The Macroeconomics of Data: Scale, Product Choice, and Pricing in the Information Age*

Vladimir Asriyan Alexandre Kohlhas

March, 2025

Abstract

We document a substantial rise in the accuracy of U.S. firms' expectations since the early 2000s, closely linked to firm-size dynamics and consistent with major advances in data-processing technologies. To study the macroeconomic implications, we develop a model of information production, in which information enables firms to optimize their *scale*, *product choice*, and *pricing strategies.* While information enhances the efficiency of resource allocation, it also facilitates price discrimination. The laissez-faire equilibrium is inefficient, warrants corrective policy interventions, and advances in data-processing technologies have ambiguous effects on social welfare. Calibrating our model to U.S. firm-level data, we find that data-processing advances have significantly increased TFP over the past two decades (5.3-6.7%), primarily by helping firms determine their optimal scale. Yet, the welfare benefits of these improvements have been modest (0.1-2.1%). Restricting data use, especially by large firms, could trigger larger welfare gains.

JEL codes: E10, E60, C53, D83, D84

Keywords: data economy, expectations, information frictions, product choice, price discrimination, rent extraction, misallocation, optimal policy, data regulation

---

*First draft: April 2024. Asriyan: CREI, ICREA, BSE, UPF (email: vasriyan@crei.cat). Kohlhas: University of Oxford (email: alexandre.kohlhas@economics.ox.ac.uk). We are thankful to Isaac Baley, Timo Boppart, Joel David, Jan Eeckhout, William Fuchs, Guido Lorenzoni, Alberto Martin, Guillermo Ordonez, Guangyu Pei, Giacomo Ponzetto, Edouard Schaal, Laura Veldkamp, Venky Venkateswaran, Jaume Ventura, Joshua Weiss, and seminar and conference participants at CREI, UCSD, USC, HKU, CUHK, HKUST, ESSEC, Oxford University, Goethe University, NHH, University of Luxembourg, NBER SI Micro Data and Macro Models, UT Austin Macro/International/Finance Conference, BSE Summer Forum 2024, SED Buenos Aires, the Annual Meetings of the Armenian Economic Association 2023 and of the AEA 2024 Meetings for their feedback and comments. Financial support from the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-00915-S) and Handelsbank Stiftelsen is gratefully acknowledged. Erfan Ghofrani provided excellent research assistance.

1 Introduction

Advances in data-processing technologies are widely believed to have the potential to transform most economic interactions. Consistent with this view, the past two decades have seen a substantial rise in the share of firms that systematically use data to inform their economic decision-making. A simple estimate based on survey data from a sample of medium-to-large firms indicates that this share has more than doubled in the past ten years alone (Mckinsey and Company, 2023). As of today, approximately 40-75 percent of manufacturing firms employ some form of *data-driven decision-making* (Brynjolfsson and McElheran, 2024).[1]

Despite the substantial rise in data use by firms, the macroeconomic consequences of these developments are, nevertheless, not fully understood. Has the apparent rise in data use led to an improved allocation of resources across and within firms, substantially enough to affect economy-wide dynamics? Or, have these developments merely allowed firms to extract ever larger rents from consumers, diluting their potential welfare benefits? Answers to these questions are important not only to understand the developments of the recent past, but also to design effective regulatory frameworks that will govern the data economy of the future.

In this paper, we explore answers to these questions. Throughout, we view *data as information* that helps firms predict economic fundamentals (Baley and Veldkamp, 2025). The rationale for this view is both *practical*—changes in information can be proxied by changes in the accuracy of firms' expectations—and *conceptual*—since the seminal work of Lucas (1972), the role of information has been shown central for the understanding of macroeconomic dynamics.[2] Using micro-level data on firms' expectations, we document a substantial increase in the accuracy of U.S. firms' expectations since the early 2000s. This improvement is closely linked to firm-size dynamics, and it aligns with a significant rise in firms' data use.

Our main contribution is to develop a *unifying*, quantitative-theoretical framework to explore the macroeconomic consequences of these developments. In our model, information production helps firms optimize their *scale*, *product choice*, and *pricing strategies*. These channels are thought to be among the most prominent ways by which data and information technologies have transformed firm behavior over the past decades.[3] In our framework, information has a dual role. On the one hand, it boosts the allocative efficiency of resources across and within firms. But, on the other hand, it also facilitates price discrimination, which may be socially detrimental. The net effect of advances in data-processing technologies depends

[1]See Brynjolfsson and McElheran (2016) and Goldfarb *et al.* (2015) for overviews on adoption rates.

[2]See, e.g., Woodford (2002), Sims (2003), Blanchard *et al.* (2013), Chahrour and Jurado (2018), Angeletos and Lian (2016), Angeletos *et al.* (2021), among others.

[3]Feng *et al.* (2020) and Babina *et al.* (2024) discuss changes to product design, while Adams *et al.* (2024) documents changes in firms' pricing; Zolas *et al.* (2021) and Jin and McElheran (2024) discuss changes in inputs associated with cloud computing and AI; lastly, O'Neill (2023) presents several business cases.

on the balance of those two forces. We show that the laissez-faire equilibrium of our economy is inefficient, warranting corrective policy regulation. Across a wide range of parameters, our theory lends support to a common concern that firms—particularly large ones—produce and use data excessively (e.g., European Commission Report, 2020).

The aggregate effect of advances in data-processing technologies, as well as the design of optimal policies, depends on several key model parameters. To identify them, we employ a calibration strategy that leverages U.S. firm-level data, allowing us to establish sharp bounds on the costs and benefits of technological advancements. Our estimates indicate that data-processing improvements have meaningfully increased total factor productivity (TFP) over the past two decades (5.3-6.7%), primarily by enabling firms to better optimize their scale of operations. Yet, the corresponding welfare gains have been modest (0.1-2.1%), as much of the benefits have been offset by excessive information production. Our findings, thus, underscore a crucial role that corrective data regulation can play in the modern information age.

To empirically motivate our work, we use micro-level data on managerial forecasts from the I/B/E/S-Compustat panel. Using the merged firm-level data set, we document a systematic rise in the accuracy of US firms' expectations over time. Over the past two decades, firms' expectations of one-year-ahead revenue have witnessed accuracy increases between 24-41%, with similar developments across most sectors and for other outcome variables (e.g., profits and capital expenditures). Crucially, we show that this rise in the accuracy of firms' expectations persists even after controlling for the volatility of firm-level and economy-wide shocks, implying that firms have become substantially more informed since the early 2000s.

The increased accuracy of firms' expectations hints at changes in firm behavior or characteristics over time. We therefore study the cross-sectional determinants of firms' accuracy in the micro data. Across a range of estimation strategies and controls, we find robust evidence that larger firms are, all else equal, systematically more accurate than smaller ones. The accuracy of firms' expectations exhibits substantial heterogeneity across the size distribution.[4] Through a simple simulation exercise, we show that the bulk of the aggregate improvement in the accuracy of firms' expectations over the past two decades can be accounted for by developments along the size-accuracy relationship. In contrast, changes in sectoral composition, or changes in the volatility of outcome variables, appear to play only a minor role.

We conduct a comprehensive set of robustness checks, demonstrating that our results hold across alternative measures, specifications, and with the inclusion of firm-level controls. Comfortingly, we show that the positive relationship between firm size and accuracy extends to an independent data source—the Duke-Richmond Fed CFO Survey—which captures firms' expectations of macroeconomic outcomes beyond their direct control. This finding further

[4] See also Senga (2018), Tanaka *et al.* (2020), Chen *et al.* (2023), and Chen *et al.* (2024).

bolsters the case that the documented size-accuracy relationship reflects differences in information rather than differences in the nature (or predictability) of the outcome variable.

To explore the macroeconomic implications of our findings, we develop a macroeconomic model of information production. We consider an economy populated by heterogeneous consumers with preferences over differentiated varieties supplied by monopolistic firms à la Dixit and Stiglitz (1977). The central friction in our economy is that firms must allocate inputs and price products under uncertainty about technologies and preferences.

In particular, firms in our economy are uncertain about their optimal (i) *scale of operations* and (ii) *product choice*. The former arises because production quantities depend not only on input choices but also on firms' ex-ante unknown productivity state. The latter instead occurs because firms want to customize varieties to best match ex-ante unknown consumer tastes. Finally, the presence of unknown consumer tastes further implies that firms also face uncertainty about their (iii) *pricing strategy*.

A key feature of our framework is that firms can mitigate such uncertainty through costly information production. By expending scarce resources, firms can obtain informative signals about the states of nature governing *demand* and *productivity*. We model advancements in data processing as stemming from either a reduction in the cost of obtaining these signals or an increase in their accuracy. This stylized approach captures key technological developments, such as, e.g., declining computing costs and improvements in processing speeds.[5]

Our analysis starts with a pedagogical *baseline economy*, where firms are exogenously constrained from price discrimination and their use of information is efficient. In equilibrium, firms select into information production based on ex-ante size differences. Information production facilitates the optimal determination of *scale* and *product choice*, driving growth in firm profits, size, and efficiency, as reflected in firm-level measures of total factor productivity (TFP) and dispersion of marginal revenue products. In general equilibrium, information production enhances aggregate TFP both directly—by improving firm-level efficiency—and indirectly—by reallocating inputs toward better-informed firms. Under this benign view, we show that advances in data-processing technologies unambiguously enhance social welfare.

We next turn to our main setting—the *rent-extracting economy*—where firms also optimize their *pricing strategies* to engage in price discrimination. In particular, we allow firms to design optimal trading mechanisms to trade their varieties with consumers. Even though, in equilibrium, firm-level outcomes in the rent-extracting economy appear similar to those in our baseline setting, there is a critical difference. Due to firms' uncertainty about consumer tastes, price discrimination introduces a conflict between profit and social surplus maximization, lead-

[5] See, e.g., Nordhaus (2008), Coyle and Hampton (2024), and Gill *et al.* (2024) for evidence of substantial declines in computing costs and improvements in processing speeds over the past two decades.

ing firms to distort allocations to extract ever larger rents from consumers. In turn, information production—through its effect on firms' uncertainty about consumer tastes—affects the severity of this conflict. We show that the laissez-faire equilibrium is, as a result, inefficient, and the welfare effects of technological advances in data processing become ambiguous, providing a rationale for regulatory intervention.

We characterize the optimal corrective policy and identify the scenarios in which firms engage in *excessive* information production. This occurs when *rent extraction is severe*, a parameter region in which firms' ability to extract rents from consumers grows faster with information production than their contribution to social surplus. This finding lends support to the recently discussed and implemented policies, such as the General Data Protection Regulation in the EU to limit firms' ability to collect, store, and exploit consumer data in ways that foster discriminatory behavior.[6] Yet, we also provide conditions under which information production is *insufficient*, providing caution that limiting data use may not always be desirable. When information production is excessive, we further show that optimal corrective policy in addition targets *firm-size concentration*, as the distortions introduced through information production intensify with firm scale.

Our framework emphasizes that the economy-wide effects of advances in data-processing technologies and the design of optimal corrective policies depend critically on several key structural parameters. To assess the consequences of improvements in data-processing technologies for the U.S. economy over the past two decades, we first validate our theoretical framework and then quantify the implications of these changes using U.S. firm-level data.

We proceed in two steps. First, we show that, consistent with the predictions of our model, more informed firms display higher and less dispersed (revenue-based measures of) total factor productivity, less dispersed marginal revenue products of labor and capital, are more profitable, and grow faster and larger. Second, we estimate the potential benefits of improvements in data-processing. Because a range of rent-extracting equilibria—differing in the severity of distortions from price discrimination—are consistent with the *same* observed firm-level data, we construct both best- and worst-case scenarios for the effects of these improvements.

For plausible parameter values, based on the observed increase in firms' accuracy over our sample period, our main estimates suggest that aggregate TFP (household welfare) in the U.S. would have been 5.3–6.7% (0.1–2.1%) lower in 2022 in the absence of the technological improvements. We further find that the majority of these TFP gains—approximately two-thirds to three-quarters—stem from firms enhanced ability to optimize their *scale* of operations. The benefits from improved *product design*, by contrast, account for at most one-third. Despite these productivity increases, our results, however, also worryingly indicate that most

[6]For details of the GDPR in the EU, see https://gdpr.eu.

of the welfare gains over the past two decades may have been offset by excessive information production, arising from firms' enhanced ability to *price discriminate* consumers.

Our quantitative findings, thus, underscore the potentially central role that data regulation can play in ensuring that advances in data-processing technologies translate into welfare improvements. In particular, we show that even a simple tax on information production can substantially improve welfare outcomes—by more closely aligning TFP gains with household welfare—even under the worst-case parametric scenario.

