-year time.

- Longer internal gaps are not interpolated (since were eliminated), consistent with the screening rule described in the previous section.

ii. Boundary gaps (start and end of the series)

- Missing values at the beginning or end of a firm's series are filled using the nearest observed level:

  ⇒ last-observation-carried-forward (LOCF) for terminal gaps

  ⇒ backward fill for initial gaps.

This step ensures that the series is defined for every firm-year once short internal gaps have been bridged, without imposing time trends beyond the observed support.

The resulting series is stored as cshr_filled. An indicator variable, cshr_imputed, flags all firm-year observations for which the original cshr was missing and subsequently recovered through interpolation or padding. A hard consistency check confirms that cshr_filled contains no missing values for the retained firms.

In the final retained sample, which comprises 12,372 firms, only 5.8% of firm-year observations are flagged as imputed.

3.4.3. Scale conversion and units

Compustat reports cshr in thousands. For interpretability and to avoid scale artifacts in subsequent non-linear transformations, the filled series is converted to absolute counts using a deterministic mapping:

$$cshr\_abs \ = \ round(1{,}000 \ x \ cshr\_filled)$$

This transformation:

- preserves all observed levels exactly up to the database's reporting precision
- places the variable on its natural counting scale.

A small number of firm-years take the value zero after conversion (99 observations, 0.079% of retained rows). These cases are rare and typically reflect genuine database zeros or rounding at very low holder counts. They are retained at this stage and handled explicitly in the next section when defining logarithmic transformations.

3.4.4. Diagnostics and implications for estimation

Two diagnostics summarize the outcome of this treatment:

i. Firm-level diagnostics: the eligibility screen yields a balanced mix of short and long panels, each with sufficient within-firm variation to support fixed-effects estimation.

ii. Observation-level diagnostics

   - The overall imputation intensity is low (5.8%)
   - Imputation is strictly confined to gaps of at most two years

- All imputed observations are fully traceable via cshr_imputed

Together, these features minimize measurement error arising from aggressive smoothing while ensuring that the breadth variable is observed for every firm-year included in the subsequent analyses (12,372 firms).

This section defines only the cleaned levels used as inputs to later constructions. Transformations of breadth (for example, logarithmic maps that are robust to zeros and left-tail trimming choices) are introduced in Section 3.5, where we discuss the rationale for log1p, the treatment of one-holder firm-years and the consistency of these choices across the performance and valuation specifications.

3.5. Construction of breadth measures

3.5.1. Starting point and diagnosis

The starting variable is the shareholder count in units is cshr_abs.
In levels, cshr_abs is not suitable for parametric or shape-sensitive analysis for three reasons:

i. Extreme right tail: a handful of very large observations dominate linear fits and compress the bulk of the distribution against the origin, as the histogram shows (Figure 1)

ii. Left-boundary artifacts: there are true zeros and a discrete spike at one holder (a reporting granularity point). These boundary values distort flexible forms and local smoothers

iii. Scale-dependent dispersion: variance rises sharply with the level of cshr_abs (strong heteroskedasticity), so regressions in levels would be dominated by high-count firm-years and would typically produce fan-shaped residuals.

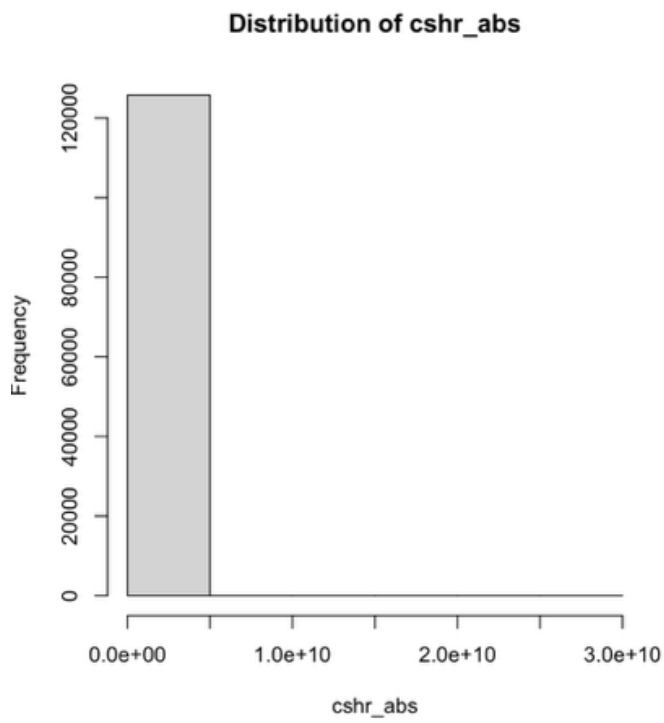

Figure 1 – Distribution of Shareholders in absolute count

3.5.2. Log mapping and rationale

A monotone, variance-stabilizing map is required to compress the right tail and remain defined at zero without arbitrary constants. The transformation adopted (see Figure 2 for distribution) is $log1p(x) = log(1 + x)$, which:

i. is well defined at $x = 0$
ii. preserves the rank order of observations
iii. approximates the natural log for moderate and large x
iv. supports elasticities and "percent-style" interpretations locally

Accordingly, the working breadth variable is created as:

$$breadth = log1p(cshr\_abs)$$

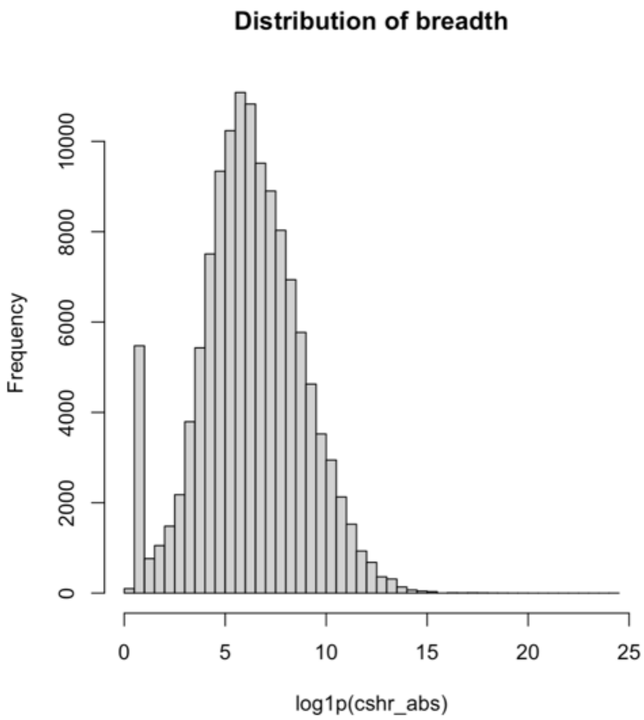

Figure 2 – Distribution of breadth (log1p of Shareholders absolute count)

3.5.3. Data-driven left-tail cut (improving local symmetry)

Before fixing the final breadth measures, a minimal left-tail threshold on cshr_abs is selected to mitigate boundary pathologies and obtain a "locally symmetric" distribution after log1p. The procedure was:

- Candidate thresholds: a small grid is considered for the raw count, e.g., {1, 10, 25, 50, 100, 250, 500}.
- For each candidate cut ct, the sub-sample $x = \{cshr\_abs \geq ct\}$ is formed and mapped to $lx = log1p(x)$.
- On lx, three diagnostics are computed: skewness, kurtosis and when $3 \leq N \leq 5{,}000$, the Shapiro–Wilk p-value (the Shapiro test is intentionally limited to moderate N).
- The selected threshold, best_cut, minimizes |skewness(lx)|. Skewness is used as the primary criterion because it directly targets left-right balance after transformation, which is what matters most for polynomial/spline approximations. Kurtosis and Shapiro p-values are reported as secondary diagnostics.
- The dataset is then filtered to $cshr\_abs \geq best\_cut$ and breadth is recomputed as $log1p(cshr\_abs)$ on the filtered sample to ensure perfect alignment between the cut and the transformation (see Figure 3).

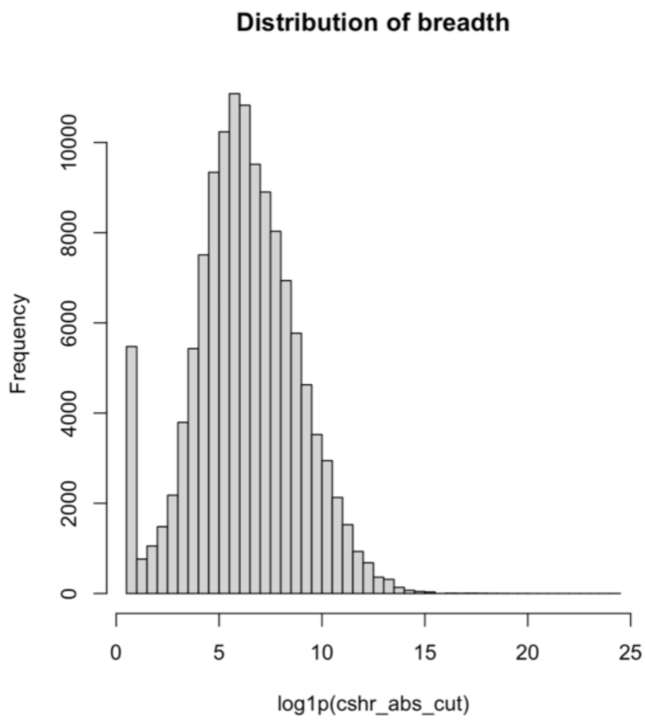

Figure 3 – Distribution of breadth (log1p of Shareholders absolute count after minimizing skewness)

3.5.4. Final mapping and operative choice of the breadth regressor

The log transformation substantially regularizes the shareholder-count series, yet a residual problem remains at the extreme left tail. Even after converting counts to $breadth\_all = log1p(cshr\_abs)$ and after applying the minimal left-tail filter discussed earlier), the distribution retains a discrete spike at $log1p(1)$. That spike is mechanical: it originates from firm-years with exactly one recorded shareholder ($cshr\_abs = 1$). The single-holder mass creates a boundary pile-up that is not informative about the interior shape of the breadth–outcome relation and can distort flexible parametric fits by pulling curvature toward the boundary.

This issue is visible in the side-by-side histogram (see Figure 4): the panel on the left (breadth_all) exhibits a prominent bar at the lower bound; the panel on the right removes only those one-holder firm-years before taking logs (breadth_no1), yielding a smooth, unimodal distribution with markedly improved behavior in the first few quantiles.

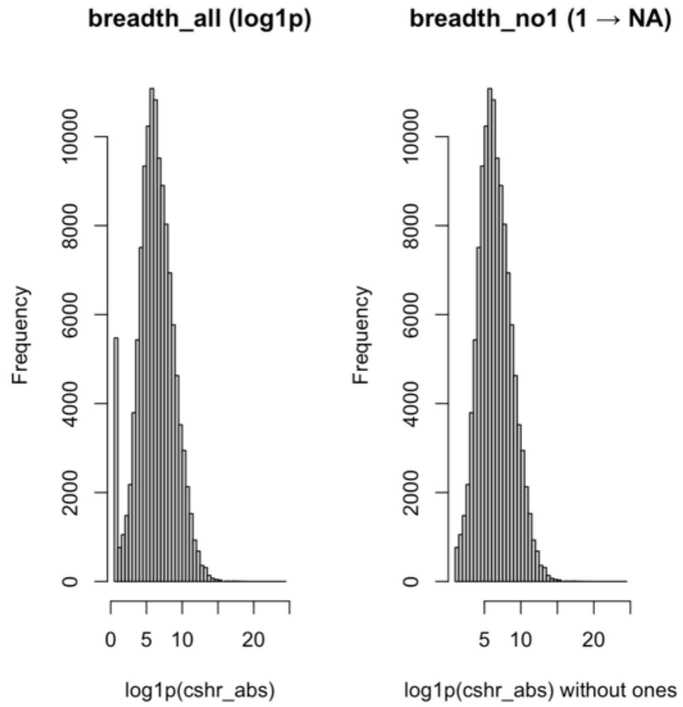

Figure 4 – Distribution of breadth (comparison between log1p of Shareholders absolute count raw and after removing *cshr_abs* = 1)

The operative choice therefore balances parsimony and transparency: no aggressive trimming is introduced beyond excluding the knife-edge case $cshr\_abs = 1$; all other observations are retained, the rank information is preserved and the left-tail artefact that would otherwise bias curvature near the boundary is removed. This is the breadth variable referenced in subsequent econometric sections.

3.5.5. Implications for estimation and figure placement

These construction choices are designed to align diagnostics and estimation:

- The breadth_no1 variable delivers a stable regressor with compressed tails and monotone interpretation. It is preferred in non-linear fixed-effects specifications to avoid curvature driven by an atom at one; breadth_all remains available for descriptives and robustness
- The left-tail cut prevents boundary artifacts from influencing polynomial/spline shapes
- All comparative figures (e.g., linear vs quadratic vs spline fits) must be produced on identical row sets to ensure that differences come from functional form, not sample composition.

3.6. Construction of supporting variables

3.6.1. Total assets

Total assets proxy the firm's economic scale, the cumulative stock of resources that underpins capacity, fixed-cost absorption, disclosure intensity, analyst coverage and index eligibility. Larger balance sheets tend to coincide with broader investor participation (more inclusion in mandates, heightened visibility) and with operating outcomes through classical scale economies and lower cash-flow volatility.

In a two-way fixed-effects setting, variation in assets within a firm over time captures contemporaneous expansions/contractions in scope, acquisitions or major investment cycles that could jointly influence breadth and performance. Including assets as a control therefore prevents breadth from acting as a stand-in for raw scale and helps isolate any participation effect from simple growth in the firm's footprint.

Formulas:

- $total\_assets = at$
- $log\_total\_assets = log(at)\ for\ at > 0;\ otherwise\ NA$

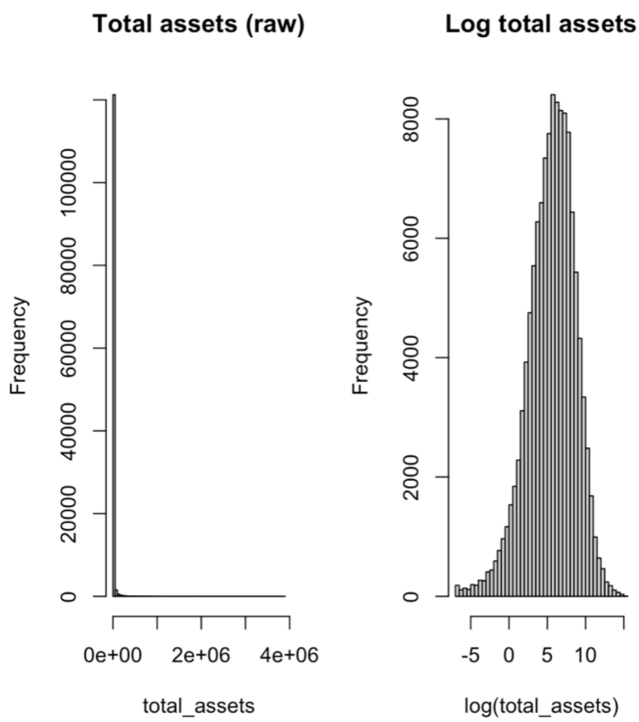

Figure 5 – Distribution of total assets (comparison between the raw and logarithmic distributions)

The log transformation (see Figure 5) turns total assets from an unusable regressor in levels into a well-behaved size control for fixed-effects estimation. In levels, a handful of mega-firms sit far out in the right tail and dominate linear fits; residuals display a clear fan shape and standard errors become overly sensitive to leverage points. Mapping to log(total_assets) yields a compact, roughly bell-shaped distribution (as in the figure), attenuates the influence of extremes and stabilizes variance across the range. This improves numerical conditioning in feols with firm and year effects, delivers coefficients with a semi-elasticity interpretation (a change Δ in the log corresponds to a $\times\, e^{\wedge}\Delta$ change in assets) and ensures that fitted relationships are not driven by outliers. The log scale also matches the economic idea that size matters in proportional rather than absolute terms, which is why it is used as the default firm-year "size" control in the subsequent models.

3.6.2. Shareholders' equity

Equity summarizes cumulative profitability and external financing choices and represents the primary loss-absorption buffer. Shifts in equity often accompany events that also affect investor recognition, public offerings, stock-financed acquisitions, option exercises, so equity is a natural state variable to hold constant when relating breadth to outcomes. Larger equity typically lowers distress risk and may support higher valuation multiples and steadier operations; however, equity growth can reflect either retained earnings (supportive for ROA) or dilution (ambiguous in the short run), so conditioning on equity helps separate participation changes from capital-structure-driven movements.

Formulas:

- $total\_equity = ceq$

- $log\_total\_equity = log(ceq)\ for\ ceq > 0;\ otherwise\ NA$

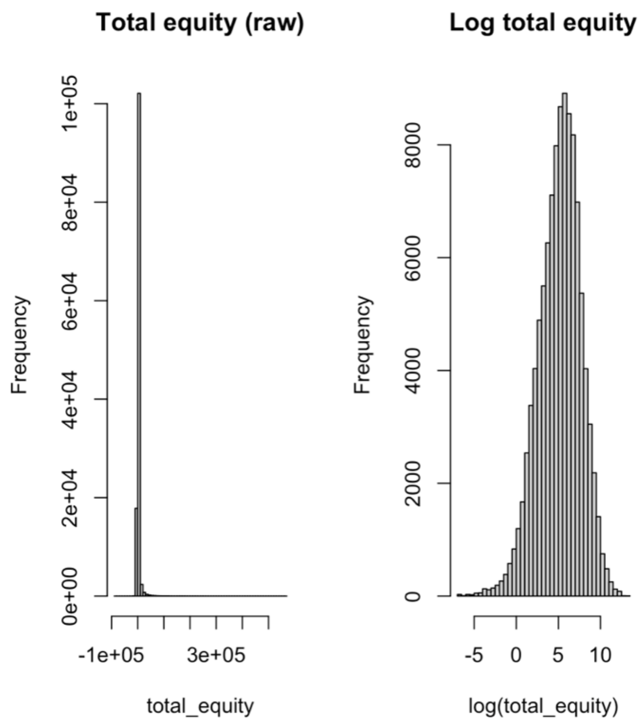

Figure 6 – Distribution of total equity (comparison between the raw and logarithmic distributions)

The log transformation (see Figure 6) likewise turns total equity into a workable covariate for fixed-effects models. In levels, equity exhibits a spike near zero, a very long right tail and frequent non-positive values (from accumulated losses or write-downs); these features make linear fits unstable and heavily influenced by a small number of observations. Mapping equity to log(total_equity) on the subset with strictly positive equity compresses the right tail, reduces leverage of outliers and yields a compact, near-symmetric distribution (as visible in the figure 6), which stabilizes residual variance and improves numerical conditioning in ordinary least squares regressions with firm and year fixed effects (feols). The log scale also provides a clean semi-elastic interpretation, coefficients read as the change in the outcome associated with a proportional change in the equity base and aligns with the economic view that capital strength matters in percentage terms rather than in absolute dollars. Observations with zero or negative equity are left as missing in the logged variable by construction, avoiding arbitrary offsets and ensuring that any specification using log_total_equity relies on well-defined scale information.

3.6.3. Cash and short-term investments

Cash is financial slack and timing flexibility: it buffers shocks, funds projects without issuance costs and can signal precautionary behavior or transitory proceeds from financing/asset sales. Because cash spikes are often news-driven and can attract investor attention independently of operations, controlling for cash helps prevent confounding between recognition shocks and liquidity positions. Economically, higher cash balances can ease financing constraints (potentially supporting investment and valuation), but may also reflect underemployment, leaving expected signs context-dependent; the role here is to partial out first-order liquidity differences that co-move with breadth.

Formulas:

- $total\_cash = ch$
- $log\_total\_cash = log(ch)\ for\ ch > 0;\ otherwise\ NA$

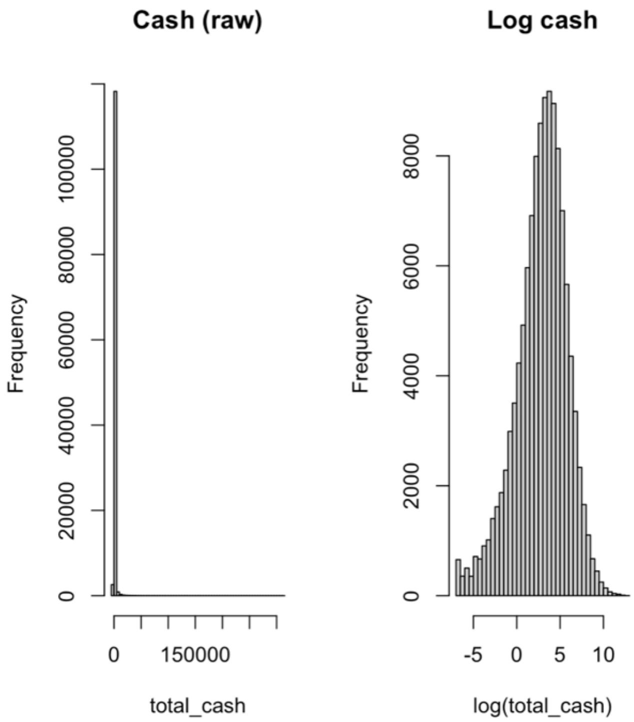

Figure 7 – Distribution of total cash (comparison between the raw and logarithmic distributions)

Cash and short-term investments proxy a firm's immediate liquidity buffer, resources available to bridge cash-flow shortfalls, fund working-capital needs or react to investment opportunities without external financing. As a control, cash captures precautionary motives and financing frictions that can co-move with both profitability and valuation (e.g., firms under tighter constraints tend to hold more cash; highly profitable firms may also accumulate cash endogenously). In raw levels the series is unusable for linear panel models: most observations lie close to zero while a thin right tail stretches to very large values, so any regression in levels would be dominated by a handful of high-cash firm-years and produce fan-shaped residuals. The natural logarithm addresses these issues (see Figure 7). Mapping the variable to log(total_cash) after masking non-positive values (so only strictly positive cash enters the log) compresses the right tail, yields a nearly symmetric, bell-shaped distribution and places the variable on a scale where coefficients admit local elasticity interpretations (a given percentage change in cash has a roughly constant marginal association). The contrast in the figure is stark: the left panel (raw cash) piles mass at the lower boundary with an extreme tail, whereas the right panel (log cash) shows a compact, well-behaved distribution, precisely the form required for stable fixed-effects estimation and shape diagnostics. Because some firm-years legitimately report zero cash, those rows are left as NA in the logged version; this avoids arbitrary offsets, preserves the accounting meaning of zeros and ensures that models including log cash rely only on informative, strictly positive observations.

3.6.4. Total liabilities

Liabilities capture financing intensity and the burden of fixed charges, linking directly to expected returns through financial risk and refinancing exposure. Changes in leverage can trigger heightened disclosure and creditor monitoring, potentially broadening the investor base for reasons orthogonal to operating fundamentals. Conditioning on liabilities therefore guards against attributing leverage-related variation in performance or valuation to breadth. Within firms, increases in liabilities usually compress ROA (higher interest burden) and can lower valuation (higher discount rates), making this a core control in both profitability and valuation specifications.

Formulas:

- $total\_liabilities = lt$

- $log\_total\_liabilities = log(lt) \ for \ lt > 0; \ otherwise \ NA$

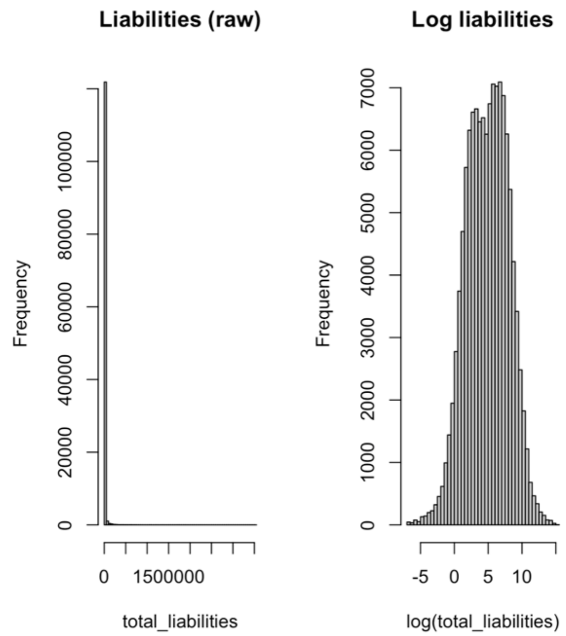

Figure 8 – Distribution of total liabilities (comparison between the raw and logarithmic distributions)

The side-by-side histograms (see Figure 8) motivate using the natural log of liabilities as the control. In levels (left panel), total liabilities are crammed against zero with a very long right tail, so a handful of very large issuers dominate the scale and any level-specification would suffer from severe heteroskedasticity and leverage by outliers. After applying log to strictly positive liabilities (right panel), the distribution becomes compact, roughly unimodal and far closer to symmetric, which stabilizes variance and preserves rank information while converting multiplicative scale differences into additive ones. Observations with lt ≤ 0 are rare in this industrial sample and are left missing rather than offset, maintaining accounting meaning and avoiding arbitrary constants. This logged measure makes the variable usable in two-way fixed-effects regressions (and in polynomial/spline terms) and coefficients admit the usual semi-elasticity reading: a given percentage change in liabilities maps to a roughly constant marginal association with the outcome.

3.6.5. Net income

Net income provides the earnings flow used to construct ROA and serves as a diagnostic anchor for the profitability environment of the panel. Because it can be negative and volatile, it is not employed as a baseline right-hand-side control in the ROA regressions to avoid “bad control” issues; rather, it structures the outcome and supports distributional checks. Even so, documenting the transformation clarifies how extremes are handled for visuals and supplementary summaries.

Formulas:

- $total\_net\_income = ni$
- $log\_total\_net\_income = log(ni)\ for\ ni > 0;\ otherwise\ NA$

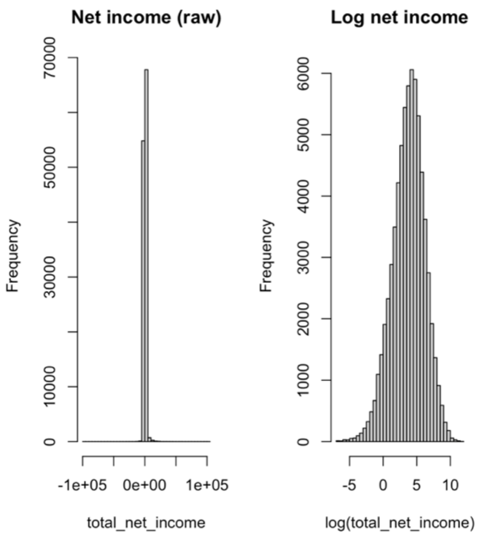

Figure 9 – Distribution of total net income (comparison between the raw and logarithmic distributions)

The side-by-side histograms (see Figure 9) make clear why a log mapping is useful for net income. In levels (left panel), total_net_income is tightly compressed around zero with long, thin tails on both sides; a small number of very large profit or loss realizations dominate the scale and the mass near zero would induce severe heteroskedasticity and leverage problems in level-specifications. After applying the natural log to strictly positive observations (right panel), the distribution becomes compact, unimodal and much closer to symmetric, which stabilizes variance and makes the variable compatible with two-way fixed-effects models and polynomial/spline terms. Importantly, the log is computed only when net income > 0, so loss and zero observations are left as missing rather than shifted by arbitrary constants; this preserves accounting meaning and ensures that the logged series reflects the intensive margin of profitability among profitable firm-years.

3.6.6. Total revenue

Revenue is a proximate measure of operating intensity and demand and sits close to channels through which recognition may operate: investor-facing initiatives, distribution expansions or product launches can raise both sales and visibility; conversely, strong sales performance can attract new holders. Controlling for revenue in robustness checks helps separate these potential mediating pathways from a direct association between breadth and the outcomes of interest. Within firms, higher revenue is expected to correlate positively with profitability (scale economies, overhead absorption) and with valuation (growth optionality).

Formulas:

- $total\_revenue = revt$
- $log\_total\_revenue = log(revt)\ for\ revt > 0;\ otherwise\ NA$

The side-by-side histograms (see Figure 10) make clear why a log mapping is useful for revenue. In levels (left panel), revenue is tightly compressed around zero with long, thin tails on both sides; a small number of very large profit or loss realizations dominate the scale and the mass near zero would induce severe heteroskedasticity and leverage problems in level-specifications. After applying the natural log to strictly positive observations (right panel), the distribution becomes compact, unimodal and much closer to symmetric, which stabilizes variance and makes the variable compatible with two-way fixed-effects models and polynomial/spline terms. Importantly, the log is computed only when revenue > 0, so loss and zero observations are left as missing rather than shifted by arbitrary constants; this preserves accounting meaning and ensures that the logged series reflects the intensive margin of profitability among profitable firm-years.

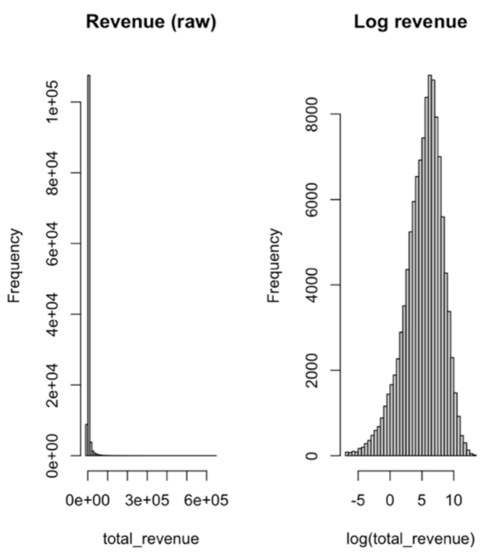

Figure 10 – Distribution of total revenue (comparison between the raw and logarithmic distributions)

3.7. Construction of Outcome Variables: ROA and Tobin's Q

This chapter details how the two dependent variables used throughout the empirical analysis are constructed from Compustat fundamentals, why each design choice is made and what those choices imply for fixed-effects estimation. Return on assets (ROA) serves as an accounting-based proxy for operating performance, anchored in audited statements and largely insulated from contemporaneous market sentiment. Tobin's Q serves as a market-based proxy for valuation, embedding investors' expectations about future cash flows and discount rates. Across both constructions, inputs must be finite and denominators strictly positive, tail behavior is managed through transparent and pre-announced rules that preserve economic ordering while stabilizing variance and raw counterparts are retained to document how transformations reshape the distribution.