Finally, we conduct two exercises that further refine our quantitative results. First, using the empirical evidence from Bornstein and Peter (2024), who measure the extent of price discrimination with scanner data from the retail sector, we provide a first-pass estimate of where the US economy locates itself within our calibrated range. Matching their estimates shows that the US economy appears close to our calibrated worst-case scenario. Clearly, the U.S. economy is comprised of sectors beyond the retail sector for which product-level data is not immediately available. Nevertheless, our estimates do provide an initial (albeit imperfect) gauge of the potential detrimental effects that information production may have through price discrimination. Second, we extend our framework to accommodate capital and variety accumulation. Both features should, in principle, amplify the economy-wide effects of advances in data processing. We show that, despite modest amplification, our main results remain robust—TFP advances are estimated to have been meaningful but the welfare benefits have been comparatively more subdued.

Related Literature. In addition to the above-cited work, our paper is related to several strands of research, which we review in order of proximity.

First, our work builds on the rapidly expanding macroeconomic literature on the data economy (e.g., Begenau *et al.*, 2018; Farboodi and Veldkamp, 2020, 2024; Eeckhout and Veldkamp, 2023). Baley and Veldkamp (2025) provides a comprehensive overview. Our study departs from this growing body of research by developing a *unifying* macroeconomic framework, disciplined with firm-level data, in which information shapes firm behavior through three distinct channels that have been particularly salient since the early 2000s. Our framework, specifically, allows us to decompose and quantify the consequences that advances in data-processing have had on macroeconomic outcomes through separate channels—and to address central normative questions about corrective data regulation. Our work is, in particular, closely related to the complementary contributions of David *et al.* (2016) and David and Venkateswaran (2019), who develop a methodology to measure firms' information use. On the empirical front, their approach primarily infers firms' information use indirectly from stock prices, whereas we take a more direct approach that exploits data on managers' expectations. On the theoretical front, their work centers on the role that information has on factor misallocation—akin to our *scale* *channel* of information—whereas we stress that information also affects firms' product choice and pricing strategies, and we further shed light on normative questions.

Second, our work naturally relates to an extensive literature in microeconomics on price discrimination, as well as its recent macroeconomic application by Bornstein and Peter (2024), who examine how non-linear pricing affects economy-wide markups and factor misallocation.[7] We contribute to this literature by analyzing the general equilibrium effects of price discrimination in settings in which it interacts with firms' incentives to produce information. In turn, our main normative results—and their implications for policy—build on the broad idea that more public information may be socially undesirable in settings with asymmetric information, as it may exacerbate the distortions that are already present.[8] In our setting, this logic appears in stark form, since information production—through its effect on product choice—induces a first-order stochastic dominance shift in firms' posterior beliefs about consumer types, rather than merely increasing the dispersion of firms' posteriors.

Third, our work relates to the growing literature that documents a rise in market power and studies the welfare costs stemming from it (e.g., De Loecker *et al.*, 2020; Edmond *et al.*, 2023; Boar and Midrigan, 2024; Eeckhout *et al.*, 2024). A common theme in this work is that large firms—those that have higher market power and markups—are inefficiently small, precisely because they constrain production to raise prices. By contrast, in our framework, monopolistic firms do not need to constrain production to extract consumer surplus; they can do so through price discrimination instead. Indeed, in the parameter region that is consistent with US micro data, we find that large firms are indeed *too large*: optimal corrective policy—besides discouraging firms from producing information—also reduces firm-size concentration.

Finally, our work contributes to the classic literature on the macroeconomic effects of information frictions, tracing back to Lucas (1972). Notable contributions in this area include Woodford (2002), Mankiw and Reis (2002), Angeletos and Pavan (2007), Ordonez (2009), Lorenzoni (2009), Maćkowiak and Wiederholt (2009), and Angeletos *et al.* (2021), among many others. We emphasize the role that firms' strategic information choices have through pricing and resource-allocation channels in amplifying the aggregate consequences of imperfect information. Our contribution, in this context, is to provide the first, to our knowledge, decomposition and quantification of the effects that recent advances in data-processing technologies, and the associated fall in information frictions, have had on macroeconomic outcomes.

[7] See Varian (1989) for a survey of the classic literature on price discrimination and Kehoe *et al.* (2018) for recent research on how big data may enhance firms' ability to discriminate among consumers. Nevo and Wong (2019) and Baker *et al.* (2020) provide additional evidence on the non-linearity of pricing.

[8] While this idea has primarily been explored in trading environments with adverse selection (e.g., Malherbe, 2012; Gorton and Ordonez, 2014; Asriyan *et al.*, 2017), the contemporaneous contributions of Farboodi *et al.* (2025) and Asriyan *et al.* (2025) explore it further in the related context of a price-discriminating monopolist facing privately informed consumers.

2 Motivating Evidence

We present new evidence on the accuracy of firm expectations over time and across the firm-size distribution. To start, we use micro data on firm expectations from the I/B/E/S managerial guidance database. The I/B/E/S data set contains, for an individual firm-year, a manager's publicly stated expectation for their firm's revenue, profits, and other performance variables for the upcoming year. We exploit one-year-ahead forecasts made concurrently with the release of the previous year's financials. We link the I/B/E/S database to Compustat, which provides detailed data on firms' financials. The merged I/B/E/S-Compustat sample spans the period 2002-2022. Appendix A.1 provides more information on the sample construction.

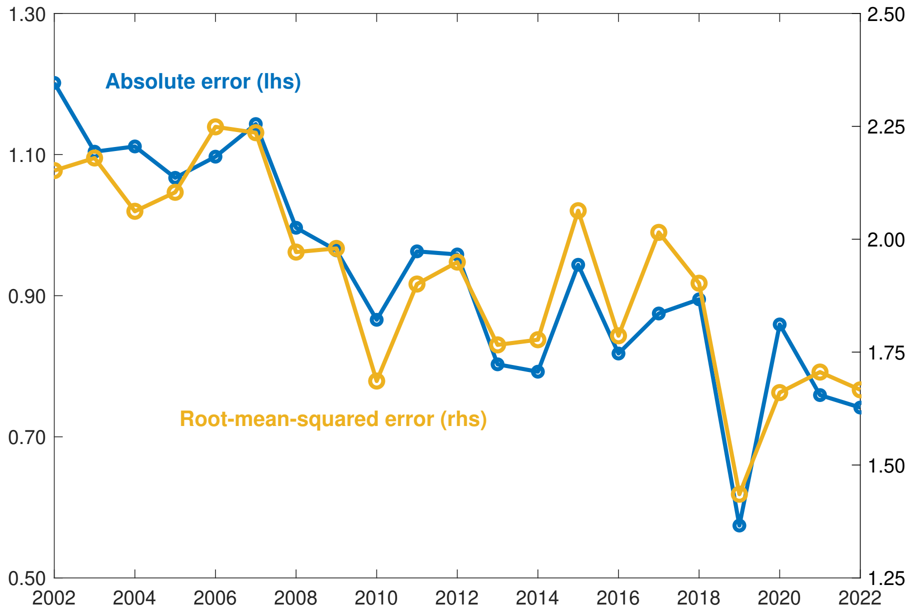
We begin by documenting changes in the accuracy of firms' expectations over time.[9] We focus on one-year-ahead revenue errors, defined as the realization minus its predicted value. A negative error thus corresponds to an over-estimate of future revenue. Figure 1 depicts the evolution of the average accuracy of firms' expectations over time. All else equal, over the past two decades, firms' expectations have improved markedly, with an average improvement in one-year-ahead accuracy between 24-41%, depending on the accuracy measure used. Firms have, on average, become substantially more accurate over time. Table A.3 in the Appendix confirms this initial finding using regressions of individual errors on time.

A natural candidate explanation for firms' increased accuracy is a decline in economy-wide volatility, such as that which occurred during the Great Moderation (e.g., Arias *et al.*, 2007). However, as Table A.4 in the Appendix shows, similar results hold after partialling out *sector*$\times$*time* fixed effects from firms' errors. Consistent with most of the uncertainty faced by firms being due to *firm-specific* rather than *aggregate* factors (Lucas, 1977), the lion's share of accuracy increases here arises from improvements in firms' expectations about firm-specific outcomes. Combined with the observation that idiosyncratic volatility has, if anything, increased over our sample period (e.g., Bloom *et al.*, 2018 and Section 6.3), we conclude that *firms must have become substantially more informed since the early 2000s.*

The increased informativeness of firms is suggestive of changes in firm behavior or characteristics. The cross-section of firms can, as such, be revealing about the drivers behind the overall improvement over time. Because of a sizable fixed-cost component to the processing of information (e.g., Brynjolfsson *et al.*, 2008; Bloom *et al.*, 2019), it is natural to ask whether there is a relationship between firm size and the accuracy of firms' expectations. To investigate this question, Panel (a) in Figure 2 plots the difference between the average accuracy of one-

[9] In Appendix A.3, we conduct several data-validation tests, akin to those in Tanaka *et al.* (2020), Chen *et al.* (2023), and Chen *et al.* (2024). In particular, we show that firms' expectations are (close to) unbiased, feature a symmetric error distribution, that more (less) optimistic firms increase (decrease) their use of factors of production, and that positive (negative) surprises result in more (less) inputs being employed subsequently.

Figure 1: Time Evolution of Revenue Accuracy



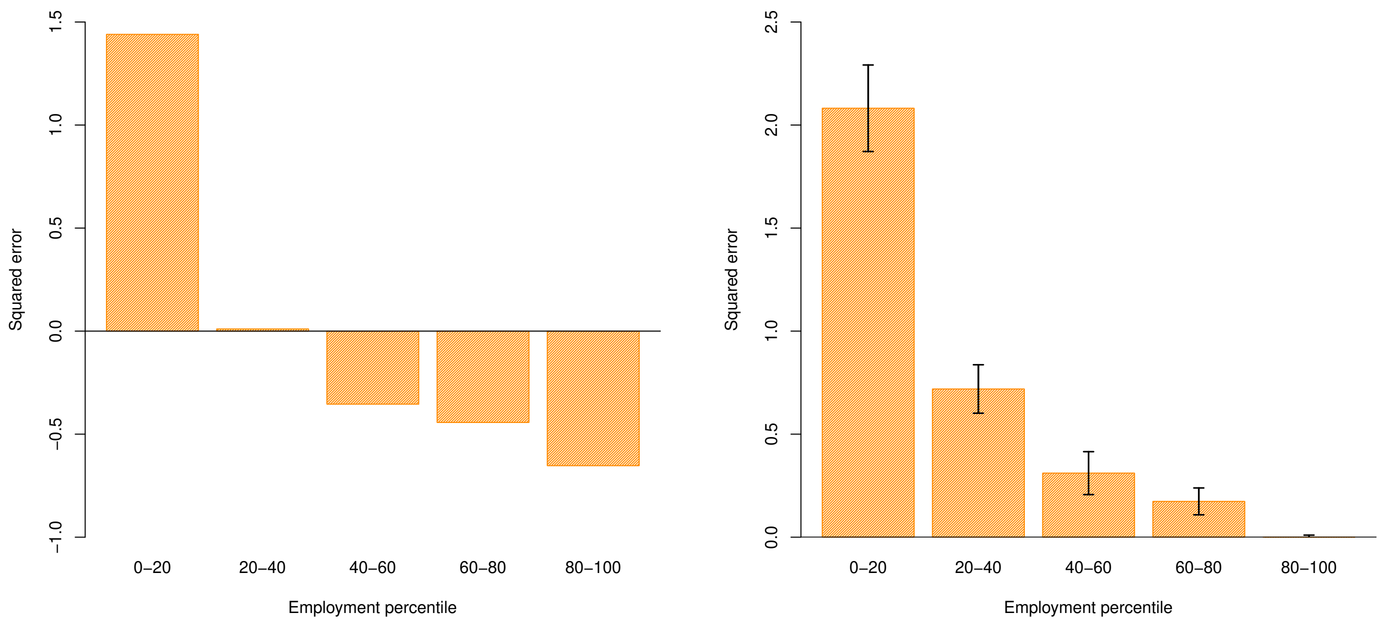
Note: Data from the I/B/E/S-Compustat sample. The panel shows the mean absolute error of one-year-ahead revenue forecasts on the left vertical axis, and the root mean-squared-error on the right axis. Revenue errors are scaled by a firm's tangible capital stock. Table A.3 in the Appendix shows the associated regression results.

year-ahead revenue expectations within size (employment) quintiles and the overall average taken across all firm sizes. The results show a marked, monotone relationship between firm size and the accuracy of firm expectations in the raw data. Larger firms have more accurate expectations—with an especially pronounced difference when moving away from the bottom quintile of the size distribution.