3.7.1. Return on assets (ROA)

ROA measures period profitability relative to the asset base that generates it, providing a firm-year signal that aggregates the outcome of numerous operating and financial decisions. In a within-firm design with year effects, ROA moves when profitability changes for reasons internal to the firm or because of transitory events captured by accounting and it does so on a scale that is comparable across firms. This comparability is attractive when relating performance to changes in investor recognition because it avoids reliance on market prices and makes the outcome less sensitive to short-run sentiment, while still reflecting the consequences of financing conditions and scrutiny that may be associated with breadth. Among available definitions, the construction here uses net income over end-of-year total assets, rather than averages of beginning and ending assets or operating-income variants, because the end-of-year denominator aligns exactly with the fiscal-year timing of breadth and other Compustat items in the panel and because net income captures bottom-line profitability that can transmit breadth-related channels operating below the operating line. Negative net income is permitted by construction, the restriction concerns only the denominator. Because accounting ratios can exhibit extreme values when the denominator is very small or when one-off items hit the numerator, the raw ROA distribution shows heavy tails. To limit the leverage of such cases while preserving cross-firm comparability, ROA is winsorized at the 1st and 99th percentiles on the pooled firm-year distribution. Pooled rather than year-wise trimming is preferred because ROA's dispersion is driven more by firm-specific accounting events than by market regimes, a common trimming rule keeps the scale comparable across years and avoids mechanically narrowing or widening tails in high-volatility periods and winsorization is monotone within tails so interior ranks are preserved for inference (see Figure 11).

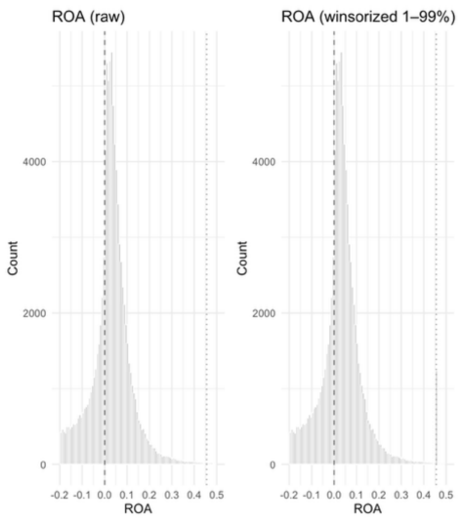

Figure 11 – Distribution of ROA (comparison between the raw and winsonorized distributions)

Formulas:

- $ROA = ni / at, computed\ only\ if\ at > 0;\ otherwise\ NA.$
- ROA_w= ROA winsorized at the 1st and 99th percentiles on the pooled distribution across all firm-years.

Adopting the pooled 1–99% winsorized ROA (roa_win) trims only 2.00% of observations (cuts = −19.8836 and 0.4551) yet removes the pathologies that would otherwise dominate least-squares: the standard deviation collapses from 386.96 to 2.42 (−99.4%), absolute skewness falls from 310.34 to 6.60 (−97.9%), kurtosis from 103,704 to 49.30 (−99.95%) and ∣mean − median∣ shrinks by 85.4% (3.684 → 0.539). These gains sharply reduce residual variance and prevent a handful of ratios driven by tiny denominators or transitory items from dictating the fit. Because the transformation is monotone on the interior, the cross-sectional ordering is essentially unchanged (Spearman ≈ 1.00), so the within-firm slope identified by fixed effects is preserved apart from the extreme tails. In practice, this yields numerically stable feols estimates with firm and year fixed effects and cluster-robust inference, while raw ROA remains available for non-parametric checks. Finally, ni, at and breadth are aligned by fiscal year, ensuring that panel variation is synchronous across outcomes and regressors and that fixed-effects estimates reflect contemporaneous within-firm movements rather than timing artifacts.

Under the cleaned panel used throughout (industrial template, consolidated, standard format, domestic source, latest filing per firm-year), valid ROA is observed for the large majority of rows and all fixed-effects specifications in Chapter 5 use ROA_w as the dependent variable while the raw ratio is retained for descriptive checks.

3.7.2. Tobin's Q

Tobin's Q is the market-based counterpart to the accounting measure used for performance. It links the market's assessment of the firm's equity claim to the scale of the asset base reported in Compustat and is therefore the natural valuation endpoint for shareholder-base breadth. Within the incomplete-information and investor-recognition perspective, broader investor participation should ease financing frictions, improve risk sharing and strengthen information aggregation. Conditional on fundamentals and firm and time effects, years in which a given firm is recognized by a wider investor base are expected to exhibit higher Q. Implementing this idea credibly requires a construction that (i) reconciles the two equity-value sources available in Compustat without introducing artificial noise, (ii) guards the ratio against ill-posed denominators, (iii) neutralizes the very long, regime-dependent right tail of Q that otherwise dominates least-squares and (iv) produces a scale that is numerically well behaved in fixed-effects regressions while preserving the central cross-sectional ordering that identifies within-firm slopes.

Compustat provides two ways to measure the market value of common equity at fiscal year-end: a reported figure (mkvalt) and a reconstructible one obtained as common shares outstanding times the fiscal-year-end price (csho × prcc_f). The two should coincide up to rounding and share-class conventions; in practice they are nearly collinear on the overlap. To maximize coverage while avoiding ad-hoc averaging of heterogeneous definitions, a deterministic tolerance rule chooses a single numerator per firm-year: when both sources exist and the absolute relative gap is within one percent, the reported mkvalt is retained, when the two disagree by more than one percent, the reconstructed csho × prcc_f is used, when only one source is present and positive, that source is kept, otherwise the observation is set to missing. The denominator is the book value of total assets (at), restricted to strictly positive values to avoid division by zero or sign inversions that have no valuation meaning. This yields the standard "average Q" proxy that is widely used in the literature; debt-augmented variants would impose extra assumptions on the market value of debt while materially shrinking sample size with little benefit for the recognition questions addressed here.

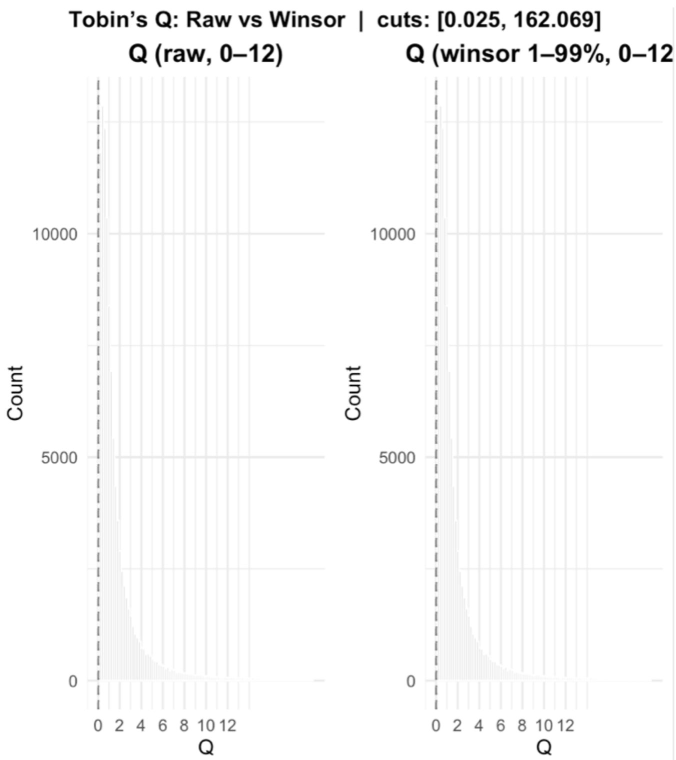

Figure 12 – Distribution of Tobin's Q (comparison between the raw and winsonorized distributions)

Even after careful numerator selection, raw Q is extremely right-skewed (see Figure 12) and its tail length varies across market regimes. Rather than impose a single, pooled cap that would let boom or crisis years dictate the effective trimming for all, the series is winsorized within fiscal year at the 1st and 99th percentiles, which equalizes tail influence year by year and preserves comparability over time. The pooled diagnostics confirm that this step targets only the extremes: exactly 2.00% of observations are clipped at pooled cuts near 0.025 and 162.069, while the median (≈ 0.9932) and the interquartile range (≈ 0.4815–2.1388) are unchanged. Because valuation is multiplicative and strictly non-negative, a log map is appropriate for regressions. We therefore take a log1p on the within-year winsorized series, log(1+Q_w). This compresses the right tail, improves numerical stability and provides coefficients that can be read as semi-elasticities in 1+Q. Moment comparisons illustrate the payoff of this pipeline: relative to raw Q, the standard deviation falls by about 99.3%, absolute skewness by about 95.7% and kurtosis by about 99.8%, while the interior rank information is essentially preserved (Spearman correlation between raw and winsorized Q ≈ 1.00, Pearson drops by design because the extremes are capped). The resulting dependent variable is single-peaked and well behaved, which is desirable for fixed-effects estimation with firm-clustered standard errors.

Formulas:

- $mktcap\_rep = mkvalt\ (finite\ and\ > 0)$
- $mktcap\_rec = csho \times prcc\_f\ (finite\ and\ > 0)$
- mktcap_final selection rule (tolerance 1%): if both available and $|(mktcap\_rec - mktcap\_rep)/mktcap\_rep| \leq 0.01,$ then $mktcap\_final = mktcap\_rep$; if both available and the absolute relative difference > 0.01, then

$mktcap\_final = mktcap\_rec$; if only one source is available and positive, use that source; otherwise NA

- $Q\_raw = mktcap\_final / at \ for \ at > 0; \ otherwise \ NA$
- $Q\_w \ (within - year) =$
  $Q\_raw \ winsorized \ at \ the \ 1st \ and \ 99th \ percentiles \ within \ each \ fiscal \ year$
- $q\_log\_w = log(1 + Q\_w)$

The figures corroborate the statistics. In the side-by-side histograms (zoomed to 0–12), both raw Q and Q_w place the overwhelming mass close to one and show a thin, rapidly decaying tail beyond two; the visual similarity on this central window reflects that only 2.00% of observations are clipped and that most of the removed mass lives far outside the plotted range.

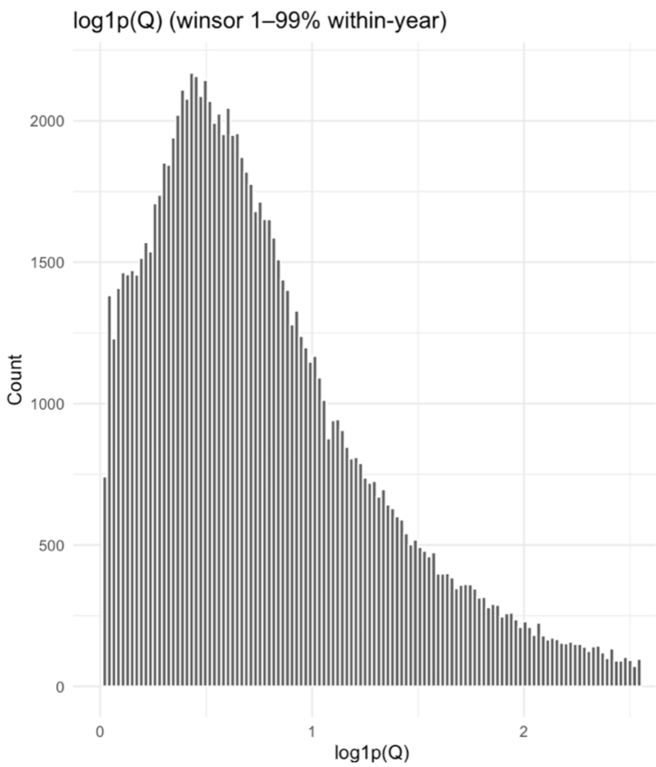

Figure 13 – Distribution of log1p of Tobin's Q winsorized

The log histogram of log(1+Q_w) is unimodal with mild right skew (see Figure 13), confirming that the within-year winsorization plus log mapping converts a regime-sensitive, heavy-tailed price ratio into a stable outcome for panel regressions. Taken together with the ROA construction in 3.7.1, this ensures that performance and valuation measures share consistent timing, filters and transparent tail rules calibrated to the statistical nature of each variable: ROA remains in levels because it can be negative and already clusters tightly around small values, whereas Q is log-mapped because it is non-negative, multiplicative and heavy-tailed. Hence, any differences between the ROA- and Q-based estimates in the next sections can be attributed to economics rather than artifacts of variable construction.

4 Empirical Strategy and Identification

4.1. Aims and organizing principles

This chapter clarifies how the empirical analysis proceeds and why each step is structured the way it is. The overarching aim is to test the link between breadth and outcomes on two fronts, operating performance and market valuation, using a single, symmetric design so that any difference in results reflects economics rather than construction or modelling artifacts. The analysis is anchored in within firm variation over time to difference out time invariant firm traits, it aligns measurement windows across regressors and outcomes to keep timing coherent and it manages distributional tails in a transparent, pre specified way to prevent a handful of extreme observations from dictating curvature or significance.

4.2. Evidence path and estimation protocol

The empirical path is intentionally sequential and parallel for both outcome blocks. We begin with non-parametric evidence on common samples, using dense scatters with local smoothers and balanced binned comparisons to display the sign and broad shape of the relationship between breadth and outcomes without committing to a functional form. These visuals serve two purposes, they provide an unconditional sense of direction and curvature and they inform whether flexible parametric terms, for example a quadratic term or a spline with low degrees of freedom, are warranted in the next step. We then move to compact fixed effects regressions with firm and year effects, which summarize the within firm slope under tight controls for unobserved heterogeneity and aggregate shocks. The same sample definition used for the visuals is carried into the regressions to ensure that the estimated slope is directly comparable to the non-parametric shape. Curvature is introduced parsimoniously, not to overfit but to test the economically motivated hypothesis that marginal effects taper at higher breadth. Finally, we layer standard balance sheet controls, size, leverage, liquidity, constructed and transformed in Chapter 3 with identical rules across outcomes, so that any attenuation or stability in the breadth coefficient has the same interpretation in the performance and valuation blocks. At each step, we preserve rank information, we keep tails handled by the same winsorization logic described earlier and we report results on the same set of rows for the specification at hand avoiding silent changes in composition.

4.3. Identification, interpretation and robustness

Identification rests on within firm changes, firm fixed effects remove time invariant attributes such as business model, disclosure style, baseline visibility, year fixed effects purge macro movements and the remaining variation comes from the firm specific evolution of breadth. To mitigate mechanical feedback, we replicate key estimates using lagged breadth, which reduces simultaneity with contemporaneous outcomes while preserving within firm leverage. We keep the timing of measurement synchronous across all variables, we apply the same tail rules to both outcomes and we document the influence of potential edge cases by re estimating after excluding observations flagged by construction, for example filled shareholder counts or extreme tail outcomes prior to winsorization. Inference relies on heteroskedasticity and autocorrelation robust standard errors clustered at the firm level, recognizing that residuals are likely serially correlated within firms. Coefficients are interpreted in a way that is appropriate for each outcome yet comparable across blocks, for ROA, slopes read as changes in profitability points associated with within firm movements in breadth, for valuation, we work on a stabilized scale, winsorized Q and where used a log mapped Q and we translate effects into economically meaningful shifts over interquartile changes in breadth. Robustness follows a simple extreme bound logic, we vary the functional form for breadth, linear versus a single curvature term versus a spline with few degrees of freedom, we perturb tail treatments within a narrow, pre announced band we verify that results are not driven by a few boom or crisis years and we replicate on size or liquidity stratified subsamples where frictions plausibly differ. All robustness checks are executed symmetrically across the performance and valuation blocks and are reported with identical formatting to make the comparison immediate.

4.4. What follows

The subsequent chapters implement this roadmap in lockstep. For operating performance and for market valuation alike, we proceed from shape to slope, first visuals on common samples, then fixed effects baselines, then curvature and controls and finally targeted robustness. Estimates are always read against the same construction rules and timing conventions established in Chapter 3, so that the narrative can focus on economics, whether and how investor breadth tracks improvements in fundamentals, in prices or in both, rather than on specification drift.

5 Operating performance: shareholder-base breadth and ROA

5.1. Non-Parametric evidence

5.1.1. Visual diagnostics: ROA versus shareholder breadth

This subsection reports unconditional evidence on the association between operating performance and investor breadth using the raw firm-year panel, operating performance is measured as $ROA = ni / at$, computed whenever total assets are strictly positive and allowing negative values to retain loss-making observations, breadth is inspected both in levels through cshr_abs, the number of registered shareholders and in compressed form through $breadth\_no1 = log(1 + cshr\_abs)$ restricted to firm-years with at least two shareholders. Because the logarithm is strictly increasing over the positive reals, the mapping to breadth_no1 preserves the rank ordering of firms by breadth on their common support, which is useful for the rank-based inference reported in the next subsection.

Figure 14 plots return on assets on the vertical axis against the number of recorded shareholders on the horizontal axis. Return on assets is computed as total net income divided by total assets, with observations retained whenever total assets are strictly positive, therefore negative realizations are possible when net income is negative. The horizontal support is extremely concentrated at small shareholder counts and quickly thins as the base grows, the vertical mass is tightly packed just above zero with a long but sparse negative tail. Conditional dispersion is large at low cshr_abs and contracts only mildly as the shareholder base expands and no stable linear or curved pattern is visually dominant once the strong imbalance of the horizontal distribution is considered. The figure therefore suggests that, if breadth relates to profitability, the association is subtle on raw scales and easily overshadowed by scale effects and a few extreme ratios.

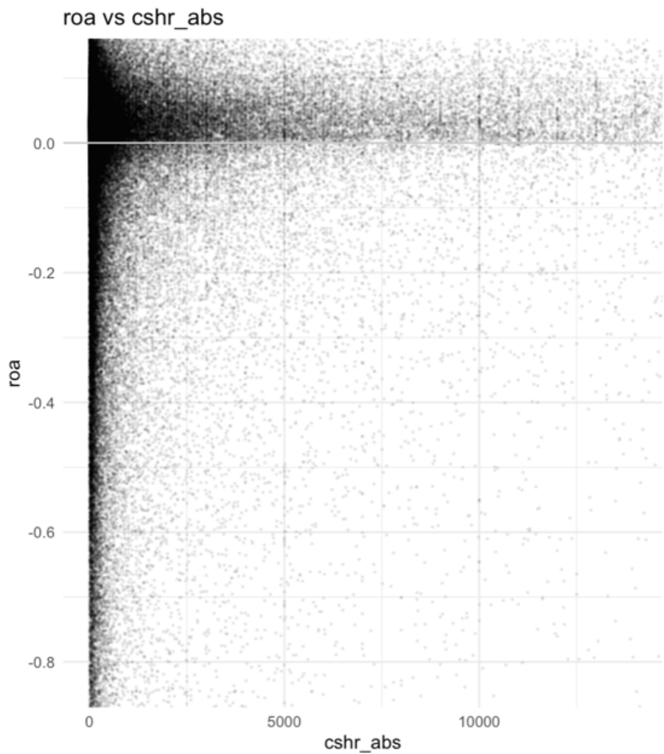

Figure 14 – ROA vs Shareholders absolute count

Figure 15 repeats the exercise using asset turnover on the vertical axis and the number of shareholders on the horizontal axis. Asset turnover is defined as total revenue divided by total assets, hence it is non-negative by construction and less sensitive than ROA to transitory income shocks. This robustness view is introduced to check whether a different operating proxy reveals a clearer relation with investor base size. The cloud again bunches near the origin for mechanical reasons, then fans out slowly with larger shareholder bases, indicating slightly greater heterogeneity in operating efficiency for firms with broader recognition. As with ROA in Figure 14, an evident functional form does not emerge in levels, which motivates the use of transformations that equalize leverage across the support before drawing inference.

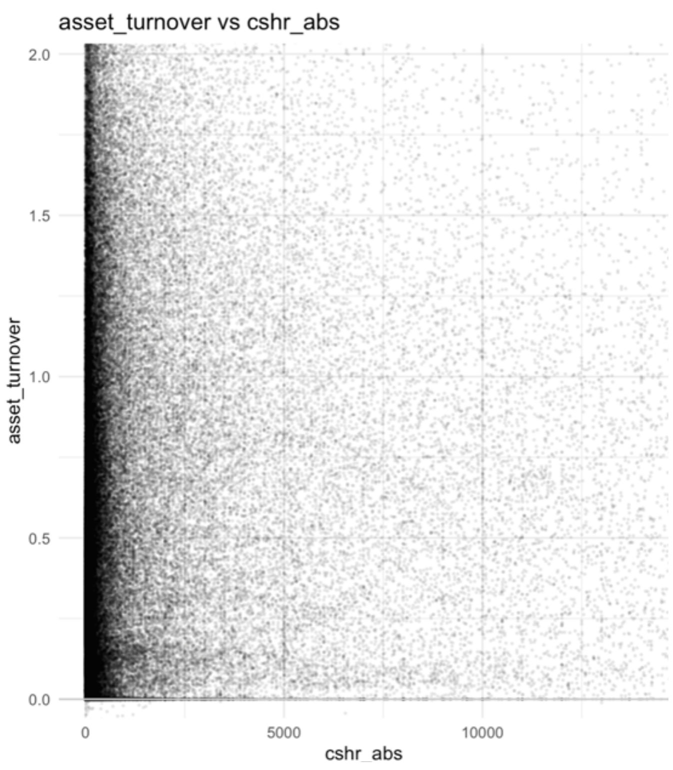

Figure 15 – Asset turnover vs Shareholders absolute count

Figure 16 implements the transformations motivated in Section 3.7. The horizontal axis reports breadth_no1, defined as the natural logarithm of one plus the number of shareholders with single-shareholder cases excluded, which spreads the heavy mass at low counts and compresses the long right tail. The vertical axis reports ROA_w, the pooled winsorized version of ROA that truncates the distribution at the first and ninety-ninth percentiles, which affects about two percent of observations and leaves the ranking of the interior mass unchanged while materially reducing skewness and kurtosis. Under these adjustments the interior becomes readable, the bunching at very small shareholder bases largely disappears and a mild upward association of profitability with breadth is visually more plausible over the central region where the data are densest. The purpose here is not to claim shape, but to construct a scale on which the interior drives the picture rather than a handful of extremes.

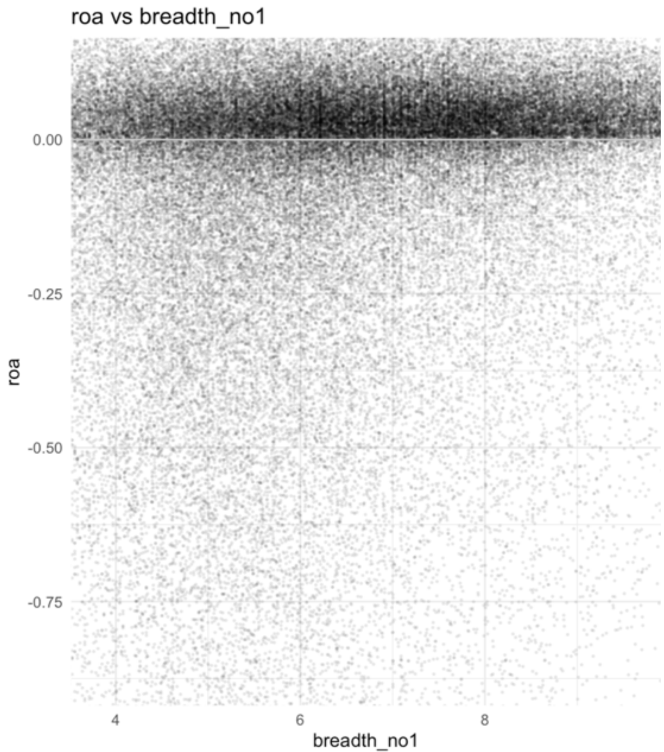

Figure 16 – ROA winsorized vs breadth_no1

Figure 17 aggregates the relation by dividing breadth_no1 into balanced bins and plotting the mean of ROA_w for each bin, with bin sizes printed next to the markers. The profile is monotone over most of the support, average profitability is more negative in the bins with the narrowest shareholder base, becomes distinctly less negative in the central bins that contain the bulk of the sample and approaches values close to zero in the bins with the broadest base. Because only the tails of ROA are trimmed, this improvement reflects movements in the interior mass rather than an artifact of winsorization. The binned curve thus provides a compact and interpretable summary that motivates the rank-based tests and the within-firm models developed in Sections 5.1.2 and 5.2, where the same patterns are quantified under transparent assumptions.

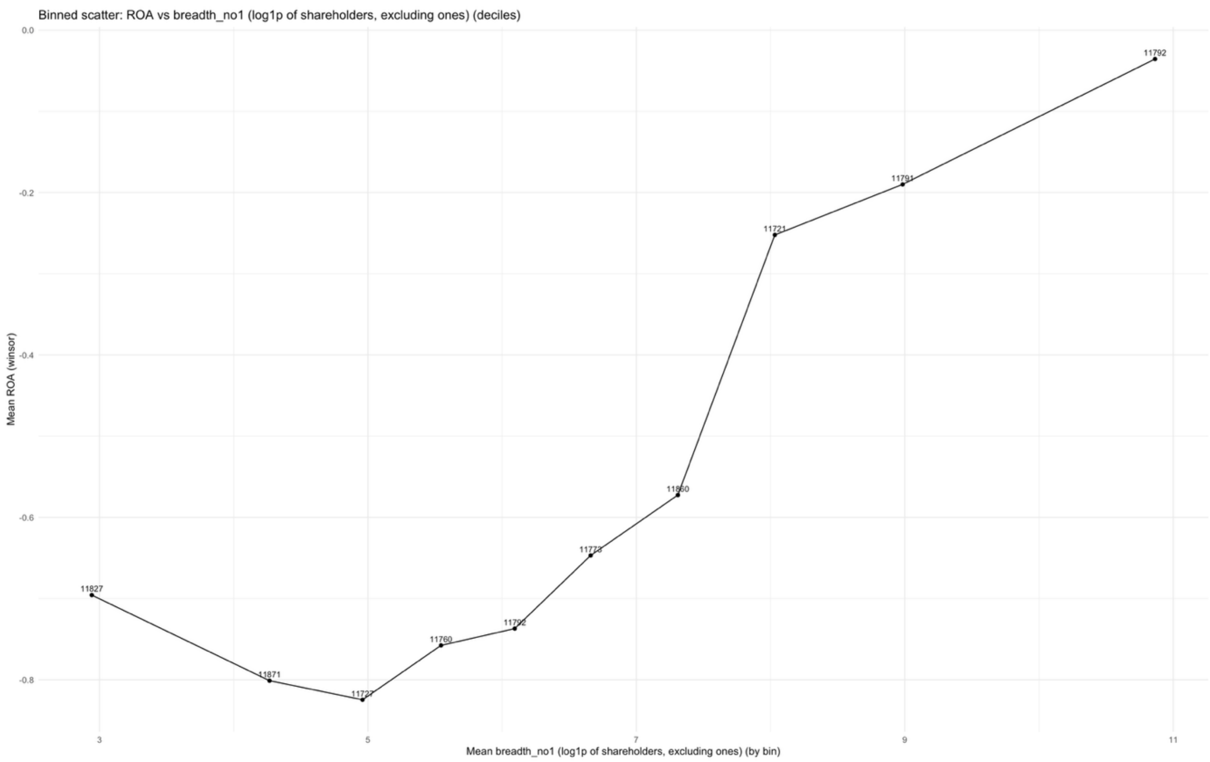

Figure 17 – Binned scatter of ROA winsorized on breadth_no1

5.1.2. Rank-based correlation and permutation tests

This subsection formalizes the descriptive evidence with a distribution-free procedure that targets monotone association without committing to a functional form or metric scale. For each pair (X, Y) the statistic used is Spearman's rank correlation, defined as the Pearson correlation between the mid-ranks of X and the mid-ranks of Y, written as: $rho\_S = cor(rank(X), rank(Y))$. Because it depends only on ordering, rho_S is robust to outliers, to nonlinear rescalings and to the heavy tails typical of accounting variables. All calculations are performed on exact finite pairs, the optical "zoom" in the figures improves legibility only and does not affect the sample used for inference. ROA here is the raw variable, not winsorized, breadth is considered both in levels as cshr_abs and in its log1p transformation as breadth_no1.

Significance is assessed with a permutation test that avoids auxiliary distributional assumptions. Holding X fixed, the Y values are randomly permuted without replacement, a new Spearman rho is computed and this operation is repeated B times. The two-sided permutation p-value is the fraction of draws with absolute statistic at least as large as the observed one, written as: $p\_perm = (1/B) * sum\_\{b = 1\}\^B\ 1(|rho\^(b)| >= |rho\_obs|)$. The number of permutations is set to B = 2000, a deliberate balance between precision and computation on the full panel; with two thousand draws the minimal attainable two-sided p-value is about 0.0005, which already exceeds the precision required given the large gaps observed between the null distribution and the realized statistics. A fixed random seed is used to guarantee exact reproducibility of the Monte Carlo draws.

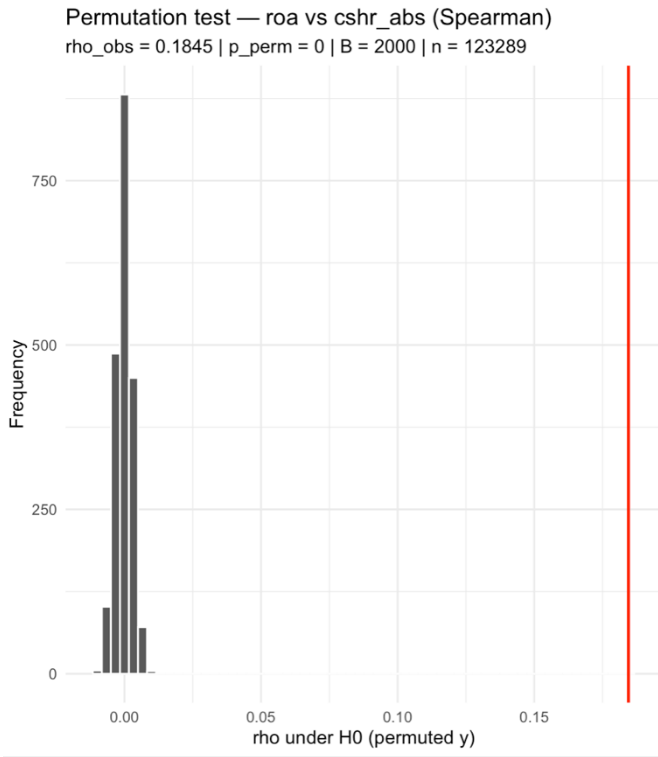

Figure 18 – Permutation test – ROA vs Shareholder absolute count

The first test relates ROA to the level of shareholder breadth using all firm-years with finite values (n = 123,289). The observed rank correlation equals rho_obs = 0.1845. The permutation distribution in Figure 18 is tightly centered near zero, with virtually all simulated statistics lying far to the left of the observed vertical line. The resulting permutation p-value is numerically indistinguishable from zero at the chosen Monte Carlo resolution. The magnitude indicates a modest but clear monotone association, meaning that firms with more shareholders tend to rank higher in the cross-sectional distribution of ROA.