The relationship in Panel (a) may, nevertheless, be contaminated by other factors, such as differences in the volatility of the outcome variable (i.e., revenues) across firm size or learning with age, which may be correlated with firm size for other reasons. To address this concern, Panel (b) in Figure 2 plots estimates from a regression of the accuracy of firm expectations on firm size, controlling for firm characteristics and time and sector fixed effects. Table I explores the effects of alternative controls and estimation methods, crucially, controlling for changes in the volatility of firm revenue and productivity over time. Our results confirm the findings from the raw data. The accuracy of expectations improves with size, even after controlling for characteristics. *Larger firms are, all else equal, more informed than smaller ones.*

Table A.5 in the Appendix shows that the documented size-accuracy relationship further extends to alternative data sets—in this case, the Duke-Richmond Fed CFO survey (Graham *et al.*, 2023; Appendix A.2)—which surveys firms' expectations of macroeconomic outcomes over which firms have no control. The latter is important as it further bolsters the case

Figure 2: Revenue Expectations Across the Size Distribution



Panel (a): Relative Accuracy

Panel (b): Coefficient on Size

Note: Panel (a) plots the difference between the squared error of one-year-ahead log-revenue expectations from the I/B/E/S-Compustat merger within size (employment) quintiles and the overall average taken across all size levels. Panel (b) plots the coefficient estimates on size from a regression of the squared value of individual errors on the size quintile the firm belongs to, controlling for firm characteristics (Table I Column 3). Revenue errors are scaled by a firm's tangible capital stock and normalized by their mean value in the sample. Whisker-intervals are one-standard deviation robust (clustered) confidence bounds. Sample: 2002-2022.

that accuracy improves with firm size due to improvements in firms' information rather than differences in the outcome variable across the size distribution.

The magnitude of the estimated effect of size in Table I is, moreover, considerable. Increasing the size of a firm by one quintile, for example, decreases the associated squared error by 47 percent of its average value (Column 1 in Table I). As documented in Figure A.2 and Tables A.3-A.4 in the Appendix, over the past two decades, firm size has increased drastically—with, for example, a close to doubling in the share of firms with employment exceeding the 80th percentile of the 2002-employment distribution.[10] Crucially, after controlling for this evolution in the firm-size distribution, the effect of time itself becomes insignificant (Column 2 in Table I). This suggest that accuracy increases unrelated to firm size have played only a minor role over this period. Indeed, Figure A.4 in the Appendix, using estimates from Table I, shows that the change in the firm-size distribution alone can account for approximately 80 percent of the observed increase in accuracy. Clearly, these estimates cannot be interpreted as causal,

[10]See also, e.g., Autor *et al.* (2020), Kwon *et al.* (2023), among others, and Appendix Figure A.3.

Table I: Revenue Expectations, Firm Size, and Time

	Squared (log.) revenue errors					
	(1)	(2)	(3)	(4)	(5)	(6)
Firm size	-0.468^{***}	-0.454^{***}		-0.416^{***}	-0.290**	-0.430^{*}
	(0.055)	(0.052)		(0.122)	(0.124)	(0.226)
Firm size (1)			2.082^{***}			
			(0.210)			
Firm size (2)			0.719^{***}			
			(0.118)			
Firm size (3)			0.311^{***}			
			(0.104)			
Firm size (4)			0.174^{***}			
			(0.065)			
Time		0.007				
		(0.007)				
Firm age		-0.063^{**}	-0.072^{**}	0.118^{**}	0.187^{**}	-0.117
		(0.032)	(0.033)	(0.057)	(0.066)	(0.093)
Log rev. volatility				-0.030		
				(0.027)		
Log TFP. volatility					1.095	
					(0.744)	
Observations	12,489	12,489	12,489	6,809	5,637	2,570
Sector FE	✓	✓	✓	×	×	×
Firm FE	×	×	×	✓	✓	✓
Time FE	×	×	✓	✓	✓	✓
Panel GMM	×	×	×	×	×	✓
F statistic	3.911^{***}	3.922^{***}	4.322^{***}	9.295^{***}	8.005^{***}	NA.

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Column (1) shows estimates of the squared value of one-year-ahead log-revenue errors on firm size (employment) and sector (NAICS-4) fixed effects. Firm size is measured by the quintile the firm's employment is at time t relative to the 2002-employment distribution. Column (2) adds time and age controls. Column (3) allows for separate coefficients on size-levels (estimates are relative to the largest firms, those in the 80-100th percentile) and time fixed effects. Column (4) allows for firm fixed effects instead of sector fixed effects and controls for the individual four-year-rolling revenue volatility. Column (5) instead controls for individual four-year-rolling TFP volatility (Appendix A.2). Finally, Column (6) provides Arellano-Bond estimates. Revenue errors are scaled by firm capital and normalized by the overall average absolute (squared) error. The top and bottom 1 percent of errors have been removed. Robust (clustered) standard errors in parentheses. Sample: 2002-2022. *p<0.1,**p<0.05,***p<0.01

as firm size and the accuracy of expectations are determined jointly (e.g., Section 4). Yet, our results demonstrate that, *over the past two decades, an intimate relationship has existed between the accuracy of firms' expectations, firms' information, and firms' size.*

We obtain similar estimates to those in Figures 1 and 2 for alternative measures of size and accuracy. Table A.6 and Figure A.5 show similar estimates when proxying size with firm assets instead of employment. Table A.7-A.8 documents that an alternative measure of accuracy (e.g., the absolute error) likewise monotonically increases with firm size, irrespective of whether size is measured by employment or assets. We further find that across all-but-one sector accuracy has improved over time and increases with size (Figure A.6).

Table A.9-A.12 in the Appendix contains additional analysis and further robustness checks. We document that our findings extend to firm expectations of other variables than revenue (profits and capex), and to different transformations of firm revenue. We also show that the accuracy of firm expectations improves after large acquisitions of other firms, lending support to the notion that there are increasing returns to information; and extends to different assumptions about sectoral and time fixed effects. Finally, consistent with improvements in firm information driving the observed patterns, Table A.12 documents that firms who have a larger stock of *acquired intangibles* (Chiavari and Goraya, 2023), which includes business software and expenditures on data processing, among others, feature more accurate expectations.

In summary, the results in this section show that a notable improvement in firms' accuracy over the past two decades has coincided with a substantial increase in firm size, reflecting the fact that larger firms have more accurate expectations. These findings clearly reject the common-expectations assumption frequently employed in full-information rational expectations models of firm behavior. Motivated by the findings in this section, we next develop a macroeconomic model of information production, which we then use to study the aggregate implications of rising firm-level accuracy.

3 The Baseline Economy

We start by developing our *baseline economy*, a central feature of which is that firms are uncertain about the optimal *scale* of their operations and their optimal *product choice*.

3.1 Environment

Preferences, Endowments, and Technology. We consider an economy populated by a unit mass of households, indexed by $i \in [0, 1]$, with CES-preferences over consumption:

$$\mathcal{U}_i = C_i = \left(\int_0^1 \left(\delta_{ij} \cdot c_{ij} \right)^{\frac{\theta-1}{\theta}} \cdot dj \right)^{\frac{\theta}{\theta-1}}, \tag{1}$$

where c_{ij} is household i's consumption of variety $j \in [0,1]$, $\delta_{ij} \in \{\delta_H, \delta_L\}$ is the household's taste shifter with $\delta_H > \delta_L > 0$, and $\theta > 1$ is the elasticity of substitution across varieties. Each household is endowed with N units of labor, the economy's only factor of production.

The taste shifter δ_{ij} is distributed identically across households and independently across varieties. The share of δ_H-households for variety j, denoted by $\gamma_j \in [0,1]$, is assumed to be random and to depend on *the type*, x_j, of the variety available in the market. In particular, households may prefer either the red- or the blue-type variety, where $\omega_j \in \{\text{red}, \text{blue}\}$ denotes the *demand state* with $\mathbb{P}(\omega_j = \text{red}) = \mathbb{P}(\omega_j = \text{blue}) = 1/2$. The households' demand for variety j is assumed to be higher when it is customized to households' tastes:

$$\gamma_j = \gamma(x_j|\omega_j) \text{ where } \gamma(x_j = \omega_j|\omega_j) = \bar{\gamma} > \underline{\gamma} = \gamma(x_j \neq \omega_j|\omega_j) \quad \forall j,\ \omega_j, \tag{2}$$

i.e., the share of δ_H-households is higher when the variety-type matches the demand state.[11]

Each variety is produced by a monopolistically competitive firm, owned by households. Firm j chooses the type, x_j, of its variety to produce at no cost, and produces the quantity y_j of it in accordance with the linear production technology:

$$y_j = A_j \cdot n_j, \tag{3}$$

where n_j are the units of labor employed by the firm and A_j denotes the firm's productivity.[12] Firm productivity is comprised of two components:

$$a_j \equiv \log(A_j) = \mu_j + \upsilon_j, \tag{4}$$

where $\mu_j \sim \mathcal{N}\left(0, \tau_\mu^{-1}\right)$ and $\upsilon_j \sim \mathcal{N}(0, \tau_a^{-1})$ are distributed independently across firms and of each other, and where τ_μ^{-1} and τ_a^{-1} capture the dispersions of each component.

Uncertainty and Information. A central friction in our economy is that firms are uncertain about their productivity and demand when making production choices. When choosing x_j and n_j, a firm knows its mean-productivity level, μ_j, but does not know the innovation to productivity, υ_j, nor the composition of its demand, ω_j.[13] As a result, the component μ_j is a source of *ex-ante heterogeneity* among firms, whereas υ_j and ω_j are sources of *ex-post uncertainty* that the firms may want to overcome through information production.[14]

[11] In Appendix B, we provide a simple micro-foundation for such consumer tastes, in which individual preferences over red and blue variety-types are modeled explicitly.

[12] Note that productivity, A_j, can equivalently be embedded into consumer preferences by supposing that the taste shifter of consumer i for variety j is $A_j \cdot \delta_{ij}$ instead of δ_{ij}.

[13] In our baseline framework, knowing the realizations of υ_j and ω_j is sufficient for all firm decisions.

[14] Ex-ante heterogeneity is not essential for our results. It is, however, useful to "purify" potential mixed-strategy equilibria, and to better connect the theory to firm-level data. See Section 4 for further discussion.

A novel feature of our framework is that the firm can produce information to overcome its uncertainty about productivity and demand. In particular, the firm can obtain signals:

$$s_j^{\upsilon} = \upsilon_j + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}\left(0, [\tau_j^{\upsilon}]^{-1}\right) \tag{5}$$

and

$$s_j^{\omega} \in \{\text{red}, \text{blue}\} \;\; \text{with } \mathbb{P}\left(s_j^{\omega} = \omega_j | \omega_j\right) = \tau_j^{\omega} \in \left[\frac{1}{2}, 1\right], \tag{6}$$

where the errors of the signals are distributed independently of each other and of the errors of other firms' signals, and where $\tau_j^{\upsilon} \in \{\underline{\tau}^{\upsilon}, \bar{\tau}^{\upsilon}\}$ and $\tau_j^{\omega} \in \{\underline{\tau}^{\omega}, \bar{\tau}^{\omega}\}$ are the signal precisions.

We assume that, whereas the firm obtains the signals with precision $(\underline{\tau}^{\upsilon}, \underline{\tau}^{\omega})$ at no cost, it must allocate $\chi > 0$ units of labor to information production to increase the precision of its signals to $(\bar{\tau}^{\upsilon}, \bar{\tau}^{\omega}) > (\underline{\tau}^{\upsilon}, \underline{\tau}^{\omega})$. We denote firm j's information choice by $\iota_j \in \{0, 1\}$, where $\iota_j = 1$ whenever the firm produces information about its scale and product choice. To save on notation, in what follows, we let $\boldsymbol{s}_j \equiv \left[s_j^{\upsilon},\, s_j^{\omega}\right]$ and $\boldsymbol{\tau}_j \equiv \left[\tau_j^{\upsilon},\, \tau_j^{\omega}\right]$.

3.2 Markets and Timing

Timing. The economy proceeds through three stages. In *Stage 1*, each firm j learns its mean-productivity level, μ_j, and chooses whether or not to produce information, ι_j. The economy then proceeds to *Stage 2*, where, conditional on its information choice, each firm observes its signals, $\boldsymbol{s}_j$, and decides the amount of labor to employ, n_j, and the type of variety to produce, x_j. The economy concludes in *Stage 3*, where each firm learns its productivity and demand states, (υ_j, ω_j), and produces in accordance; each household i, in turn, learns its tastes, $\{\delta_{ij}\}_j$, chooses its demand for each variety and supplies labor to firms; markets clear according to the protocols discussed next; and consumption takes place.

Markets. The market for labor is perfectly competitive: each household i supplies labor and each firm j hires the units of labor it desires to employ, taking the market wage, w, as given. In the market for variety j, each household takes its price, p_j, as given and chooses how many units of the variety to consume. By contrast, when choosing its production, the monopolistically competitive firm internalizes that its output affects the clearing price, p_j.