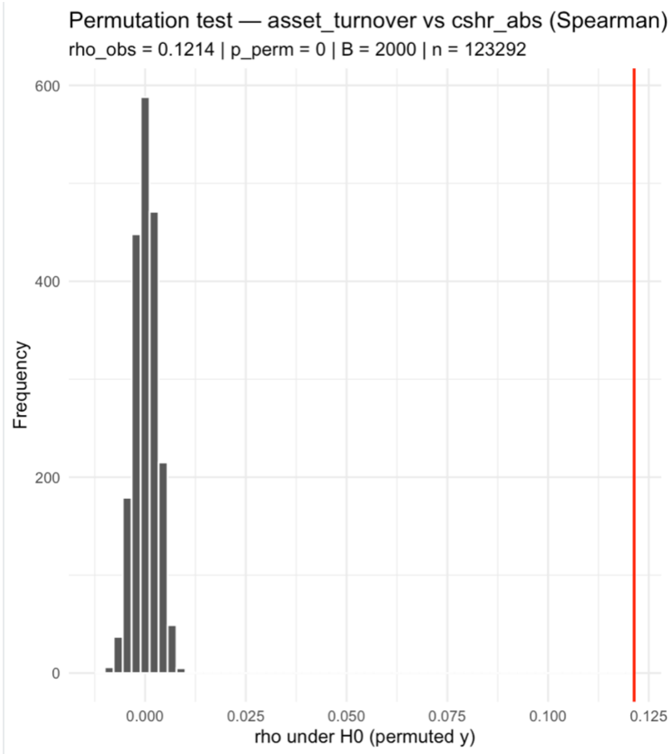

Figure 19 – Permutation test – Asset turnover vs Shareholder absolute count

The second test pairs asset turnover with cshr_abs on all finite observations (n = 123,292). The observed statistic is rho_obs = 0.1214. As displayed in Figure 19, the null histogram again concentrates around zero and remains far from the observed value, producing a permutation p-value that rounds to zero with B = 2000. The correlation is weaker than for ROA, which is informative, breadth appears more closely aligned with profitability than with sheer throughput, consistent with channels that operate through margins, cost discipline and governance rather than through proportional expansion of sales over assets.

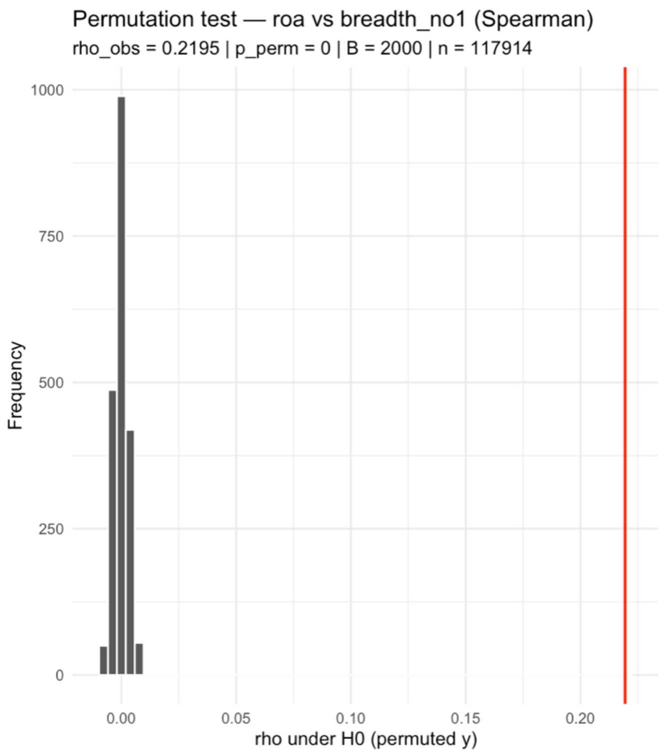

Figure 20 – Permutation test – ROA winsorized vs breadth_no1

The third test uses ROA against the log1p transformation of breadth, breadth_no1, after excluding one-holder firm-years to avoid degenerate logs (n = 117,914). The observed rank correlation increases to rho_obs = 0.2195. Figure 5.1.G shows that the permuted distribution remains narrowly centred near zero, with no overlap with the observed statistic, again yielding a permutation p-value effectively equal to zero at our simulation precision. The higher rho under the log mapping indicates that the empirically informative variation in recognition resides in the move from very low to moderate breadth rather than in the extreme right tail, which the logarithm compresses.

From an economic perspective, these rank-based results provide a clean, model-free statement. In the pooled panel, broader investor recognition is positively associated with operating performance, more strongly with profitability than with asset turnover and the association is most pronounced when breadth is considered on a logarithmic scale that accords more weight to gains at the low-recognition margin. Several non-causal explanations remain viable at this stage, including reverse causality if profitable firms attract more investors and common drivers such as size, leverage or industry shocks that comove with both breadth and performance. The point of this subsection is not to adjudicate causality, but to document that the association is present in the raw data, is not an artifact of a few extreme points and does not depend on linearity assumptions. The next subsection therefore turns to firm- and year-fixed-effects regressions with standard controls and flexible shapes, where identification is driven by within-firm changes over time and where effect sizes can be read as partial associations conditional on time-invariant heterogeneity and common shocks.

5.2. Parametric evidence: firm- and year-fixed-effects models for ROA

This section examines the within-firm association between breadth and operating performance using linear panel regressions with firm and year fixed effects. The dependent variable is the winsorized return on assets, denoted roa_win, constructed as described in Chapter 3. ROA is defined as ni divided by at, winsorization is applied at the pooled first and ninety-ninth percentiles and standard errors are clustered at the firm level throughout. The analysis proceeds in two steps, first estimating models that include breadth as the only regressor and fixed effects, then augmenting the specification with time-varying controls for size and leverage that are consistently available in the panel. All models are run on the samples created in the code block you provided and the exact outputs, including coefficients, standard errors, p-values and goodness-of-fit diagnostics, are referenced directly.

5.2.1. Breadth-only fixed-effects models

The baseline within-firm specification relates changes in breadth to changes in operating performance while absorbing permanent firm heterogeneity and common shocks by year. The regression estimated is:

$$roa\_win\_it = \alpha\_i + \tau\_t + \beta 1 * breadth\_no1\_it + u\_it$$

where α_i are firm fixed effects, τ_t are year fixed effects, breadth_no1 is the log1p transformation of the shareholder count excluding single-owner cases and u_it is the error term clustered by firm.

	Linear FE		
	Est.	Std. Error	p-value
breadth_no1	0.040***	0.008	<0.001
Num.Obs.	117914		
RMSE	1.64		
R2	0.559		
R2 Adj.	0.510		

Table 1 - Results of linear fixed-effects regression (ROA, breadth only)

On the breadth-eligible sample that excludes one-shareholder observations and contains finite outcomes, the estimation uses 117,914 firm-years, the within R^2 is approximately 0.00040 and RMSE is about 1.641. The coefficient on breadth_no1 is positive and statistically significant, β1 = 0.039603 with a clustered standard error of 0.00832 and a two-sided p-value of 1.96e-06. This indicates that, within the same firm and after absorbing year effects, years with higher recognition as captured by breadth_no1 tend to be associated with slightly higher operating profitability.

To allow for local curvature while keeping the functional form parsimonious, a centered quadratic is estimated by subtracting the sample mean of breadth on the modeling sample, μ_b = 6.56357, so that the linear term captures the marginal effect at the center of support. The regression is:

$$roa\_win\_it = \alpha\_i + \tau\_t + \gamma 1 * (breadth\_no1\_it - \mu\_b) + \gamma 2 * (breadth\_no1\_it - \mu\_b)\text{^}2 + u\_it$$

	Quadratic FE		
	Est.	Std. Error	p-value
I(breadth_no1 - 6.56357030137323)	0.042***	0.009	<0.001
I((breadth_no1 - 6.56357030137323)^2)	−0.003+	0.002	0.055
Num.Obs.	117914		
RMSE	1.64		
R2	0.559		
R2 Adj.	0.510		

Table 2 - Results of quadratic fixed-effects regression (ROA, breadth only)

The linear term is positive and significant, $\gamma1$ = 0.042127 with a standard error of 0.009095 and a p-value of 3.66e-06, while the quadratic term is negative and borderline significant, $\gamma2$ = −0.003202 with a standard error of 0.001671 and a p-value of 0.0554. The implied turning point in original breadth_no1 units is 13.142, computed as μ_b − $\gamma1$ divided by twice $\gamma2$, which lies very near the extreme right tail of the empirical support. The within R^2 increases modestly to about 0.000446 and RMSE remains 1.641, which is consistent with a gentle concavity that manifests only at very high breadth.

As a flexible robustness check that avoids polynomial extrapolation, a cubic B-spline with three degrees of freedom is introduced while retaining firm and year fixed effects:

$$roa\_win\_it = \alpha\_i + \tau\_t + s(breadth\_no1\_it) + u\_it$$

where s(·) denotes a B-spline basis.

	Spline FE (df=3)		
	Est.	Std. Error	p-value
splines = bs(breadth_no1, df = 3)1	0.387	0.348	0.266
splines = bs(breadth_no1, df = 3)2	1.020*	0.452	0.024
splines = bs(breadth_no1, df = 3)3	−0.825	0.761	0.279
Num.Obs.	117914		
RMSE	1.64		
R2	0.559		
R2 Adj.	0.510		

Table 3 - Results of spline fixed-effects regression (ROA, breadth only)

In the breadth-only sample of 117,914 firm-years, the spline terms jointly provide a similarly small but non-zero increase in within R^2, the second basis coefficient is positive and significant at the five percent level and the first and third basis terms are not statistically different from zero at conventional thresholds. RMSE is again 1.641. This pattern aligns with the centered quadratic, namely a positive slope over the interior of the distribution with mild flattening in the far right tail.

Across the three fixed-effects specifications estimated on the same firm-year sample with gvkey and fyear effects and firm-clustered standard errors, the evidence is remarkably consistent. The linear model yields a positive and statistically significant association between investor breadth and profitability, with the coefficient on breadth_no1 around 0.04 and stable RMSE and adjusted R^2 that mirror the alternative forms. Introducing curvature does not overturn this result. In the quadratic specification, the centered linear term remains positive and significant at roughly 0.042, while the squared term is negative and only weakly significant at the ten percent level, implying at most a mild concavity concentrated near the upper tail; the implied turning point lies around breadth_no1 ≈ 13.1, a region where data density is thin and inference is therefore fragile. The spline with three degrees of freedom paints a similar picture, preserving a positive average gradient and hinting at a slight attenuation at very high breadth, with only the central basis coefficient reaching conventional significance and the others not differing clearly from zero. Within-firm explanatory power is virtually identical across models, with RMSE near 1.64, adjusted R^2 around 0.51 and very small within R^2 in all cases, indicating that allowing modest nonlinearity adds little incremental fit relative to the linear benchmark. Taken together, these results support a robust, predominantly linear within-firm association between a broader shareholder base and higher contemporaneous ROA, with any concavity at high breadth levels being modest and statistically tenuous; consequently, the linear specification captures the essential relationship, while the quadratic and spline formulations function as confirmatory robustness checks rather than sources of qualitatively different conclusions.

5.2.2. Fixed-effects models with controls

The second step augments the regression with balance-sheet controls in order to address the possibility that changes in breadth co-move with changes in firm scale and leverage, which are known to be strongly related to profitability and could confound the breadth coefficient. The controls are the natural logarithm of total assets and the natural logarithm of total liabilities, both centered on their modeling-sample means to stabilize interpretation and numerics. The linear controlled specification is:

$$roa\_win\_it = \alpha\_i + \tau\_t + \beta 1 breadth\_c\_it + \beta 2 size\_c\_it + \beta 3 * liab\_c\_it + u\_it$$

where breadth_c is breadth_no1 minus μ_b, size_c is log_total_assets minus μ_s, liab_c is log_total_liabilities minus μ_l and μ_b, μ_s, μ_l are the corresponding sample means computed in the code. Estimation on the complete-regressor sample uses 117,536 firm-years.

	Linear FE + controls		
	Est.	Std. Error	p-value
breadth_c	–0.033***	0.007	<0.001
size_c	1.542***	0.031	<0.001
liab_c	–0.916***	0.024	<0.001
Num.Obs.	117536		
RMSE	1.32		
R2	0.715		
R2 Adj.	0.683		

Table 4 - Results of linear fixed-effects regression with controls (ROA)

The within R^2 rises substantially to 0.351, RMSE declines to 1.319 and the coefficients on the controls are large in magnitude and precisely estimated, β2 = 1.542 with a standard error of 0.0311 and β3 = −0.916 with a standard error of 0.0238, both with p-values below 2.2e-16, which is consistent with larger firms exhibiting higher operating margins while higher leverage is associated with lower contemporaneous ROA after fixed effects. The coefficient on breadth_c turns negative and significant, β1 = −0.033354 with a standard error of 0.00676 and a p-value of 8.15e-07. This sign reversal relative to the breadth-only models indicates that the positive association in the simpler fixed-effects regressions is accounted for by concurrent changes in size and leverage, once these controls are included the partial within-firm association between breadth and ROA is small and adverse on average.

To probe for curvature in a controlled setting, quadratic terms are added for breadth and for size:

$$roa\_win\_it = \alpha\_i + \tau\_t + \gamma 1 breadth\_c\_it + \gamma 2 breadth\_c\_it\hat{\ }2 + \delta 1 size\_c\_it + \delta 2 size\_c\_it\hat{\ }2 + \beta 3 * liab\_c\_it + u\_it$$

	Quadratic FE + controls		
	Est.	Std. Error	p-value
breadth_c	−0.005	0.006	0.376
I(breadth_c^2)	−0.001	0.001	0.669
size_c	0.812***	0.024	<0.001
I(size_c^2)	−0.087***	0.002	<0.001
liab_c	−0.607***	0.021	<0.001
Num.Obs.	117536		
RMSE	1.23		
R2	0.753		
R2 Adj.	0.725		

Table 5 - Results of quadratic fixed-effects regression with controls (ROA)

The within R^2 increases to 0.438 and RMSE declines to 1.228. The size terms display the expected concavity, δ1 = 0.8117 with a standard error of 0.0237 and δ2 = −0.08663 with a standard error of 0.00242, both highly significant, while leverage remains strongly negative, β3 = −0.6067 with a standard error of 0.0215. The breadth linear and quadratic terms are statistically indistinguishable from zero at conventional levels, γ1

= −0.00546 with a p-value of 0.376 and $\gamma 2$ = −0.000519 with a p-value of 0.669, suggesting that once size and leverage dynamics are held constant, a flexible quadratic in breadth does not contribute explanatory power for ROA.

As a non-parametric robustness within the regression framework, a cubic B-spline in breadth is combined with the quadratic in size and the linear leverage term:

$$roa\_win\_it = \alpha\_i + \tau\_t + s(breadth\_no1\_it) + \delta 1 size\_c\_it + \delta 2 size\_c\_it\text{^}2 + \beta 3 * liab\_c\_it + u\_it$$

	Spline FE (df=3) + controls		
	Est.	Std. Error	p-value
splines = bs(breadth_no1, df = 3)1	0.456+	0.244	0.061
splines = bs(breadth_no1, df = 3)2	–0.982**	0.376	0.009
splines = bs(breadth_no1, df = 3)3	1.737*	0.777	0.025
size_c	0.812***	0.024	<0.001
I(size_c^2)	–0.087***	0.002	<0.001
liab_c	–0.607***	0.021	<0.001
Num.Obs.	117536		
RMSE	1.23		
R2	0.753		
R2 Adj.	0.725		

Table 6 - Results of spline fixed-effects regression with controls (ROA)

The within R^2 and RMSE match the quadratic-controls model to three decimal places, the second and third spline basis coefficients are significant at conventional levels while the first is marginal and the control coefficients remain virtually unchanged. This outcome confirms that allowing a flexible shape in breadth does not overturn the conclusion that size and leverage are the primary covariates of within-firm changes in ROA and that the partial association of breadth is weak once these channels are absorbed.

Interpretation follows directly from the estimates. The breadth-only models document a small positive within-firm association between recognition and profitability, consistent with non-parametric plots that show a weak upward slope for ROA over the interior mass of breadth. Once time-varying balance-sheet conditions are introduced, the breadth coefficient either turns negative in the linear specification or becomes statistically negligible in the quadratic and spline specifications, while the controls absorb substantial variation with economically intuitive signs and strong significance. The progression in within R^2 from roughly 0.0004 to 0.351 and then to 0.438, coupled with the corresponding decline in RMSE from 1.641 to 1.319 and then to 1.228, indicates that changes in scale and leverage dominate the predictable component of within-firm movements in ROA. The centered quadratic and spline results point to clear concavity in the scale–profitability relationship and to, at most, mild curvature in the breadth–ROA association.

Two implementation notes complete the discussion. First, centering breadth and size before introducing squares improves numerics and interpretation since the linear terms can be read as marginal effects at the mean of the regressors. Second, winsorizing ROA at the pooled first and ninety-ninth percentiles reduces variance and skewness by orders of magnitude without altering the median or the IQR, as shown in Chapter 3, which stabilizes least-squares estimation and makes clustered inference well behaved, especially in the presence of rare extreme accounting realizations.

5.3. Concluding remarks

Taken together, the fixed-effects estimations portray a coherent but nuanced picture of the ROA–breadth relation. In the breadth-only specifications, using winsorized ROA as the dependent variable and firm and year effects with standard errors clustered at the firm level, the linear model delivers a positive and statistically significant association between breadth and profitability (coefficient ≈ 0.0396 on breadth_no1). Allowing for flexible shapes does not overturn this message: the quadratic term is weakly concave and only marginally significant and the spline fit shows limited curvature concentrated in the sparsest parts of the support. These models exhibit very small within R^2, which is expected for noisy firm-year accounting outcomes once persistent firm heterogeneity and common time shocks are absorbed; their role is to identify sign and shape rather than to maximize in-sample fit.

Once standard controls are included, especially size and leverage, most of the association between breadth and ROA is absorbed. The linear controlled model shows the breadth coefficient shrinking markedly and turning negative, the quadratic specification renders the breadth terms statistically indistinguishable from zero and the spline with controls yields mixed, economically small components. At the same time, within R^2 rises substantially into the 0.35–0.44 range, indicating that the additional regressors capture meaningful within-firm variation in ROA. Economically, this pattern suggests that the positive breadth–ROA link observed in the parsimonious models is largely mediated by firm scale and capital structure: firms tend to expand their investor base as they grow and deleverage and those balance-sheet changes, rather than breadth per se, are the proximate drivers of profitability dynamics. The quadratic turning point implied by the breadth-only model lies at an extreme of the breadth distribution, in a region with very low data density, which cautions against attributing substantive economic content to the mild concavity detected there.

In sum, the fixed-effects evidence is consistent with a modest, positive within-firm association between investor breadth and operating performance when examined in isolation, but this association is not robust to the inclusion of standard balance-sheet controls and appears to proxy for concurrent shifts in size and leverage.

6 Market Valuation: shareholder-base breadth and Tobins Q

6.1. Non-parametric evidence

6.1.1. Visual diagnostics: Tobin's Q versus shareholder breadth

The first cross-plot relates Q in levels to the level measure of the shareholder base, cshr_abs (see Figure 21).

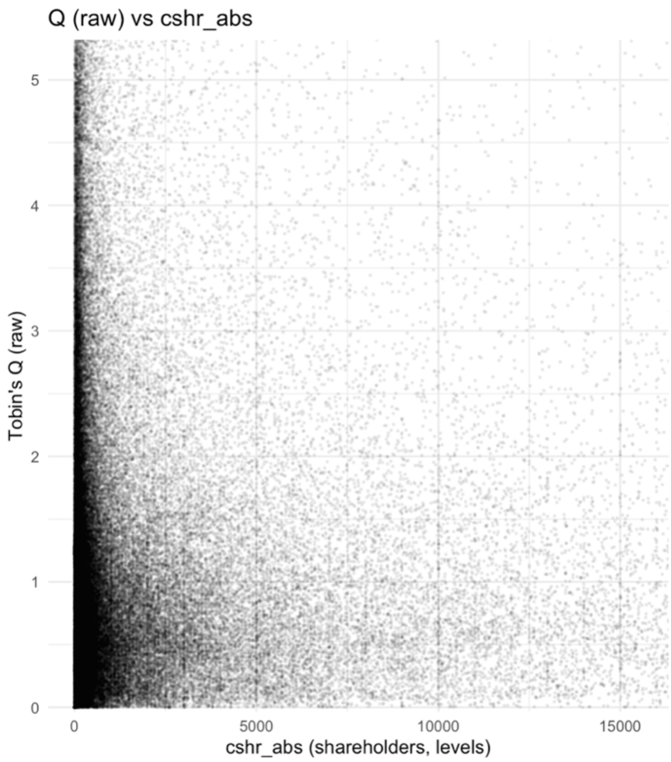

Figure 21 – Tobin's Q vs Shareholder absolute count

To prevent extremes from dictating perception, an optical zoom focuses the axes on their central deciles while keeping all points. The cloud is diffuse and nearly patternless, with only a mild negative slope on average, indicating that proportional changes in the number of shareholders are not associated with large systematic movements in raw Q once the visual influence of extremes is moderated. This is consistent with the notion that, without controls, differences in scale and in the cycle can mask the relationship of interest

The second cross-plot (see Figure 22) uses the transformed outcome q_log_w against the logarithmic breadth measure breadth_no1, defined as $log(1 + shareholders)$ and excluding one-holder cases, thereby aligning both sides on relative scales.

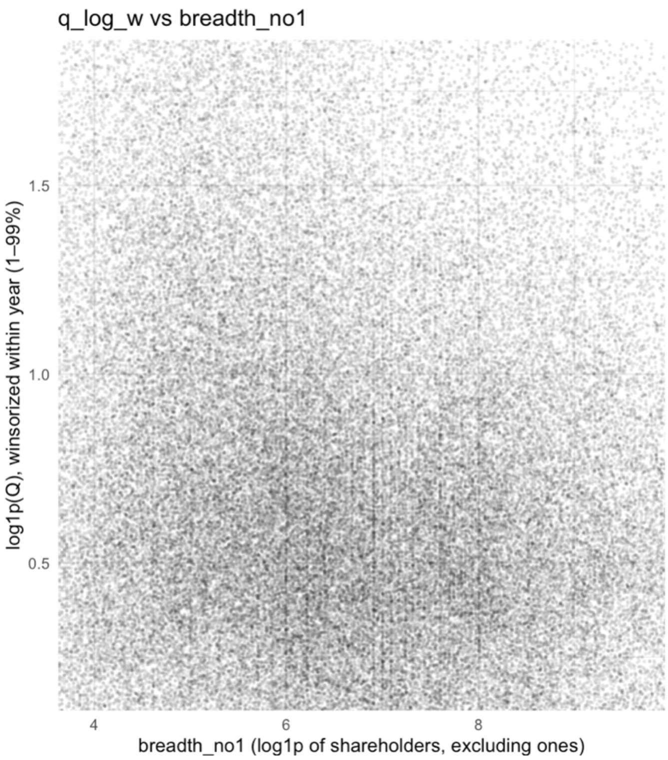

Figure 22 – Log1p of Tobin's Q winsorized vs breadth_no1

The scatter again concentrates around the center with a weak negative slope. The transformation improves readability with respect to levels, preserving ranks while limiting pointwise influence from the tails and revealing the structure of the interior mass.

6.1.2. Rank-based correlation and permutation test

Unconditional association is quantified with a rank statistic that is robust to outliers and monotone transformations. For variables X and Y, the Spearman correlation is $\rho\_S = corr(rank(X), rank(Y))$, computed on the valid pairwise sample. The null hypothesis of no association is tested by randomization, permuting Y without replacement and recomputing ϱ_S for B independent replications. The two-sided p-value is the fraction of permuted statistics whose absolute value is at least as large as the observed one. The choice B = 2000 is deliberate and pragmatic, it keeps computation fast and the Monte Carlo standard error of a small p about $sqrt(p(1 - p) \div B)$, which remains well below one percentage point at the reporting precision.

Applied to Q in levels versus cshr_abs, the statistic equals ϱ_S = −0.1347 on n = 112,201 valid pairs, with p_perm = 0.0000 at the displayed precision (Figure 23). Repeating the exercise for q_log_w versus breadth_no1 yields ϱ_S = −0.1370 on n = 111,997, with p_perm = 0.0000 (Figure 24). The sign is consistently negative across pairs and scales, yet the magnitude is small, which is coherent with the very high cross-sectional dispersion of valuation ratios and with the presence of confounding factors. In particular, breadth and Q both covary with firm size and leverage and market regimes shift over time. These unconditional results are therefore not interpreted causally. The parametric analysis with firm and year fixed effects, complemented by essential controls, is designed precisely to purge unobserved heterogeneity and to assess whether the sign persists, attenuates or changes once conditioning on state variables and on time-invariant firm characteristics is imposed.

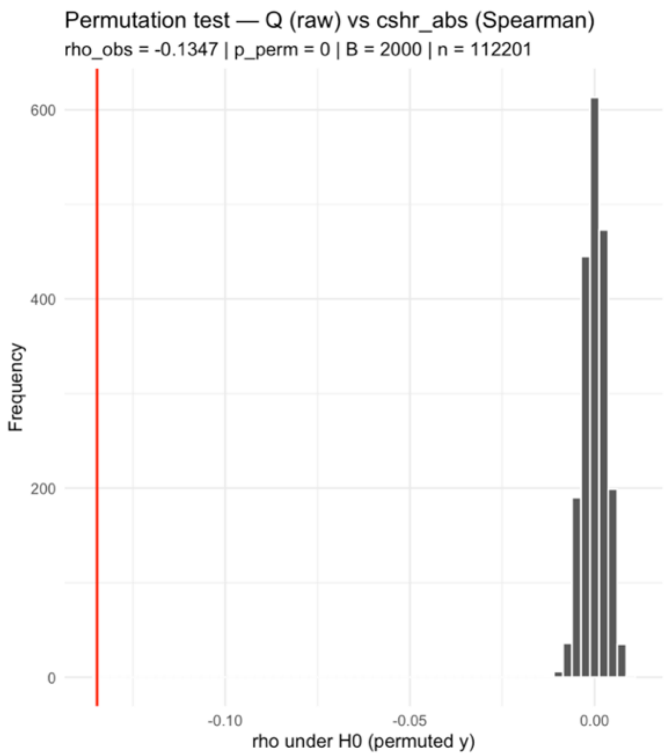

Figure 23 – Permutation test – Tobin's Q vs Shareholder absolute count

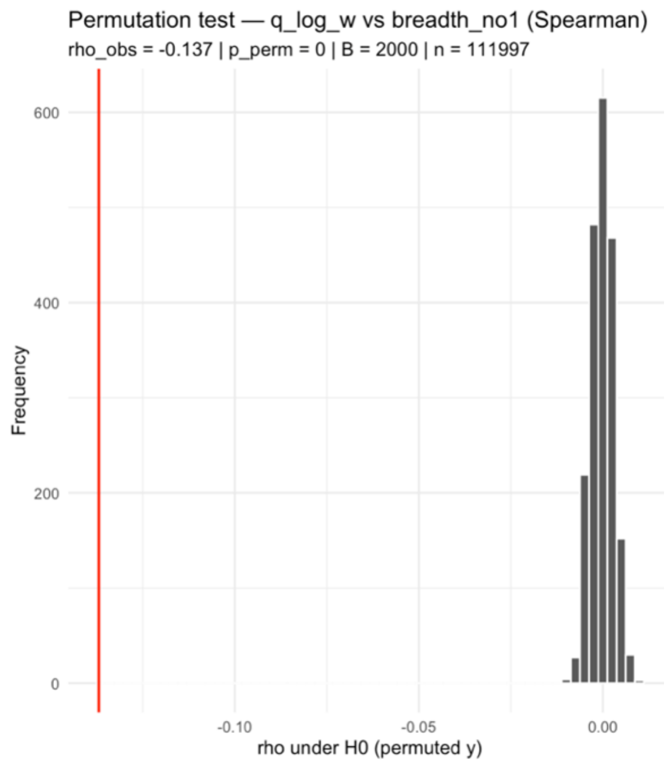

Figure 24 – Permutation test – Log1p of Tobin's Q winsorized vs breadth_no1

From an economic standpoint, the negative unconditional association can reflect multiple mechanisms operating simultaneously. Mature, widely followed firms may exhibit broader shareholder bases yet lower price-to-book ratios than young, high-growth firms, which generates a negative relation between breadth and Q when lifecycle and investment opportunities are not controlled. During strong market upswings the dispersion of Q also widens and can temporarily decouple from the dynamics of accounting fundamentals, further flattening any average slope. To evaluate the relationship relevant for the recognition hypothesis, the next section conditions on firm and year effects and on size and leverage and operates on q_log_w where variance is more regular and curvature is not dominated by episodic tails.

6.2. Parametric evidence: firm- and year-fixed-effects models for valuation

This section mirrors the ROA analysis and studies the association between investor recognition and valuation using Tobin's Q. The dependent variable is constructed as described earlier, year-wise winsorization at the first and ninety-ninth percentiles controls episodic extremes and a log1p transformation is used where indicated to stabilize right-tail dispersion while preserving zeros. All specifications include firm fixed effects and fiscal-year fixed effects and standard errors are clustered at the firm level. The focus is on within-issuer movements, therefore coefficients are interpreted as conditional associations for the same firm over time.

6.2.1. Breadth-only fixed-effects models

The starting point considers recognition in isolation, without additional time-varying controls. The baseline linear fixed-effects regression for the log specification is:

$$q\_log\_w\_it \;=\; \alpha\_i \;+\; \gamma\_t \;+\; \beta \;\cdot\; breadth\_no1\_it \;+\; u\_it$$

where q_log_w_it denotes log1p of within-year winsorized Q for firm i in year t, α_i and γ_t are firm and year effects and u_it is the error term clustered by firm. The estimate for β is negative and statistically significant, β ≈ −0.015 with a t statistic around −4.40 on a sample of about 112 thousand firm-years. The within R^2 is very small, as expected when a single regressor is used with a noisy dependent variable in a high-dimensional fixed-effects design. This first result is an unconditional benchmark that does not yet partial out other dimensions correlated with breadth and valuation, it anticipates the role of size and balance sheet composition that will be addressed below.