3.3 Remarks on Modeling Approach

In our baseline framework, information affects firm behavior through two channels.

First, the production of information helps a firm learn its productivity state, υ_j, and hence better choose its overall *scale* of production. This channel captures the canonical approach to modeling information frictions in macroeconomics, going back to at least Lucas (1972).

Second, the production of information also helps a firm learn its demand state, ω_j, and hence allocate its factors of production towards the variety-type most preferred by consumers. Although we have modeled this channel through *product choice*, it is isomorphic to several natural alternatives: e.g., the allocation of factors between different plants, or input sourcing from different suppliers.[15] Common to all these interpretations is that the production of information helps a firm improve its internal factor allocation.

Combined, the two channels through which information affects firm behavior are thought to have been among the most potent ways by which advances in data-processing technologies have improved firm decision-making over the past two decades (Brynjolfsson and McElheran, 2016; Ali *et al.*, 2020; Adams *et al.*, 2024; Veldkamp and Chung, 2024; Abis and Veldkamp, 2024). Yet, a common concern among academics and policymakers is that advances in data-processing technologies have also facilitated discriminatory practices, enabling firms to extract ever larger rents from consumers. We will capture this third channel of information in Section 5, where the production of information also helps firms optimize their *pricing strategies*. As we will show, the interaction between all three channels shape the normative properties of our economy, yielding valuable lessons about how to best regulate data production.

Finally, our baseline economy features both a fixed supply of production factors (i.e., labor) and a fixed set of firms or varieties. In Section 6, we show that standard approaches to introduce factor supply elasticity through capital accumulation and through entry and exit of varieties merely amplify the aggregate consequences of information production.

3.4 Optimization and Equilibrium

Household Problem. Household $i \in [0,1]$ chooses consumption of individual varieties $\{c_{ij}\}_{j\in[0,1]}$ in *Stage 3* to maximize its utility in (1) subject to the budget constraint:

$$\int_0^1 p_j \cdot c_{ij} \cdot dj = w \cdot N + \int_0^1 \pi_{ij} \cdot dj, \tag{7}$$

where p_j is the price of variety j, w is the wage, and π_{ij} are the profits from firm j. The household takes prices, $\{p_j\}_{j\in[0,1]}$, and the wage rate, w, as given when solving its problem. The solution yields:

$$c_{ij} = \delta_{ij}^{\theta-1} \cdot p_j^{-\theta} \cdot C_i, \tag{8}$$

[15] See, for example, the different business cases reported in O'Neill (2023). See also The Wall Street Journal's report on Levi's, whose use of information technologies to analyze fashion trends among young consumers was instrumental in the introduction of 'baggy jeans' (https://www.wsj.com/articles/how-tech-helped-levis-ride-the-baggy-jeans-trend-f290721d).

where we have normalized the ideal price index to one. Since all households are ex-ante identical, $C_i = C$ for all i, and the aggregate demand for variety j equals:

$$c_j \equiv \int_0^1 c_{ij} \cdot di = \delta_j^{\theta-1} \cdot p_j^{-\theta} \cdot C \quad \text{with} \quad \delta_j \equiv \delta(\gamma_j) = \left(\gamma_j \cdot \delta_H^{\theta-1} + (1-\gamma_j) \cdot \delta_L^{\theta-1}\right)^{\frac{1}{\theta-1}}. \tag{9}$$

Thus, it is *as if*, for each variety j, there is a *representative consumer* with demand shifter δ_j; but this shifter depends on the share γ_j of H-type consumers that the firm actually faces.

Firm Problem. The ex-post profits of firm $j \in [0,1]$ are given by the firm's revenue net of its expenditures on labor and information:

$$\pi_j = p_j \cdot y_j - w \cdot n_j - w \cdot \chi \cdot \iota_j. \tag{10}$$

In *Stage 2*, conditional on its information set $(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j)$, the firm chooses labor, n_j, and variety-type, x_j, to maximize its expected profits from goods production:

$$\widehat{\pi}_j\left(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right) \equiv \max_{n_j, x_j} \mathbb{E}\left[p_j \cdot y_j - w \cdot n_j | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right], \tag{11}$$

subject to feasibility (3) and the output of the variety being equal its demand (9). Here, $\mathbb{E}\left[\cdot | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]$ denotes the expectations operator conditional on $(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j)$. When choosing n_j, the firm optimally equates its expected marginal revenue product of labor to the wage:

$$\frac{\theta-1}{\theta} \cdot \frac{\mathbb{E}\left[p_j \cdot y_j | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]}{n_j} = w. \tag{12}$$

When choosing the variety-type, the firm optimally sets:

$$x_j = \arg\max_{s_j^\omega \in \{\text{red}, \text{blue}\}} \mathbb{E}[p_j \cdot y_j | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j]. \tag{13}$$

In *Stage 1*, conditional on its mean-productivity, μ_j, the firm chooses information production, ι_j, to maximize its expected profits. As a result, it optimally sets:

$$\iota_j \begin{cases} = 1 & \text{if } \mathbb{E}\left[\widehat{\pi}_j\left(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right) | \mu_j, \boldsymbol{\tau}_j = \bar{\boldsymbol{\tau}}\right] - w \cdot \chi \geq \mathbb{E}\left[\widehat{\pi}_j\left(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right) | \mu_j, \boldsymbol{\tau}_j = \underline{\boldsymbol{\tau}}\right] \\ = 0 & \text{otherwise} \end{cases}, \tag{14}$$

where $\mathbb{E}\left[\cdot | \mu_j, \boldsymbol{\tau}_j\right]$ denotes the expectations operator conditional on μ_j and $\boldsymbol{\tau}_j$.

Equilibrium Notion. An equilibrium consists of allocations $\left\{\{c_{ij}\}_{i\in[0,1]}, y_j, x_j, n_j, \iota_j\right\}_{j\in[0,1]}$, prices $\{p_j\}_{j\in[0,1]}$, and a real wage w such that:

- Given the prices and the wage, the allocations solve the household and the firm problems, i.e., Equations (7)-(14) hold, and;

- Given the allocations, the goods and the labor markets clear, i.e., $c_j = y_j$ for all $j \in [0,1]$ and $\int_0^1 (n_j + \chi \cdot \iota_j) \cdot dj = N$.

4 Equilibrium Characterization

We proceed by studying cross-sectional outcomes at the firm level, taking as given all economy-wide variables (Section 4.1). We then turn to the aggregate consequences of firms' information choices (Section 4.2). We end this section by analyzing the aggregate effects of improvements in data-processing technologies within our baseline framework (Section 4.3).

4.1 Information in the Cross-Section

An advantage of our framework is that all firms' choices are driven by two simple objects. The first captures the relevant notion of *market size* faced by firms in our economy and encodes all relevant general-equilibrium interactions. The second, by contrast, measures the profitability boost that a firm can expect if it were to produce information. :

Definition 1. *Let* $\Omega \equiv C \cdot \left(\frac{\theta}{\theta-1} \cdot w\right)^{-\theta}$ *be the* ***market size*** *faced by firms in the economy.*

Thus, market size, Ω, is large when aggregate consumption demand, C, is large, or when labor costs, w, are low. Market size is a key determinant of firms' information production choices, in conjunction with what we denote as firms' *information shifter.*

Definition 2. *For* $\boldsymbol{\tau} = (\tau^{\upsilon}, \tau^{\omega})$, *define the firms'* ***information shifter*** *as:*

$$g(\boldsymbol{\tau}) \equiv \left[\tau^{\omega} \cdot \delta(\bar{\gamma})^{\frac{\theta-1}{\theta}} + (1-\tau^{\omega}) \cdot \delta(\underline{\gamma})^{\frac{\theta-1}{\theta}}\right]^{\frac{\theta}{\theta-1}} \cdot \exp^{\frac{1}{2} \cdot \frac{\theta-1}{\theta} \cdot \frac{1}{\tau_a} \cdot \frac{\tau_a + \theta \cdot \tau^{\upsilon}}{\tau_a + \tau^{\upsilon}}}, \tag{15}$$

where $\delta(\cdot)$ *is given by Equation (9).*

As we will see, $g(\cdot)$ compactly summarizes all the benefits that information has both at firm level and in the aggregate. We are now ready to characterize firms' optimal choices.

Proposition 1. *In any equilibrium:*

(i) Firm j *with mean productivity* μ_j, *which observes signals* $\boldsymbol{s}_j$ *with precisions* $\boldsymbol{\tau}_j$, *chooses:*

$$x_j = s_j^{\omega} \quad \text{and} \quad n_j = \mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^{\theta} \cdot \Omega. \tag{16}$$

(ii) Firm j with mean productivity μ_j chooses:

$$\iota_j = \begin{cases} 1 & \textit{if} \quad \frac{1}{\theta-1} \cdot \left(\mathbb{E}\left[n_j | \mu_j, \boldsymbol{\tau}_j = \bar{\boldsymbol{\tau}}\right] - \mathbb{E}\left[n_j | \mu_j, \boldsymbol{\tau}_j = \underline{\boldsymbol{\tau}}\right] \right) \geq \chi \\ 0 & \textit{otherwise} \end{cases}, \tag{17}$$

where:

$$\mathbb{E}\left[n_j | \mu_j, \boldsymbol{\tau}_j\right] = \exp^{(\theta-1)\cdot\mu_j} \cdot g(\boldsymbol{\tau}_j)^{\theta-1} \cdot \Omega. \tag{18}$$

Proposition 1 has two sets of important implications for firm behavior.

1) Information and Resource Allocation. Part (i) of Proposition 1 implies that information boosts the efficiency of resource allocation both *within* and *across* firms.

First, by helping a firm allocate its factors of production towards the variety-type most preferred by households, information increases firm-level total-factor productivity. To see this, observe that for given input and information choices (x_j, n_j, ι_j), the total surplus (or utility) generated by firm j rises monotonically with $(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}}$.[16] Thus, an appropriate measure of *(log-)total factor productivity* for firm j is:

$$\text{tfp}_j \equiv \frac{\theta - 1}{\theta} \cdot \log(\delta_j \cdot A_j). \tag{19}$$

Because the demand shifter, δ_j, faced by firm j is affected by the firm's choice of product customization, x_j, the distribution of tfp_j becomes endogenous to the firm's information choice:

Corollary 1. *For all j, the conditional mean and variance of tfp is:*

$$\mathbb{E}[\text{tfp}_j | \mu_j, \boldsymbol{\tau}_j] = \frac{\theta-1}{\theta} \cdot \left[\tau_j^{\omega} \cdot \log\left(\delta(\bar{\gamma})\right) + (1 - \tau_j^{\omega}) \cdot \log\left(\delta(\underline{\gamma})\right) + \mu_j \right] \tag{20}$$

$$\mathbb{VAR}[\text{tfp}_j | \mu_j, \boldsymbol{\tau}_j] = \left(\frac{\theta-1}{\theta}\right)^2 \cdot \left[\tau_j^{\omega} \cdot (1 - \tau_j^{\omega}) \cdot \left[\log\left(\delta(\bar{\gamma})\right) - \log\left(\delta(\underline{\gamma})\right)\right]^2 + \frac{1}{\tau_a} \right], \tag{21}$$

where $\delta(\cdot)$ is given by Equation (9).

All else equal, tfp is higher and less volatile among firms with more precise demand information, i.e., a larger τ_j^{ω}. This is because firms with more precise demand information are able to customize their products more effectively to consumer tastes, thereby raising and stabilizing the effective demand for them. We note that a variant of this "internal allocation" channel of information also appears in Farboodi and Veldkamp (2024), albeit modeled in a reduced-form manner that directly modifies the process for firm-level tfp.

[16] The total surplus generated by firm j is equal to the total utility generated by the firm, $(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} \cdot n_j^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}}$, net of its labor costs, $w \cdot (n_j + \chi \cdot \iota_j)$, all measured in the numeraire.

Second, by helping a firm correlate its overall employment of factors with the realized shocks to the firm's tfp, information also reduces (ex-post) factor misallocation across firms. To see this, let mrp_j denote the (*log-*)*marginal-revenue product* of labor for firm j:

$$\text{mrp}_j \equiv \log(p_j \cdot y_j) - \log(n_j). \tag{22}$$

Absent information frictions, $\text{mrp}_j = \text{mrp}_{j'}$ for all j, j' (Equation 12). Any dispersion in mrp_j is thus a tell-tale sign of inefficiency resulting from information frictions.