	Linear FE		
	Est.	Std. Error	p-value
breadth_no1	−0.015***	0.003	<0.001
Num.Obs.	111997		
RMSE	0.45		
R2	0.711		
R2 Adj.	0.677		

Table 7 - Results of linear fixed-effects regression (log Tobin's Q, breadth only)

To allow for flexible marginal effects of recognition, a quadratic specification is estimated using a centered regressor to improve interpretability and numerical stability, with the turning point computed on the original scale. The model is:

$$q\_log\_w\_it = \alpha\_i + \gamma\_t + \beta\_1 \cdot (breadth\_no1\_it - \mu\_b) + \beta\_2 \cdot (breadth\_no1\_it - \mu\_b)\text{^}2 + u\_it$$

where μ_b is the sample mean of breadth_no1 on the estimation sample. The linear term is negative and significant, $\beta\_1 \approx -0.0177$, the squared term is positive and significant, $\beta\_2 \approx 0.0027$. The implied turning point is at breadth_no1 ≈ 9.913. Since the empirical mass of breadth_no1 is concentrated well below this threshold, the in-sample slope is largely negative but becomes less negative as breadth increases, indicating mild convexity without a pronounced U shape in the region where data are dense.

	Quadratic FE		
	Est.	Std. Error	p-value
I(breadth_no1 - 6.63454093506704)	−0.018***	0.004	<0.001
I((breadth_no1 - 6.63454093506704)^2)	0.003***	0.001	<0.001
Num.Obs.	111997		
RMSE	0.45		
R2	0.711		
R2 Adj.	0.677		

Table 8 - Results of quadratic fixed-effects regression (log Tobin's Q, breadth only)

A nonparametric spline provides an additional shape check. The specification is:

$$q\_log\_w\_it = \alpha\_i + \gamma\_t + \theta\_1 \cdot bs\_1(breadth\_no1\_it) + \theta\_2 \cdot bs\_2(breadth\_no1\_it) + \theta\_3 \cdot bs\_3(breadth\_no1\_it) + u\_it$$

with a B-spline basis of degree and degrees of freedom matching the ROA chapter for comparability. The first basis coefficient is negative and marginally significant, the remaining terms are not precisely estimated and overall the fitted curve is nearly linear over the interior support. These results corroborate the quadratic model's indication of weak curvature and a predominantly flat to slightly negative unconditional association.

	Spline FE (df=3)		
	Est.	Std. Error	p-value
splines = bs(breadth_no1, df = 3)1	−0.336*	0.157	0.032
splines = bs(breadth_no1, df = 3)2	−0.306	0.257	0.233
splines = bs(breadth_no1, df = 3)3	0.450	0.588	0.444
Num.Obs.	111997		
RMSE	0.45		
R2	0.711		
R2 Adj.	0.677		

Table 9 - Results of spline fixed-effects regression (log Tobin's Q, breadth only)

Taken together, the breadth-only evidence suggests that recognition is not strongly informative for valuation once firm and year effects are partialled out and that unconditional slopes can be slightly negative. Given the known positive cross-sectional relation between size, recognition and valuation and the mechanical effect of leverage on Q, the breadth-only patterns are best treated as a preliminary baseline subject to omitted-variable concerns.

6.2.2. Fixed-effects models with controls

The next step augments the recognition regressions with time-varying controls that capture firm scale and balance sheet structure. The dependent variable is kept as q_log_w to preserve the stabilized distribution. The main specification is

$$q\_log\_w\_it = \alpha\_i + \gamma\_t + \beta \cdot breadth\_c\_it + s\_1 \cdot size\_c\_it + s\_2 \cdot size\_c\_it\hat{}2 + \ell \cdot lev\_c\_it + u\_it$$

where $breadth\_c\_it = breadth\_no1\_it - \mu\_b$, $size\_c\_it = log\_total\_assets\_it - \mu\_s$ and l$ev_c_it = log_total_liabilities_it - \mu_\ell$. Centering follows the ROA chapter and improves interpretability by making the intercept correspond to the mean firm in the estimation sample, it also reduces collinearity in the quadratic terms.

In the linear specification with controls the coefficient on breadth turns positive and statistically significant, $\beta \approx 0.010$ with a t statistic around 3.24, while size enters with a large negative linear coefficient and a positive quadratic term in the subsequent models, consistent with concavity in log assets. Leverage enters with a small positive coefficient in the linear specification and with a negative coefficient when quadratic size terms are included, reflecting the different partitioning of variance once curvature in size is allowed. The within R^2 increases materially relative to breadth-only models, to roughly 18 percent in the baseline with linear controls and around 26 percent in the quadratic and spline variants, indicating that time-varying fundamentals explain a nontrivial share of within-firm valuation movements.

	Linear FE + controls		
	Est.	Std. Error	p-value
breadth_c	0.010**	0.003	0.001
size_c	–0.262***	0.008	<0.001
lev_c	0.013*	0.006	0.037
Num.Obs.	111660		
RMSE	0.41		
R2	0.763		
R2 Adj.	0.735		

Table 10 - Results of linear fixed-effects regression with controls (log Tobin's Q)

Allowing for curvature in breadth and size, the specification becomes:

$$q\_log\_w\_it = \alpha\_i + \gamma\_t + \beta\_1 \cdot breadth\_c\_it + \beta\_2 \cdot breadth\_c\_it\hat{}2 + s\_1 \cdot size\_c\_it + s\_2 \cdot size\_c\_it\hat{}2 + \ell \cdot lev\_c\_it + u\_it$$

The linear term in breadth is close to zero and not statistically different from zero, β_1 ≈ −0.0015, while the squared term is positive and significant, β_2 ≈ 0.0021. This pattern implies a convex relation where the marginal association of recognition with valuation increases with breadth. Given the centering of breadth, the turning point for the marginal effect is at breadth_c ≈ 0.36 on the centered scale, which corresponds to breadth_no1 modestly above its sample mean. In practice, for firms with recognition below the mean the conditional association is near zero to slightly negative, whereas for firms at or above the mean the association turns positive and becomes economically meaningful as breadth increases. Quadratic size terms enter with the expected signs, a negative linear coefficient and a positive quadratic coefficient, capturing diminishing returns of scale on valuation and leverage is negative and statistically significant once size curvature is included. The within R^2 rises to about 26 percent and the RMSE declines relative to the linear-controls model, pointing to improved fit through flexible controls rather than through overfitting breadth.

	Quadratic FE + controls		
	Est.	Std. Error	p-value
breadth_c	–0.001	0.003	0.628
I(breadth_c^2)	0.002**	0.001	0.003
size_c	–0.059***	0.007	<0.001
I(size_c^2)	0.024***	0.001	<0.001
lev_c	–0.075***	0.006	<0.001
Num.Obs.	111660		
RMSE	0.39		
R2	0.786		
R2 Adj.	0.761		

Table 11 - Results of quadratic fixed-effects regression with controls (log Tobin's Q)

A spline in breadth is finally overlaid on the same control set to let the data choose local curvature with minimal structure:

$$q\_log\_w\_it = \alpha\_i + \gamma\_t + \theta\_1 \cdot bs\_1(breadth\_no1\_it) + \theta\_2 \cdot bs\_2(breadth\_no1\_it) + \theta\_3 \cdot bs\_3(breadth\_no1\_it) + s\_1 \cdot size\_c\_it + s\_2 \cdot size\_c\_it\text{^}2 + \ell \cdot lev\_c\_it + u\_it$$

The spline coefficients are individually imprecise, the fitted relation is close to the quadratic shape and the global fit statistics are virtually identical to the quadratic-controls model. The absence of strong nonlinearity beyond a mild convex pattern supports using a quadratic summary for parsimony and interpretability.

	Spline FE (df=3) + controls		
	Est.	Std. Error	p-value
splines = bs(breadth_no1, df = 3)1	–0.281*	0.129	0.029
splines = bs(breadth_no1, df = 3)2	0.164	0.208	0.430
splines = bs(breadth_no1, df = 3)3	0.150	0.469	0.750
size_c	–0.059***	0.007	<0.001
I(size_c^2)	0.024***	0.001	<0.001
lev_c	–0.075***	0.006	<0.001
Num.Obs.	111660		
RMSE	0.39		
R2	0.786		
R2 Adj.	0.761		

Table 12 - Results of spline fixed-effects regression with controls (log Tobin's Q)

The term $splines::bs(breadth\_no1, df = 3)$ replaces a single slope coefficient with three basis components, each capturing variation in a different region of the breadth distribution. In this parametrization, the individual spline coefficients are not marginal effects in the usual sense, because each coefficient multiplies a basis function rather than breadth itself and the basis functions overlap. Nevertheless, the pattern and statistical content of the three coefficients is informative about where in the breadth distribution the conditional association is most pronounced. In the reported output, the first spline coefficient is negative and statistically different from zero at conventional levels, whereas the second and third coefficients are closer to zero and not statistically distinguishable from zero. This configuration implies that, once firm and year fixed effects and the controls are held constant, the economically relevant curvature is concentrated in the low to intermediate breadth region, while the incremental flexibility allowed by the spline at higher breadth values does not materially improve fit relative to a smoother, more parsimonious shape.

An economically coherent interpretation follows by comparing the spline pattern to the quadratic specification estimated on the same controlled sample, which indicates mild convexity in breadth and a turning point around breadth_no1 ≈ 9.9, that is, log(1 + shareholders excluding ones) near 10. In the controlled design, the negative loading on the first spline basis is consistent with the idea that, at low breadth, increases in measured recognition can be associated with lower valuation conditional on size and leverage, because early increases in breadth often coincide with transitions in firm status such as distress episodes, restructuring or risk repricing, where attention rises but valuation does not necessarily increase contemporaneously. At the same time, the absence of statistically strong additional spline components at higher breadth is consistent with diminishing heterogeneity in the conditional relation once the shareholder base becomes sufficiently deep, meaning that after a certain point, additional recognition tends to operate through smoother channels such as lower participation frictions, broader risk sharing and improved information aggregation, which are adequately captured by a simple convex shape rather than requiring multiple turning points. In this reading, the spline results do not contradict the quadratic evidence, they reinforce it by indicating that the non linearity is limited in complexity, the data support a single broad change in slope rather than oscillating patterns and the sign of the association becomes more favorable in the upper part of the breadth distribution, where recognition is plausibly linked to thicker investor markets and stronger valuation responses.

The controls play a central role in giving this interpretation economic content. The coefficients on size and leverage indicate that, within firms over time, valuation in log scale is strongly related to scale and financing conditions, with a negative linear size effect combined with positive curvature in size and a negative leverage effect. Because breadth tends to comove with both firm scale and balance sheet evolution, the breadth basis terms in the spline model should be interpreted as capturing the component of recognition that is orthogonal to these fundamental shifts. The fact that at least one breadth basis component remains statistically meaningful after absorbing size and leverage suggests that recognition contains incremental information for valuation beyond mechanical scale effects, but it also indicates that the valuation relevance of recognition is not uniform across the breadth distribution, it is more consistent with a regime where additional breadth matters more once the shareholder base is already sufficiently developed, rather than a strictly linear effect that applies equally at all levels of recognition.

6.3. Concluding remarks

Taken together, the valuation results deliver a coherent synthesis once the distinction between unconditional rank-based evidence and within firm-controlled evidence is kept explicit. The non-parametric analysis, based on Spearman correlations paired with permutation tests, points to a negative rank association between valuation and breadth proxies in the pooled data, both for Q (raw) versus cshr_abs and for q_log_w versus breadth_no1. This result is not surprising in a large panel where cross firm heterogeneity dominates, firms with very large shareholder counts tend to be mature, large and lower growth, whereas firms with very high Q values often correspond to smaller, high growth, high option value issuers, so a negative pooled rank relation can arise mechanically even if recognition is beneficial within a firm over time. The fixed effects framework addresses precisely this issue by removing time invariant firm heterogeneity and common year shocks, thereby shifting identification away from cross sectional differences and toward within firm changes in recognition.

Within that controlled fixed effects design, the breadth only specifications serve as a baseline rather than a decisive economic test, because they can still mix breadth movements with contemporaneous shifts in firm size and financing. Once size and leverage are included, the results become systematically more supportive of the recognition hypothesis, the conditional association between breadth and valuation turns positive on average and the shape evidence points to mild non-linearity rather than instability. The sign change between the breadth only and controlled specifications is therefore interpreted as an omitted variables phenomenon consistent with the economic structure of valuation, breadth, firm scale and leverage comove through risk, growth opportunities and financing channels and failing to account for those fundamentals loads shared variation onto the breadth term. The controlled spline and quadratic results refine the interpretation further, the data support a simple curvature where the marginal valuation relevance of recognition strengthens at higher breadth levels, consistent with the idea that recognition effects become more economically meaningful once participation frictions are already low enough that additional investors can translate into thicker trading, improved information aggregation and a more robust market valuation relative to book assets.

The overall reconciliation is therefore that the negative pooled rank evidence documents a cross-firm composition effect, whereas the fixed effects results, especially in the controlled models, target the within firm recognition channel central to the thesis question. Under that interpretation, the evidence is not internally inconsistent, it is complementary, the non-parametric tests describe the unconditional sorting of firms in the joint distribution of valuation and breadth, while the controlled fixed effects models isolate the economically relevant conditional association for the same issuer over time.

7 Conclusion

7.1. Summary of the Topic and Research Questions

This thesis investigates whether investor recognition, proxied by shareholder-base breadth, is associated with firm-level outcomes. The analysis is grounded in incomplete-information and limited-participation frameworks, where broader recognition improves risk sharing, reduces participation frictions and potentially affects both real performance and market valuation. Investor recognition is operationalized through the number of shareholders, transformed into a log-based breadth measure after careful data cleaning, interpolation constraints and consistent tail treatment. The empirical analysis addresses three central research questions.

- First, whether increases in investor recognition are associated with improvements in operating performance within firms over time.
- Second, whether increases in investor recognition are associated with higher market valuation within firms over time.
- Third, whether the relationship between breadth and these outcomes exhibits non-linear patterns.

The empirical strategy applies identical timing conventions, row filters, transformations and firm- and year-fixed effects across both outcomes, ensuring that differences in results reflect underlying economics rather than construction choices.

7.2. Empirical Findings on Operating Performance

The empirical evidence on accounting performance suggests that the association between investor recognition and contemporaneous operating profitability is limited in magnitude and should be interpreted with caution.

In the non-parametric analysis, the rank-based correlations between shareholder breadth and ROA are positive and statistically significant under permutation tests. However, their economic size is small. The visual diagnostics do not reveal strong structural patterns and the dispersion of ROA remains substantial across all levels of breadth. This indicates that, while the association is unlikely to be purely random, it is not economically large in unconditional terms.

The fixed-effects models provide a more structured assessment. In breadth-only specifications, increases in shareholder breadth within firms over time are associated with modest increases in winsorized ROA. This positive sign is consistent across linear and flexible functional forms. Nevertheless, the explanatory power of these models remains very low in within-firm terms, which is not unexpected given the volatility and firm-specific nature of accounting outcomes.

When firm size and leverage are introduced as controls, the estimated coefficient on breadth becomes smaller and, in some specifications, statistically weaker. This attenuation suggests that part of the initial positive association may reflect the joint evolution of breadth with underlying firm characteristics rather than an independent effect of recognition on profitability. Larger and more established firms may both attract broader shareholder bases and display different profitability dynamics, creating an apparent association in simpler specifications.

Taken together, the evidence does not support a strong or robust link between investor recognition and accounting profitability. The results are directionally consistent with a positive association, but the magnitude is modest and sensitive to the inclusion of controls. From a cautious perspective, the findings indicate that recognition does not appear to systematically improve short-run operating performance once time-invariant heterogeneity and key fundamentals are accounted for.