Corollary 2. *For all j, the conditional variance of mrp is:*

$$\begin{aligned} \mathbb{VAR}[\text{mrp}_j|\mu_j, \boldsymbol{\tau}_j] &= \mathbb{VAR}[\text{tfp}_j|\mu_j, \boldsymbol{\tau}_j] + \left(\frac{\theta-1}{\theta}\right)^2 \cdot \left(\frac{1}{\tau_a + \tau_j^v} - \frac{1}{\tau_a}\right) \\ &= \mathbb{VAR}[\text{error}_j|\mu_j, \boldsymbol{\tau}_j], \end{aligned} \tag{23}$$

where $\text{error}_j \equiv \log(p_j \cdot y_j) - \log\left(\mathbb{E}[p_j \cdot y_j|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j]\right)$.

All else equal, factor misallocation is therefore lower among firms with more precise information, i.e., a larger τ_j^v and τ_j^ω. This is because firms with more precise information are also those that have more accurate revenue expectations. Indeed, the cross-sectional dispersion in mrp_j is equal to the variance of a firm's log-revenue error, a mapping we will later utilize to pin down the implied decline in mrp-dispersion from the observed increase in accuracy in the data (Section 2). We note that the role of information through its effect on factor misallocation has also been studied in David *et al.* (2016) and David and Venkateswaran (2019), albeit in a setting where firms do not themselves decide on their information production.

2) Information and Firm-size Distribution. Part (ii) of Proposition 1 implies that—by improving efficiency of resource allocation—information also alters the *firm-size distribution*.

To see this, consider the benefits to a firm from producing information, which are given by the change in the firm's expected profits from goods production resulting from a more efficient resource allocation (Corollaries 1 and 2). From the optimality condition (12), a firm's expected profits from goods production are proportional to its expected employment:

$$\mathbb{E}\left[p_j \cdot y_j - w \cdot n_j|\mu_j, \boldsymbol{\tau}_j\right] = \frac{1}{\theta-1} \cdot w \cdot \mathbb{E}[n_j|\mu_j, \boldsymbol{\tau}_j]. \tag{24}$$

It then follows from Equation (18) in Proposition 1 that the information shifter $g(\cdot)$ is key to understanding the benefits of information. In particular, given the properties of $g(\cdot)$:

$$\frac{\partial \mathbb{E}[n_j|\mu_j, \boldsymbol{\tau}_j]}{\partial \tau_j^v} > 0, \ \frac{\partial \mathbb{E}[n_j|\mu_j, \boldsymbol{\tau}_j]}{\partial \tau_j^\omega} > 0, \ \frac{\partial^2 \mathbb{E}[n_j|\mu_j, \boldsymbol{\tau}_j]}{\partial \tau_j^v \partial \tau_j^\omega} > 0. \tag{25}$$

Thus, not only does each channel of information alone—whether learning about demand or productivity—boost a firm's size and profits, but the two channels interact and *reinforce* each other, amplifying the gains from information production.

We have so far shown that, holding fixed a firm's ex-ante mean productivity, μ_j, information production—by improving the efficiency of resource allocation—boosts that firm's expected size and profitability. In equilibrium, however, there is also a selection of firms into information production, which we must consider when studying any relationship between information and firm characteristics. Since the benefits from information production scale with a firm's expected size—which grows with μ_j (Equation 18),—but the information cost χ does not, it is the ex-ante more productive firms that choose to produce information. As it turns out, such a selection, if anything, reinforces the effects of information production that we outlined above.

Corollary 3. *In equilibrium, firm j produces information if and only if:*

$$\mu_j \geq \bar{\mu} \equiv \frac{1}{\theta - 1} \cdot \log \left[\frac{(\theta - 1) \cdot \chi}{\left(g\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} - g\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1} \right) \cdot \Omega} \right]. \tag{26}$$

Further, more informed firms, i.e., $\{j : \boldsymbol{\tau}_j = \bar{\boldsymbol{\tau}}\}$, have on average higher and less dispersed tfp, less dispersed mrp, and they grow larger and more profitable as measured by:

$$\mathbb{E}[\mathcal{Z}_j | \boldsymbol{\tau}_j = \bar{\boldsymbol{\tau}}] > \mathbb{E}[\mathcal{Z}_j | \boldsymbol{\tau}_j = \underline{\boldsymbol{\tau}}], \tag{27}$$

for $\mathcal{Z}_j \in \{n_j, n_j + \chi \cdot \iota_j, p_j \cdot y_j, p_j \cdot y_j - w \cdot n_j, \pi_j\}$.

To conclude, it is worth noting that, although in our setting selection into information production is driven by ex-ante productivity differences, alternative specifications are also possible. For example, the cross-sectional patterns described in Corollary 3 would be the same if we instead assumed that firms were heterogeneous in information costs (i.e., in χ_j); or if there was no heterogeneity whatsoever, but we instead focused on the parameter region in which the equilibrium is in mixed strategies (i.e., where an interior share of firms produce information). Compared to these alternatives, our current approach, nevertheless, has the distinct advantage of allowing us to more closely discipline the link between firm size and the accuracy of firms' information, using our empirical results documented in Section 2.

We have characterized how information production affects firm choices and outcomes at the micro level. The next natural step is to study the general equilibrium of the economy. We return to validate the cross-sectional predictions, described in Corollaries 1-3 above, using the I/B/E/S-Compustat sample in Section 6.

4.2 Information in General Equilibrium

To analyze the general equilibrium, we first define the relevant notion of *aggregate total factor productivity* (TFP) for our economy, which captures the units of aggregate consumption (or utility) our economy creates from the workers that it allocates to goods production.

Definition 3. *Given aggregate consumption, C, and aggregate employment in goods production, $\mathcal{N} \equiv \int_0^1 n_j \cdot dj$, we define aggregate* ***total factor productivity*** *as $\mathcal{A} \equiv C \cdot \mathcal{N}^{-1}$.*

Using households' goods demands, market clearing for variety j, and firms' labor choices in Proposition 1, we can express TFP solely as a function of firms' information sets, $\{(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j)\}_j$, and hence their information choices:

$$\mathcal{A} = \left(\int_0^1 (\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} \cdot \left(\frac{n_j}{\mathcal{N}}\right)^{\frac{\theta-1}{\theta}} \cdot dj\right)^{\frac{\theta}{\theta-1}} = \left(\int_0^1 \mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^{\theta} \cdot dj\right)^{\frac{1}{\theta-1}}, \tag{28}$$

where to obtain the last equality we have made use of the fact that:

$$\mathcal{N} = \int_0^1 \mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^{\theta} \cdot dj \cdot \Omega. \tag{29}$$

Since from Corollary 3, we know that firms' information choices are fully pinned down by the mean productivity $\bar{\mu}$ of the marginal firm that is just indifferent to producing information (henceforth, the *marginal-type*), we have that:

Lemma 1. *Given the marginal-type, $\bar{\mu}$, aggregate total factor productivity, $\mathcal{A}$, equals:*

$$\mathcal{A}(\bar{\mu}; g) \equiv \exp^{\frac{\theta-1}{2} \cdot \frac{1}{\tau_\mu}} \cdot \left[g(\underline{\boldsymbol{\tau}})^{\theta-1} \cdot (1 - \xi(\bar{\mu})) + g(\bar{\boldsymbol{\tau}})^{\theta-1} \cdot \xi(\bar{\mu})\right]^{\frac{1}{\theta-1}}, \tag{30}$$

where $\xi(\bar{\mu}) \equiv \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu} + (\theta - 1) \cdot \frac{1}{\sqrt{\tau_\mu}}\right)$ and $\Phi(\cdot)$ is the standard normal c.d.f.

The main implication of Lemma 1 is that, by boosting the efficiency of resource allocation—both within and across firms—information production raises the productivity of the economy as a whole. Notice that the weight $\xi(\bar{\mu})$ in Equation (30) captures the share of information producing firms in the economy, adjusted for the fact that these firms are larger (Corollary 3) and thus have a larger impact on the aggregate allocation of resources. As more firms produce information—that is, as $\bar{\mu}$ falls and $\xi(\bar{\mu})$ rises—more firms feature the larger information shifter and, as a result, TFP increases (since $g(\bar{\boldsymbol{\tau}}) > g(\underline{\boldsymbol{\tau}})$ by Definition 2). Notice that we have made the dependence of TFP on the information shifter, $g(\cdot)$, explicit in Equation (30), a notational device that will later prove useful in Section 5.

Aggregate productivity is an important determinant of market size, $\Omega = C \cdot \left(\frac{\theta}{\theta-1} \cdot w\right)^{-\theta}$, and hence all economy-wide objects in our economy. Combining the definition of TFP with Equation (29) and the labor market-clearing condition shows that:

$$\Omega = \frac{\mathcal{A}(\bar{\mu}, g) \cdot \left[N - \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right) \cdot \chi\right]}{\mathcal{A}(\bar{\mu}, g)^{\theta}}. \tag{31}$$

Recall from Section 4.1 that a firm's incentive to produce information depends on the market size, Ω. Equation (31) shows that, in general equilibrium, the market size itself also depends on the collective information choices of firms, as summarized by the marginal-type, $\bar{\mu}$.

Information production affects market size through its effect both on aggregate consumption and on firms' production costs. First, the denominator in Equation (31) captures the fact that more information production, i.e., a lower $\bar{\mu}$, raises firms' production costs, since by boosting TFP information also raises the demand for labor and hence the equilibrium wage ($w = \frac{\theta-1}{\theta} \cdot \mathcal{A}$). Second, the numerator in Equation (31), in part, captures the fact that more information production nevertheless also directly raises aggregate consumption, as it raises the economy's TFP. Finally, the last term in the numerator shows that information production also has another depressing effect on overall consumption, as it diverts scarce labor away from the production of goods. It is straightforward to see that, since $\theta > 1$, the net effect of these opposing forces is negative, and market size decreases with information production (i.e., as $\bar{\mu}$ falls). This, in turn, implies that, in equilibrium, firms' information production choices are *strategic substitutes*, thereby guaranteeing the uniqueness of any equilibrium.

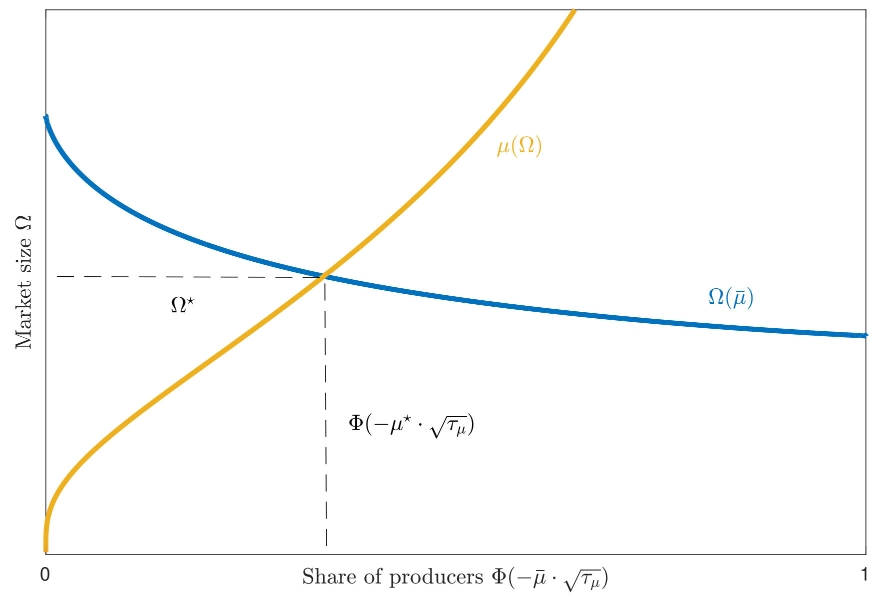
A convenient feature of our framework is that the economy's equilibrium can be studied through the intersection of just two schedules. Equation (31) defines a continuous schedule $\Omega : \mathbb{R} \to \mathbb{R}^{+}$, which maps a given marginal-type $\bar{\mu}$ into a market size $\Omega(\bar{\mu})$ that is consistent with market clearing. By contrast, Equation (26) defines a continuous schedule $\bar{\mu} : \mathbb{R}^{+} \to \mathbb{R}$ that for a given market size, Ω, yields the marginal-type $\bar{\mu}(\Omega)$ that is indifferent to producing information. An equilibrium is given by the fixed point of the composite map: $\bar{\mu} \circ \Omega : \mathbb{R} \to \mathbb{R}$. The following proposition states that such a fixed point exists and is always unique.

Proposition 2. *An equilibrium exists, is unique, and in it the marginal-type that is just indifferent to producing information solves:*

$$\bar{\mu}(\Omega(\mu^{\star})) = \mu^{\star}, \tag{32}$$

where $\bar{\mu}(\cdot)$ and $\Omega(\cdot)$ are defined by Equations (26) and (31), respectively. Aggregate TFP, $\mathcal{A}^{\star}$,

Figure 3: Equilibrium Determination



Note: The upward-sloping (orange) locus depicts the relationship between market size, Ω, and share of firms that produce information, $\Phi(-\bar{\mu} \cdot \sqrt{\tau_\mu})$, as defined by Equation (26). The downward-sloping (blue) locus instead depicts the relationship between Ω and $\Phi(-\bar{\mu} \cdot \sqrt{\tau_\mu})$ as defined by Equation (31).

and consumption, $C^\star$, are, in turn, equal to:

$$\mathcal{A}^\star = \mathcal{A}(\mu^\star, g) \text{ and } C^\star = \mathcal{A}^\star \cdot \left[N - \Phi\left(-\mu^\star \cdot \sqrt{\tau_\mu}\right) \cdot \chi\right], \tag{33}$$

where $\mathcal{A}(\cdot)$ is stated in Lemma 1 and $\Phi(\cdot)$ is the standard normal c.d.f.

The equilibrium's determination is illustrated in Figure 3, which depicts the equilibrium relationship between market size, Ω, and the share of information producers, as given by $\Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)$. The orange locus depicts the combinations of $\Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)$ and Ω that are consistent with the indifference condition in Equation (26). This locus is upward sloping because, as we discussed in Section 4.1, a firm's incentive to produce information increases with the market size. The blue locus, by contrast, depicts the combinations of $\Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)$ and Ω, which are consistent with market clearing, as described in Equation (31). As we discussed above, this locus is instead downward sloping. An equilibrium is characterized by the unique intersection of the orange and blue loci, which pins down the equilibrium marginal-type $\mu^\star$ and the market size $\Omega^\star$ corresponding to it.

4.3 On Advances in Data-Processing Technologies

We examine the macroeconomic consequences of advances in data-processing technologies using our baseline framework. The past two decades have seen firms increasingly adopt and rely on large-scale, data-intensive algorithms to enhance their economic decision-making (e.g., Baley and Veldkamp, 2025). This shift in the nature by which firms make their economic choices has, in turn, been driven by substantial declines in computing costs and improvements in processing speeds (e.g., Nordhaus, 2008; Coyle and Hampton, 2024; Gill *et al.*, 2024).

We analyze the aggregate effects of such technological advances through two comparative static exercises: (i) a reduction in the information cost parameter, χ; and (ii) an increase in the precision of information that can be obtained through information production, $\bar{\boldsymbol{\tau}} = (\bar{\tau}^{\upsilon}, \bar{\tau}^{\omega})$. These changes serve as stylized representations of declines in computing costs and improvements in processing capabilities. The following proposition summarizes their effects:

Proposition 3. *An improvement in data-processing technologies, such as a fall in χ, a rise in $\bar{\tau}^{\upsilon}$, or a rise in $\bar{\tau}^{\omega}$, leads to an increase in the share of firms producing information, $\Phi\left(-\mu^{\star} \cdot \sqrt{\tau_{\mu}}\right)$, an increase in aggregate TFP, $\mathcal{A}^{\star}$, and an increase in welfare, $C^{\star}$.*

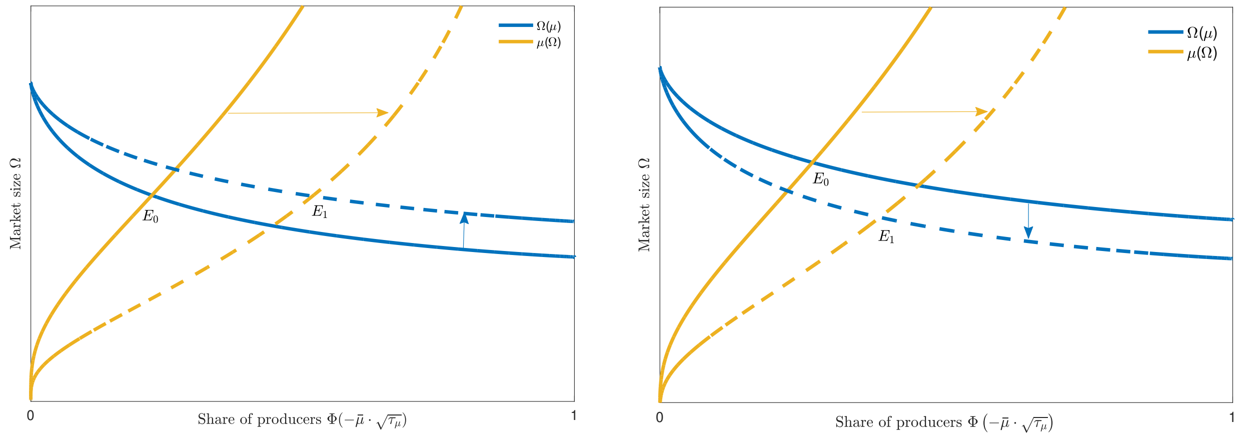
Figure 4 illustrates the equilibrium effects of improvements in data-processing technologies on the share of information-producing firms, $\Phi(-\mu^{\star} \cdot \sqrt{\tau_{\mu}})$, and on the market size, $\Omega^{\star}$, respectively. Panel (a) depicts the impact of a decline in χ, while Panel (b) illustrates the consequences of an increase in $\bar{\boldsymbol{\tau}}$.[17] Both improvements make information production more attractive from a firm's perspective (Proposition 1), leading to a rightward shift in the orange loci in Figure 4. However, their general equilibrium consequences differ.

A reduction in χ shifts the blue locus upward, as lower information costs allow the economy to allocate more labor to goods production, thereby increasing consumption and expanding market size, $\Omega(\mu^{\star})$. In contrast, an increase in $\bar{\boldsymbol{\tau}}$ shifts the blue locus downward, as the higher accuracy of firms' information enhances total factor productivity, which, as discussed in Section 4.2, has a depressing effect on the equilibrium market size, $\Omega(\mu^{\star})$.

Despite these differing mechanisms, the overall macroeconomic effects of both technological improvements are, nevertheless, similar. In both cases, TFP, consumption, and welfare increase. This should not come as a surprise, however, given the benign nature of information assumed so far. Indeed, we formally establish in Section 5.5 that the equilibrium of our baseline economy is efficient. As a result, in this setting, improvements in data-processing technologies are inherently beneficial, leading invariably to higher productivity and welfare.

[17] The aggregate effects of an increase in $\bar{\tau}^{\upsilon}$ are similar to those of an increase in $\bar{\tau}^{\omega}$.

Figure 4: Improvements in Data-Processing Technologies



Panel (a): a fall in χ | Panel (b): a rise in $\bar{\boldsymbol{\tau}}$

Note: Panel (a) depicts the effects of a fall in the cost of information, χ, whereas Panel (b) depicts the effects of a rise increase in the accuracy of information, $\bar{\boldsymbol{\tau}} = (\bar{\tau}^{\upsilon}, \bar{\tau}^{\omega})$. The solid (dashed) loci depicts the relationship between market size, Ω, and share of firms producing information, $\Phi(-\bar{\mu} \cdot \sqrt{\tau_{\mu}})$, before (after) the improvement in information technologies. Equilibria before (after) the change are denoted by E_0 (E_1).

5 The Rent-Extracting Economy

We have thus far taken a benign view of information, emphasizing its role in enhancing resource allocation within and across firms. Yet, the production and use of information need not be universally beneficial. A growing concern about recent advances in data-processing technologies is their potential to increase firms' capacity to extract consumer surplus by facilitating more sophisticated discriminatory practices (European Commission Report, 2020).[18] To examine this issue, we extend our baseline framework to analyze how information influences firms' pricing strategies. We refer to this extended framework as the *rent-extracting economy*, which, as we will show, encompasses our baseline economy as a special parametric case.

5.1 Information and Price Discrimination

We have up to now assumed that firms are restricted to sell varieties at uniform prices. However, given that each monopolistic firm faces consumers with heterogeneous tastes—i.e., a share $\gamma_j = \gamma(\omega_j | x_j)$ with demand-shifter δ_H and a share $1 - \gamma_j$ with demand-shifter δ_L—such a trading mechanism is generally sub-optimal.[19] We therefore now depart from the baseline

[18] See, e.g., O'Neill (2023) and Eeckhout and Veldkamp (2023).

[19] As it is trivial for the mechanism to elicit the common component, ω_j, of consumer preferences (see, e.g., Crémer and McLean (1988)), we assume without loss that the firm knows it when designing the mechanism.

framework by allowing each firm to design *the optimal trading mechanism* by which to allocate its goods to the different types of consumers. While we will effectively focus on *second-degree price discrimination through quantities*,[20] our main qualitative results also hold in settings where price discrimination arises due to *bilateral bargaining* or *quality differentiation*. We refer the interested reader to Appendix C for these alternative specifications.

In what follows, we discuss the main features of a firm's problem and relegate detailed derivations to Appendix C. The only change to our baseline framework is that, in *Stage 3*, each firm j now proposes a menu $\mathcal{M}_j = \{(t_j, q_j)\}$ to consumers, each of whom then decides whether and which allocation to accept. If the consumer accepts allocation (t_j, q_j), she receives q_j units of variety j in exchange for a payment of t_j; otherwise, she does not trade.

The firm's problem at this stage is to maximize its revenues, as all cost are sunk:

$$\gamma_j \cdot t_j^H + (1-\gamma_j) \cdot t_j^L, \tag{34}$$

where (t_j^l, q_j^l) denotes the allocation accepted by the type-$l \in \{H, L\}$ consumer. This maximization problem is subject to incentive compatibility, individual rationality, and feasibility constraints. Since the firm does not observe individual consumer-types, in equilibrium, each consumer-type must select its intended allocation (*incentive compatibility*):

$$(t_j^l, q_j^l) = \arg\max_{(t_j, q_j) \in \mathcal{M}_j} (\delta_l \cdot q_j)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - t_j \quad \text{for} \quad l \in \{H, L\}, \tag{35}$$

where the right-hand side of equation (35) is the total surplus that the consumer gains by accepting the allocation (t_j, q_j). Moreover, to participate in the mechanism, type-l consumer must obtain a non-negative surplus from it (*individual rationality*):

$$\left(\delta_l \cdot q_j^l\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - t_j^l \geq 0 \quad \text{for} \quad l \in \{H, L\}. \tag{36}$$

Finally, as the firm supplies its variety inelastically in *Stage 3*, the aggregate quantity allocated to the consumers cannot exceed the quantity that the firm has produced (*feasibility*):

$$\gamma_j \cdot q_j^H + (1-\gamma_j) \cdot q_j^L \leq A_j \cdot n_j. \tag{37}$$

The solution to the firm's problem admits a standard form. At the optimum, the L-type's individual rationality, the H-type's incentive compatibility, and the feasibility constraints bind; in turn, the quantities allocated to consumers are set to equalize the marginal revenues extracted across types. It follows that the share of produced goods, α_j^l, that firm j allocates

[20] See, also, Bornstein and Peter (2024) for a recent macroeconomic application of discriminatory pricing.

to each type-l consumer satisfies:

$$\alpha_j^H \equiv \alpha^H(\gamma_j) = \frac{\delta_H^{\theta-1}}{\gamma_j \cdot \delta_H^{\theta-1} + (1-\gamma_j) \cdot (\psi_j \cdot \delta_L)^{\theta-1}}, \tag{38}$$

$$\alpha_j^L \equiv \alpha^L(\gamma_j) = \frac{(\psi_j \cdot \delta_L)^{\theta-1}}{\gamma_j \cdot \delta_H^{\theta-1} + (1-\gamma_j) \cdot (\psi_j \cdot \delta_L)^{\theta-1}}, \tag{39}$$

where the sole departure from allocative efficiency is the "micro-level wedge":

$$\psi_j \equiv \psi(\gamma_j) = \begin{cases} \left[\frac{1-\gamma_j \cdot \left(\frac{\delta_H}{\delta_L}\right)^{\frac{\theta-1}{\theta}}}{1-\gamma_j}\right]^{\frac{\theta}{\theta-1}} & \text{if } \delta_L^{\frac{\theta-1}{\theta}} \geq \gamma_j \cdot \delta_H^{\frac{\theta-1}{\theta}} \\ 0 & \text{otherwise} \end{cases}, \tag{40}$$

which arises due to the binding incentive constraint of the H-type consumer.

The wedge $\psi_j \leq 1$ captures the fact that, to extract rents from the H-type more effectively, the firm chooses to inefficiently restrict trade with the L-type consumer. Indeed, at the optimum, the ratio of marginal utilities of consumption of the L-type to the H-type consumer equals $\psi_j^{-\frac{\theta-1}{\theta}}$, whereas efficiency requires equalization of marginal utilities. Since ψ_j further decreases with γ_j, the departure from efficiency becomes even *larger* when the firm faces a *more favorable* selection of consumers, a feature that will be important later.