7.3. Empirical Findings on Market Valuation

The valuation results present a more complex but still nuanced picture, which requires careful reconciliation between unconditional and controlled estimates.

In the non-parametric analysis, the rank-based correlations between shareholder breadth and Tobin's Q are negative and statistically significant. In unconditional terms, firms with broader shareholder bases tend to display slightly lower valuation ratios. However, the magnitude of this correlation remains economically small. Moreover, unconditional associations do not distinguish between cross-sectional heterogeneity and within-firm dynamics.

Breadth, firm size and leverage are correlated across firms and over time and all three variables are related to valuation through risk, growth opportunities and capital structure considerations. Consequently, models that omit size and leverage may conflate the role of recognition with structural differences between firms.

The fixed-effects regressions without controls replicate the slightly negative association observed in the non-parametric analysis. Once size and leverage are introduced, the sign of the breadth coefficient becomes positive in the linear specification and the quadratic specification indicates convexity. This suggests that, conditional on firm fundamentals and fixed effects, increases in shareholder breadth within the same firm are associated with higher market valuation.

However, these effects remain economically moderate and the within R^2 values indicate that most of the variation in valuation is explained by firm effects and fundamentals rather than by breadth itself. The convexity observed in some specifications should also be interpreted cautiously, as it may reflect local curvature rather than a strong structural non-linearity.

The reconciliation of results can therefore be summarized as follows. The unconditional negative association appears to be driven by cross-sectional heterogeneity related to size and capital structure. Once this heterogeneity is controlled for, the within-firm association between recognition and valuation becomes positive, but its magnitude remains modest.

Overall, the evidence is consistent with the hypothesis that investor recognition may be more closely linked to pricing mechanisms than to operating performance. At the same time, the results do not establish a strong causal effect. They indicate an association that is statistically robust across specifications but economically moderate and dependent on proper conditioning. Any interpretation should therefore remain cautious and avoid attributing valuation dynamics solely to changes in shareholder breadth.

7.4. Contributions in light of the Dual Evidence

This thesis set out to investigate whether investor recognition, proxied by shareholder breadth, is systematically associated with firm performance and firm valuation and whether these associations operate through real operating channels or primarily through market pricing mechanisms. The empirical results allow these research questions to be addressed directly.

The first research question concerned whether increases in investor recognition are associated with improvements in accounting performance. The evidence suggests that the answer is limited and conditional. In non-parametric tests, shareholder breadth is positively correlated with ROA and this association is statistically significant. However, the magnitude is small. In fixed-effects regressions, the positive association persists in breadth-only specifications but becomes weaker and less stable once firm fundamentals are controlled for. This pattern indicates that part of the observed relationship reflects underlying firm characteristics that co-evolve with breadth rather than a direct operational impact of recognition. Therefore, while recognition does not appear unrelated to accounting outcomes, the results do not support a strong or economically large real effect on profitability.

The second research question examined whether investor recognition is associated with market valuation. Here the evidence is more structured. Unconditional correlations show a negative association between breadth and Tobin's Q. However, once firm and year fixed effects and key controls such as size and leverage are included, the sign reverses and becomes positive in the linear specification, with some evidence of convexity in the quadratic model. This shift suggests that cross-sectional heterogeneity masks the underlying within-firm relationship. Conditional on fundamentals, increases in shareholder breadth are associated with higher valuation. Although the magnitude remains moderate, the association is more robust and consistent than for ROA. This finding aligns with participation-based theoretical frameworks, in which broader investor bases reduce participation frictions and improve risk sharing, thereby affecting prices more directly than operating efficiency.

The third research question concerned the functional form of the relationship. The results show that linear specifications capture only part of the pattern. In the valuation regressions with controls, the quadratic term in breadth is positive and statistically significant, indicating convexity. This implies that marginal increases in recognition may have stronger valuation effects at higher levels of breadth. While spline estimates confirm some non-linearity, the evidence does not point to extreme curvature. Rather, it suggests moderate amplification at higher participation levels. For accounting performance, non-linearities are weaker and less consistent.

Taken together, the dual evidence clarifies the economic channel through which recognition operates. The results suggest that investor recognition is more closely linked to how firms are priced than to how they perform operationally. Recognition appears to function primarily as a pricing mechanism rather than a productivity mechanism. The absence of a strong real effect on ROA, combined with a conditional positive association with valuation, supports the interpretation that recognition affects discount rates, risk sharing or investor demand more than it affects internal efficiency.

These contributions are methodological as well as substantive. First, the thesis implements a consistent construction of shareholder breadth under identical timing conventions and econometric design across outcomes, allowing direct comparison between performance and valuation effects. Second, it explicitly separates unconditional associations from within-firm fixed-effects evidence, thereby clarifying the role of cross-sectional heterogeneity. Third, it tests for non-linearities within a unified framework, showing that potential convex effects emerge primarily in valuation models with controls.

Importantly, the findings do not imply a causal effect of recognition on firm outcomes. They document robust within-firm associations that are economically moderate and sensitive to conditioning. The contribution lies in clarifying where recognition appears to matter most and in distinguishing pricing dynamics from operating dynamics within a unified empirical design.

In light of the research questions posed in the introduction, the evidence suggests a nuanced answer: investor recognition is associated with firm outcomes, but its primary domain of relevance is market valuation rather than accounting profitability and its magnitude remains economically contained once firm fundamentals are accounted for.

7.5. Limitations

As with any empirical study, several limitations remain.

First, despite the use of firm and year fixed effects, causal interpretation remains limited. Reverse causality cannot be fully excluded, since improvements in fundamentals may attract investors.

Second, breadth measures the quantity of investors but not their composition. Differences between institutional and retail investors, ownership concentration or holding horizons are not captured.

Third, valuation remains influenced by sentiment and macroeconomic cycles, even after winsorization and log transformation. Tobin's Q is an imperfect proxy for expected growth and discount rates.

Fourth, the analysis focuses on contemporaneous associations and does not explore dynamic lag structures.

7.6. Avenues for Future Research

The empirical analysis developed in this thesis clarifies several dimensions of the relationship between investor recognition and firm outcomes, yet it also leaves open a number of questions that naturally invite further investigation. The results should therefore be interpreted not as a closed account of the role of shareholder breadth, but as a step within a broader research trajectory concerned with participation, information frictions and asset pricing.

A central issue that deserves deeper exploration concerns the causal interpretation of the documented associations. The empirical design relies on within-firm variation combined with firm and year fixed effects, which substantially reduces concerns related to time-invariant heterogeneity and common macroeconomic shocks. Nevertheless, the analysis remains observational. Changes in shareholder breadth may reflect shifts in firm quality, growth opportunities or risk that are only imperfectly captured by observable controls. Future research could therefore seek institutional or regulatory settings in which variation in investor participation is plausibly exogenous to firm fundamentals. Episodes such as index inclusions or exclusions, regulatory reforms affecting ownership disclosure or trading access, brokerage entry or exit or changes in investment mandates could provide quasi-experimental variation in recognition. Exploiting such contexts would allow a sharper distinction between recognition as a causal determinant of valuation and recognition as an endogenous outcome of firm performance and visibility.

Further progress could also be made by refining the measurement of recognition itself. In this thesis, shareholder breadth is constructed as a participation-based proxy, capturing the dispersion of ownership across investors. While this approach aligns closely with theoretical participation models, it abstracts from the heterogeneity of investors. Institutional, retail, foreign and long-term blockholders may differ in their informational role, monitoring capacity and impact on pricing. Disaggregating breadth across investor types would make it possible to test whether the documented valuation effects are concentrated among specific categories of shareholders. Moreover, combining participation-based measures with alternative proxies of visibility or attention, such as analyst coverage, media intensity or search-based indicators, could help disentangle whether recognition operates primarily through reduced participation constraints, enhanced information diffusion or broader salience in financial markets.

The temporal dimension of the relation between recognition and outcomes also warrants closer examination. The models estimated in this thesis focus on contemporaneous associations within firms. Although this strategy is well suited to identifying conditional relationships, it does not fully address the dynamic ordering between changes in breadth and changes in valuation. It remains possible that increases in recognition precede valuation adjustments, that they follow periods of strong performance or that both variables evolve jointly in response to unobserved shocks. Incorporating richer lag structures, dynamic panel specifications or event-study methodologies around substantial shifts in ownership composition could provide insight into the persistence and sequencing of recognition effects.

Broadening the set of dependent variables would further enhance the understanding of the economic mechanisms at play. While this study concentrates on accounting profitability and Tobin's Q, investor recognition may influence additional dimensions of firm behavior and market outcomes. Investment intensity, financing costs, payout policies, stock return volatility and direct measures of the cost of capital represent natural extensions. In particular, estimating the relationship between breadth and implied or realized cost of capital measures would offer a more direct empirical link to participation-based asset pricing theories and could help reconcile the stronger valuation effects with the weaker evidence on operating performance.

Heterogeneity across firms constitutes another dimension along which the analysis could be deepened. The evidence presented here suggests that the association between recognition and valuation may exhibit non-linear patterns. Future research could examine whether these effects are amplified among firms characterized by high information asymmetry, limited analyst coverage, smaller scale or greater opacity. Interactions between breadth and such characteristics would clarify whether participation effects are particularly relevant in environments where informational frictions are more pronounced.

Finally, extending the framework beyond a single institutional context would strengthen external validity. Financial markets differ markedly in terms of investor protection, disclosure standards, trading infrastructure and participation constraints. Applying the same empirical design to other countries or regulatory environments would shed light on whether the role of recognition in valuation is universal or contingent on institutional features.

Taken together, these avenues suggest that the present findings should be viewed as a structured empirical foundation rather than as a final answer. By documenting systematic within-firm associations and distinguishing between performance and pricing channels, the thesis provides a coherent starting point for more refined identification strategies, richer measures of recognition and broader analyses of corporate and market outcomes.

Bibliography

[1] R. C. Merton, “A Simple Model of Capital Market Equilibrium with Incomplete Information,” The Journal of Finance, vol. 42, no. 3, pp. 483–510, 1987. [Online]. Available: https://doi.org/10.1111/j.1540-6261.1987.tb04565.x

[2] D. Easley and M. O’Hara, “Information and the Cost of Capital,” The Journal of Finance, vol. 59, no. 4, pp. 1553–1583, 2004. [Online]. Available: https://doi.org/10.1111/j.1540-6261.2004.00672.x

[3] R. A. Lambert, C. Leuz and R. E. Verrecchia, “Accounting Information, Disclosure and the Cost of Capital,” Journal of Accounting Research, vol. 45, no. 2, pp. 385–420, 2007.

[4] Y. Amihud, “Illiquidity and Stock Returns: Cross-Section and Time-Series Effects,” Journal of Financial Markets, vol. 5, no. 1, pp. 31–56, 2002.

[5] V. V. Acharya and L. H. Pedersen, “Asset Pricing with Liquidity Risk,” Journal of Financial Economics, vol. 77, no. 2, pp. 375–410, 2005.

[6] A. Arbel, “Generic Stocks: An Old Product in a New Package,” The Journal of Portfolio Management, vol. 8, no. 4, pp. 4–13, 1982.

[7] A. Arbel, S. Carvell and P. Strebel, “Giraffes, Institutions and Neglected Firms,” Financial Analysts Journal, vol. 39, no. 3, pp. 57–63, 1983.

[8] J. Chen, H. Hong and J. C. Stein, “Breadth of Ownership and Stock Returns,” Journal of Financial Economics, vol. 66, no. 2–3, pp. 171–205, 2002.

[9] R. Lehavy and R. G. Sloan, “Investor Recognition and Stock Returns,” Review of Accounting Studies, vol. 13, no. 2–3, pp. 327–361, 2008.

[10] H. Hong and M. Kacperczyk, “The Price of Sin: The Effects of Social Norms on Markets,” Journal of Financial Economics, vol. 93, no. 1, pp. 15–36, 2009.

[11] A. Bodnaruk and P. Östberg, “The Shareholder Base and Payout Policy,” Journal of Financial and Quantitative Analysis, vol. 48, no. 6, pp. 2087–2122, 2013.

[12] D. Lou, “Attracting Investor Attention through Advertising,” The Review of Financial Studies, vol. 27, no. 6, pp. 1797–1829, 2014.

[13] B. M. Green and R. Jame, “Company Name Fluency, Investor Recognition and Firm Value,” Journal of Financial Economics, vol. 109, no. 3, pp. 813–834, 2013.

[14] H. Demsetz and B. Villalonga, “Ownership Structure and Corporate Performance,” Journal of Corporate Finance, vol. 7, no. 3, pp. 209–233, 2001.

[14] E. F. Fama and K. R. French, “Common Risk Factors in the Returns on Stocks and Bonds,” Journal of Financial Economics, vol. 33, no. 1, pp. 3–56, 1993.

[15] C. Wooldridge, Econometric Analysis of Cross Section and Panel Data, MIT Press, 2010.

List of Figures

List of Tables

Acknowledgment

Con questo lavoro si conclude un percorso che non è stato solo universitario, ma più in generale di crescita personale. Un percorso lungo, a tratti complesso, ma ricco di soddisfazioni. È un cammino che arriva a una fine chiara, lasciando però aperte molte strade diverse e tutte ugualmente affascinanti.

Oggi non è possibile, e forse sarebbe persino controproducente, immaginare una strada unica e definita davanti a sé. Ed è proprio questa incertezza a rendere questo passaggio ancora più stimolante.

Le uniche vere certezze sono che, con i propri valori, con la consapevolezza, con l'impegno e la giusta serenità, ogni percorso può essere affrontato e portato avanti.
Questa consapevolezza, però, nasce anche dall'aver già camminato: dall'esperienza di sapere che, passo dopo passo, ogni strada può essere percorsa dall'inizio alla fine.

Allo stesso tempo, ignorare il ruolo che la fortuna e la vita possono avere nel facilitare o nell'ostacolare qualsiasi cammino significherebbe non riconoscere la realtà.
C'è chi percorre la propria strada con pesi alle caviglie, chi trova ogni buca già colmata, chi viaggia da solo e chi, invece, è accompagnato da persone pronte a sostenere nei tratti in salita.

Non riconoscere la fortuna che ho avuto non sarebbe corretto nei confronti di chi, in modi diversi, mi ha aiutato a portare a termine questo percorso e continuerà a farlo anche in quelli che verranno.

A mamma e papà, grazie per avermi permesso di affrontare questo cammino con serenità, senza altre preoccupazioni, e per avermi dato tutto ciò che mi serviva per viverlo al meglio. Grazie per il vostro supporto incondizionato: questo traguardo è dedicato a voi.

Alla mia famiglia, Pietro, i nonni, gli zii, i cugini, grazie per il vostro supporto, per esserci stati e perché so che, quando davvero conterà, sarete presenti.

A Federica, grazie per la fiducia che riponi in me e per l'amore con cui mi accompagni ogni giorno. Grazie per esserci stata in questo percorso, nei momenti di entusiasmo come in quelli di maggiore incertezza, e per avermi sostenuto con la tua presenza, le tue parole e il tuo esempio. Sei stata una fonte costante di forza e di ispirazione, e sapere di poterti avere accanto ha reso questo cammino più leggero e significativo.

Ai miei amici, grazie per esserci stati lungo questo percorso, per i momenti di leggerezza che hanno dato equilibrio anche a quelli più impegnativi e per il sostegno, spesso silenzioso ma sempre presente. Grazie per aver condiviso successi, dubbi e sacrifici, e per aver reso questo cammino non solo un obiettivo da raggiungere, ma anni indimenticabili vissuti insieme.