5.2 Profit vs Social-Surplus Maximization

The upshot from the above characterization is that firm j's ex-post profits, which govern the firm's optimal information and input choices in *Stages 1* and *2* can be expressed as:

$$\pi_j = \left(\delta_j^R \cdot A_j\right)^{\frac{\theta-1}{\theta}} \cdot n_j^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - w \cdot n_j - w \cdot \chi \cdot \iota_j, \tag{41}$$

where it is *as if* firm j faces a representative consumer with a "*revenue-based demand shifter*":

$$\delta_j^R \equiv \delta^R(\gamma_j) = \left(\gamma_j \cdot \left[\delta_H \cdot \alpha^H(\gamma_j)\right]^{\frac{\theta-1}{\theta}} + (1-\gamma_j) \cdot \left[\psi(\gamma_j) \cdot \delta_L \cdot \alpha^L(\gamma_j)\right]^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}}. \tag{42}$$

By contrast, the social surplus produced by firm j is given by:

$$u_j = \left(\delta_j^S \cdot A_j\right)^{\frac{\theta-1}{\theta}} \cdot n_j^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - w \cdot n_j - w \cdot \chi \cdot \iota_j, \tag{43}$$

where the "*surplus-based demand shifter*" that the firm faces equals:

$$\delta_j^S \equiv \delta^S(\gamma_j) = \left(\gamma_j \cdot \left[\delta_H \cdot \alpha^H(\gamma_j)\right]^{\frac{\theta-1}{\theta}} + (1-\gamma_j) \cdot \left[\delta_L \cdot \alpha^L(\gamma_j)\right]^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}}. \tag{44}$$

The properties of the two demand shifters, δ_j^R and δ_j^S, are central for the workings of the rent-extracting economy. The difference between them summarizes the conflict between profit and social surplus maximization that is inherent to the rent-extracting economy.

Lemma 2. *The revenue-based demand shifter, δ_j^R, is increasing in the share of high-demand types, γ_j, whereas the surplus-based demand shifter, δ_j^S, can be non-monotonic in γ_j. Moreover, $\delta_j^S \geq \delta_j^R$ with strict inequality if and only if $0 < \gamma_j < (\delta_L/\delta_H)^{\frac{\theta-1}{\theta}}$.*

Figure 5 illustrates these properties. First, the revenue-based demand shifter, δ_j^R, increases monotonically in γ_j, since the firm can always extract a larger surplus from consumers when their willingness to pay increases. Second, the difference between the two demand shifters effectively captures the "information rents" that the firm must leave to the H-type consumer to satisfy incentive compatibility. As a result, this difference is positive whenever there is a positive mass of H-types and the firm allocates some units to the L-types—i.e., when $0 < \gamma_j < (\delta_L/\delta_H)^{\frac{\theta-1}{\theta}}$. An immediate implication of this is that the ratio of the two demand shifters, δ_j^S/δ_j^R, must *increase at first* and *then decrease* with γ_j. In addition, as the figure shows, the departure from allocative efficiency—as captured by the ratio of marginal utilities across the two types of consumers (see earlier discussion)—can be so severe that the surplus-based demand shifter, δ_j^S, actually *decreases* with γ_j.

Definition 4. *We say rent extraction is **severe (mild)** if δ_j^S/δ_j^R falls (rises) as γ_j increases from $\underline{\gamma}$ to $\bar{\gamma}$, and it is **socially destructive** if δ_j^S falls as γ_j increases from $\underline{\gamma}$ to $\bar{\gamma}$.*

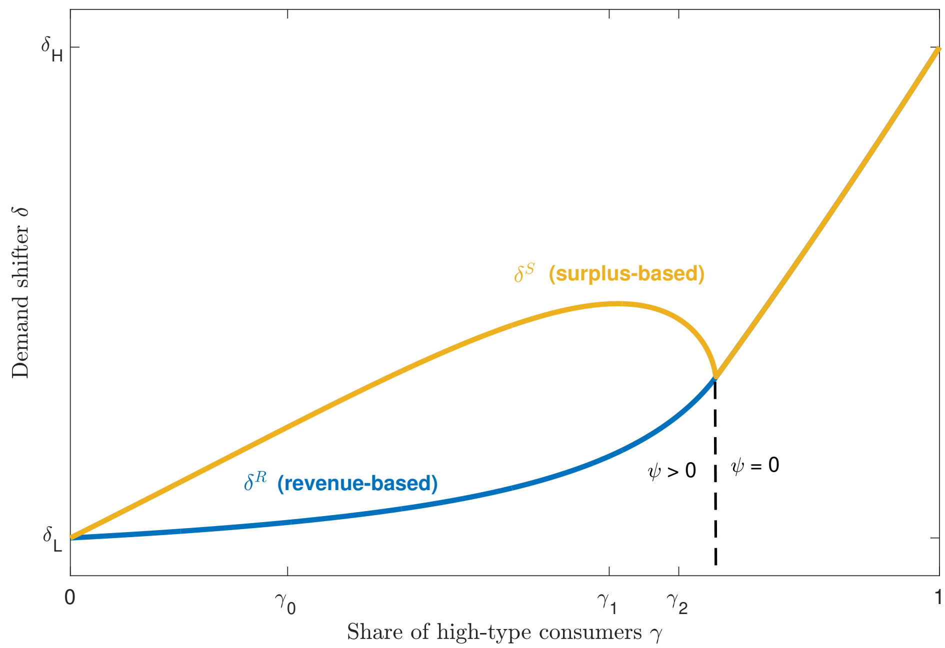
Recall that by producing information and better customizing its products, a firm raises the likelihood that it faces a share $\bar{\gamma}$ (rather than $\underline{\gamma}$) of H-type consumers from $\underline{\tau}^\omega$ to $\bar{\tau}^\omega$. Hence, when rent extraction is severe (mild), the social return to information about demand is lower (higher) than the private return; when rent extraction is socially destructive, the social return to information about demand is instead outright negative. These implications of price discrimination are key for the results that follow.[21]

5.3 Equilibrium of the Rent-Extracting Economy

We are ready to characterize the equilibrium of the rent-extracting economy. To do so, we define two main objects. First, we introduce a modified version of the information shifter,

[21] In Appendix C, moreover, we show that similar implications also obtain in alternative, plausible trading environments, where discrimination occurs through bilateral bargaining or quality differentiation.

Figure 5: Revenue- vs Surplus-Based Demand Shifters



Note: The figure depicts the demand shifters $\delta^R(\gamma_j)$ and $\delta^S(\gamma_j)$ as a function of γ_j. In the figure, rent extraction is mild if $\underline{\gamma} = \gamma_0 < \gamma_1 = \bar{\gamma}$, whereas it is both severe and socially destructive if $\underline{\gamma} = \gamma_1 < \gamma_2 = \bar{\gamma}$.

which will be instrumental in describing firm-level choices and outcomes:

$$g^R(\boldsymbol{\tau}) \equiv \left[\tau^\omega \cdot \delta^R(\bar{\gamma})^{\frac{\theta-1}{\theta}} + (1-\tau^\omega) \cdot \delta^R(\underline{\gamma})^{\frac{\theta-1}{\theta}}\right]^{\frac{\theta}{\theta-1}} \cdot \exp^{\frac{1}{2} \cdot \frac{\theta-1}{\theta} \cdot \frac{1}{\tau_a} \cdot \frac{\tau_a + \theta \cdot \tau^\upsilon}{\tau_a + \tau^\upsilon}}, \tag{45}$$

where we have merely substituted the demand shifter δ_j in Equation (15) with the revenue-based demand shifter, δ_j^R, as defined in Equation (42). As before, $g^R(\bar{\boldsymbol{\tau}})$ increases in $\bar{\boldsymbol{\tau}}$.

Second, we define a measure of the misalignment between profit and social surplus maximization for a firm with information precision $\boldsymbol{\tau} = (\tau^\upsilon, \tau^\omega)$:

$$\Delta(\boldsymbol{\tau}) \equiv \left[\frac{\tau^\omega \cdot \delta^S(\bar{\gamma})^{\frac{\theta-1}{\theta}} + (1-\tau^\omega) \cdot \delta^S(\underline{\gamma})^{\frac{\theta-1}{\theta}}}{\tau^\omega \cdot \delta^R(\bar{\gamma})^{\frac{\theta-1}{\theta}} + (1-\tau^\omega) \cdot \delta^R(\underline{\gamma})^{\frac{\theta-1}{\theta}}}\right]^{\frac{\theta}{\theta-1}}. \tag{46}$$

Crucially, this misalignment measure decreases (increases) with information precision τ^ω whenever rent extraction is severe (mild), as formalized in Definition 4.

With these definitions in place, we now analyze firm-level choices and outcomes. An equilibrium is defined analogously to that in our baseline economy, except that the exchange at *Stage 3* between the firm and the consumers occurs as described in Section 5.1.

Proposition 4. *An equilibrium of the rent-extracting economy exists, is unique, and in it:*

(i) Firm j's choices of x_j, n_j and ι_j are given by Proposition 1, with the only modification that the demand-shifter δ_j in Equation (16) is replaced by δ_j^R.

(ii) The equilibrium marginal-type, $\mu^\star$, is given by Proposition 2, with the only modification that the information shifter $g(\cdot)$ in Equations (26) and (31) is replaced by $g^R(\cdot)$.

Thus, when it comes to choices of variety-type, x_j, employment, n_j, and information production, ι_j, the behavior of firms in the rent-extracting economy is observationally equivalent to that in our baseline economy.[22] This should not come as a surprise: the only difference in the revenues and profits, associated with those choices, arises because firms now act *as if* facing the demand shifter δ_j^R, as opposed to δ_j. It then follows directly that the amount of information produced in the aggregate, as before, is pinned down by the unique fixed point of the composite map $\bar{\mu} \circ \Omega$, with the only difference that the schedules $\bar{\mu}(\cdot)$ and $\Omega(\cdot)$, defined in Equations (26) and (31), now use the modified information shifter $g^R(\cdot)$ in place of $g(\cdot)$.

An immediate consequence of Proposition 4 is that by looking only at *firm-level data* one cannot distinguish the rent-extracting economy from our baseline economy. The cross-sectional predictions, as described in Corollary 3, also hold for firms in the rent-extracting economy. There is, nevertheless, an important caveat. Since in the rent-extracting economy there is a conflict between profit and surplus maximization, revenue-based measures of tfp and factor misallocation can provide misleading information about the efficiency of resource allocation. This, in turn, implies that—in the aggregate—the rent-extracting economy can behave very differently from our baseline economy.

We can express aggregate TFP in the rent-extracting economy as follows:

$$\mathcal{A} = C \cdot \mathcal{N}^{-1} = \frac{\left(\int_0^1 \Delta(\boldsymbol{\tau}_j)^{\frac{\theta-1}{\theta}} \cdot \mathbb{E}\left[(\delta_j^R \cdot A_j)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^\theta \cdot dj\right)^{\frac{\theta}{\theta-1}}}{\int_0^1 \mathbb{E}\left[(\delta_j^R \cdot A_j)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^\theta \cdot dj}, \tag{47}$$

where $\Delta(\boldsymbol{\tau}_j)$ is the misalignment measure defined in Equation (46). We observe two differences compared to the expression for TFP in our baseline economy in Equation (28). First, as we have already discussed, firm choices reflect the revenue-based demand shifter δ_j^R as opposed to the original demand shifter δ_j in Equation (9). Second, the part of the social surplus that firms fail to capture from consumers still enters into aggregate consumption, and it is captured by the misalignment measure $\Delta(\boldsymbol{\tau}_j)$ present in the numerator. As the next proposition shows, both departures alter the aggregate behavior of the rent-extracting economy:

[22] Formally, all other parameters fixed, there is a continuum of rent-extracting economies indexed by $(\delta_L, \delta_H, \underline{\gamma}, \bar{\gamma})$, which coincide in firm-level choices $\{(x_j, n_j, \iota_j)\}_j$ and revenues $\{p_j \cdot y_j\}_j$. This continuum includes, in particular, the economy without asymmetric information, i.e., $0 = \underline{\gamma} < \bar{\gamma} = 1$, which as we will see coincides with our baseline economy.

Proposition 5. *Let $\mu^\star$ be the marginal type that is indifferent to producing information in equilibrium. Aggregate TFP and consumption in the rent-extracting economy then equal:*

$$\mathcal{A}^\star = \mathcal{E}\left(\mu^\star\right) \cdot \mathcal{A}(\mu^\star, g^R) \quad \text{and} \quad C^\star = \mathcal{A}^\star \cdot \left[N - \Phi\left(-\mu^\star \cdot \sqrt{\tau_\mu}\right) \cdot \chi\right], \tag{48}$$

in which the "macro-level wedge" satisfies:

$$\mathcal{E}(\mu^\star) \equiv \left[\Delta(\underline{\boldsymbol{\tau}})^{\frac{\theta-1}{\theta}} \cdot (1 - \zeta(\mu^\star)) + \Delta(\bar{\boldsymbol{\tau}})^{\frac{\theta-1}{\theta}} \cdot \zeta(\mu^\star)\right]^{\frac{\theta}{\theta-1}}, \tag{49}$$

where $\zeta(\mu^\star) \equiv \frac{g^R(\bar{\boldsymbol{\tau}})^{\theta-1} \cdot \xi(\mu^\star)}{g^R(\underline{\boldsymbol{\tau}})^{\theta-1} \cdot (1-\xi(\mu^\star)) + g^R(\bar{\boldsymbol{\tau}})^{\theta-1} \cdot \xi(\mu^\star)} \in (0,1)$ *is decreasing in* $\mu^\star$*,* $\mathcal{A}(\cdot)$ *and* $\xi(\cdot)$ *are defined in Lemma 1, and* $g^R(\cdot)$ *and* $\Delta(\cdot)$ *are defined by Equations (45) and (46), respectively.*

Proposition 5 reveals that aggregate TFP and welfare in the rent-extracting economy are affected by a macro-level wedge, $\mathcal{E}(\mu^\star)$, which is the aggregate counterpart to the firm-level measure of misalignment.[23] In the special case where firms are perfectly informed about consumer preferences—i.e., when $0 = \underline{\gamma}$ and $\bar{\gamma} = 1$—this misalignment disappears: $g^R(\cdot) = g(\cdot)$ and $\mathcal{E}(\cdot) = 1$, and the equilibrium mirrors that of the baseline economy. However, in the more general case, the macro-wedge $\mathcal{E}(\mu^\star)$ captures a pecuniary externality that emerges in equilibrium due to firms' rent-extracting behavior. Each firm makes its information, input, and pricing choices to maximize its private rent extraction from consumers. Yet, when all firms engage in such behavior, it becomes detrimental to all consumers, who, as the ultimate owners of the firms, would otherwise benefit from a more efficient allocation of resources.[24]

This pecuniary externality, in turn, provides a rationale for corrective policy interventions. But, before turning to policy interventions, we establish a paradoxical result that highlights the inefficiencies embedded in the laissez-faire equilibrium.

5.4 On Advances in Data-Processing Technologies: Redux

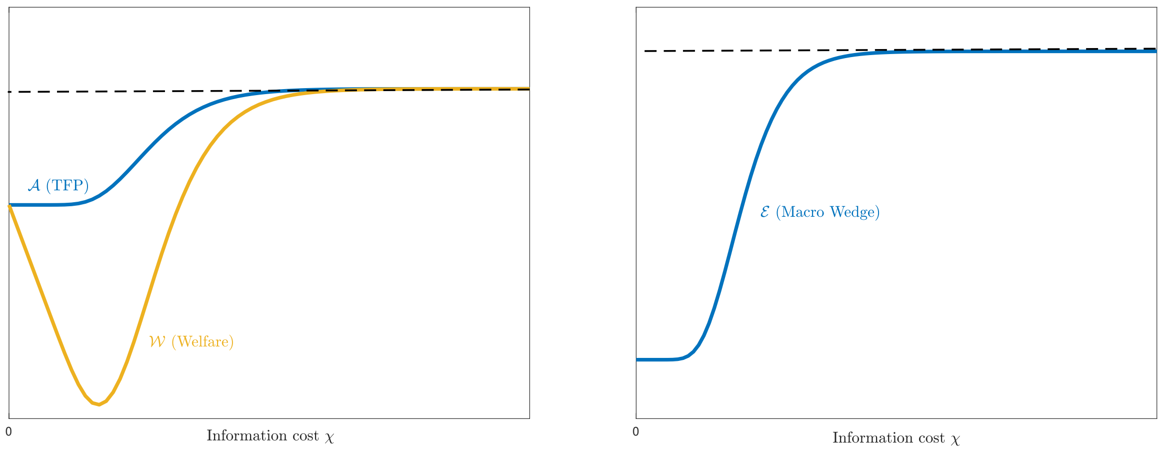
Advances in data-processing technologies—such as a reduction in the cost parameter χ or a rise in the precision $\bar{\tau}^\omega$—alter firms' ability to extract rents from consumers. As a result, these technological improvements have countervailing effects. This is in contrast to our baseline economy, in which advances in information processing were unambiguously beneficial.

Corollary 4. *An improvement in data-processing technologies, such as a fall in* χ*, or a rise in* $\bar{\tau}^\omega$*, leads to an increase in the share of information producers,* $\Phi(-\mu^\star \cdot \sqrt{\tau_\mu})$*, but has ambiguous*

[23]For instance, as $\mu^\star \to -\infty$ and all firms produce information, $\mathcal{E}(\mu^\star) \to \Delta(\bar{\boldsymbol{\tau}})$; instead, as $\mu^\star \to +\infty$ and no firm produces information, $\mathcal{E}(\mu^\star) \to \Delta(\underline{\boldsymbol{\tau}})$.

[24]Note that a related pecuniary externality emerges in monopolistic settings à la Dixit and Stiglitz (1977), when firms underproduce to extract rents from the representative consumer, who ultimately owns these firms. Crucial for our setting, however, the externality depends on the firms' information production.

Figure 6: Socially Destructive Improvements in Data-Processing Technologies



Note: The figure depicts the aggregate TFP, consumption and the macro-level wedge, as they depend on the information cost χ, for the parametric case when rent-extraction is socially destructive (see Definition 4).

effects on aggregate TFP, $\mathcal{A}^\star$, and consumer welfare, $C^\star$.

The counterintuitive nature of this result can be understood through two simple examples. First, consider the case in which firms have complete information about consumer preferences—that is, $\underline{\gamma} = 0$ and $\bar{\gamma} = 1$. As discussed earlier, in this situation, the allocations of the rent-extracting economy coincide with those of our baseline economy. Consequently, improvements in information-processing technologies—whether through reductions in χ or increases in $\bar{\tau}^\omega$—necessarily increase both TFP and welfare (Proposition 3).

Now, consider a different scenario in which information production provides little additional information about productivity, i.e., $\bar{\tau}^\upsilon \approx \underline{\tau}^\upsilon$. Suppose that the cost of information, χ, in this case falls from a prohibitively high level to zero, under conditions where price discrimination is *socially destructive*, i.e., $\delta^S(\bar{\gamma}) < \delta^S(\underline{\gamma})$ (Definition 4). In this setting, the misalignment is more pronounced among information-producing firms—i.e., $\Delta(\bar{\boldsymbol{\tau}}) < \Delta(\underline{\boldsymbol{\tau}})$—leading to a decline in the macro-level wedge, $\mathcal{E}$, as shown in Figure 6. As a result, the rise in the share of H-type consumers reduces overall social surplus, causing TFP and welfare to decline despite the fall in χ (Figure 6).[25] In this same scenario, a rise in $\bar{\tau}^\omega$ from a value close to $\underline{\tau}^\omega$ to a higher level can be shown to produce qualitatively similar effects to the decline in χ. These findings suggest that improvements in data-processing technologies can sometimes exacerbate inefficiencies from rent-extracting behavior, leading to potential welfare losses.

[25] Indeed, TFP is equal to $b \cdot \delta^S(\underline{\gamma})$ before the fall in χ, for some constant $b > 0$, and it declines to $b \cdot \delta^S(\bar{\gamma})$ after the fall. Since $C^\star = \mathcal{A}^\star \cdot N$ before and after the fall in χ, changes in consumption track changes in TFP.

5.5 Optimal Corrective Policy

We turn to the study of the normative properties of the economy. We consider the problem of a benevolent social planner who maximizes aggregate welfare, $\int_i \mathcal{U}_i di$, by making optimal production and consumption choices, $\{x_j, n_j, \iota_j\}_j$ and $\{c_{ij}\}_{ij}$. The planner operates under the same technological and informational constraints as agents in the decentralized economy. For each j, the timing of the planner's choices is as follows. In *Stage 1*, the planner chooses information production, ι_j, conditional on μ_j. In *Stage 2*, she chooses the variety-type, x_j, and employment, n_j, conditional on $(\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j)$. In *Stage 3*, the planner designs the mechanism to allocate consumption, (c_{Hj}, c_{Lj}), conditional on $(\mu_j, \upsilon_j, \omega_j)$. These choices are subject to the resource constraints: $\int_j (n_j + \chi \cdot \iota_j) \cdot dj \leq N$ and $\gamma_j \cdot c_{Hj} + (1 - \gamma_j) \cdot c_{Lj} \leq A_j \cdot n_j$ for all j.

We say that the decentralized equilibrium is *efficient* if its allocations coincide with those of the planner; otherwise, it is *inefficient*. We next state our first normative result:

Proposition 6. *The social planner's allocations coincide with those of our baseline economy. Hence, the laissez-faire equilibrium of the rent-extracting economy is generically inefficient.*

Proposition 6 establishes two key results. First, that the equilibrium of our baseline economy is efficient. This result aligns with the well-established normative properties of CES economies in the tradition of Dixit and Stiglitz (1977). It is well known that when factors of production are inelastic, such economies generally achieve efficiency despite the presence of market power. Importantly, the additional features of our framework—information production and preference heterogeneity—do not alter this fundamental property.

Second, a direct implication of Proposition 6 is that the inefficiency of the rent-extracting economy stems from the misalignment between profit maximization and social surplus maximization (Section 5.2). Indeed, if firms were to design trading mechanisms with the objective of maximizing social surplus rather than profits, the resulting allocations would coincide with those of the baseline economy and, consequently, be efficient.[26]

Overall, Proposition 6 serves as a useful normative benchmark. Yet, it also underscores a practical challenge in achieving the efficient outcome. Implementing this outcome may require *direct* policy interventions in firms' pricing strategies—an approach that may be difficult to implement in practice. Rather than pursuing such interventions, we focus below on more constrained (and arguably more realistic) policies. In particular, we explore interventions that target firms' information choices, which we broadly refer to as *data-regulation policies*.

Data-Regulation Policies. Suppose the only instrument at the planner's disposal is a tax

[26]Note that efficiency would generally not be attained by forcing firms to post unit prices. When the economy's factor supply is elastic, such a policy would simply replace distortions due to rent-extraction with more conventional distortions due to monopolistic markups.

$w \cdot T$ on information production, with proceeds rebated lump-sum to households.[27] Under this intervention, the only modification to the equilibrium conditions is in the determination of the marginal type $\mu^\star$ for a given market size, Ω: firms now face an information cost (in units of labor) of $\chi + T$ rather than χ. As a result, choosing T is equivalent to directly selecting $\mu^\star$. The next proposition characterizes the properties of optimal interventions.

Proposition 7. *The optimal tax, T, is positive (negative) if rent extraction is severe (mild).*

A simple perturbation argument helps illustrate this result. Consider a small change in the tax leading to a marginal shift in information production, $d\mu^\star$. Aggregate consumption and welfare are affected through two channels (Proposition 5):

- A change in the macro-wedge, $\mathcal{E}'(\mu^\star) \cdot d\mu^\star$.
- A change in residual welfare, $\frac{d\left(\mathcal{A}(\bar{\mu}, g^R)\cdot\left[N - \Phi(-\bar{\mu}\cdot\sqrt{\tau_\mu})\cdot\chi\right]\right)}{d\bar{\mu}}\Big|_{\bar{\mu}=\mu^\star} \cdot d\mu^\star$.

In equilibrium, the latter effect is fully internalized by firms, and is thus zero. The change in the macro-wedge then captures the non-internalized welfare effects of information production:

$$dC^\star = \mathcal{E}'(\mu^\star) \cdot \mathcal{A}^\star \cdot d\mu^\star \tag{50}$$

Thus, information production is excessive (insufficient) whenever $\mathcal{E}'(\cdot) > 0$ ($\mathcal{E}'(\cdot) < 0$), which holds if rent extraction is *severe (mild)*, as formalized in Definition 4. Recall that increases in $\mu^\star$ lower the share of information-producing firms.

When $\mathcal{E}'(\cdot) > 0$, by producing information and—through product customization— raising the share of H-type consumers they face, firms extract consumer surplus at a faster pace than they contribute to social welfare, leading to excessive information production. The optimal tax, therefore, discourages firms from information production in this case.

That said, a simple information tax may be too blunt an instrument. When rent extraction is so severe as to be *socially destructive* a conflict arises: information production provides firms with both socially valuable information (about productivity, υ_j) and socially harmful information (about demand, ω_j).[28] A more refined policy can better address this conflict.

Suppose that, in addition to the information tax, the planner can also "garble" the signals received by firms. Formally, suppose the planner increases the noise in the signals obtained through information production from $(\bar{\tau}^\upsilon, \bar{\tau}^\omega)$ to:

$$(z^\upsilon \cdot \underline{\tau}^\upsilon + (1 - z^\upsilon) \cdot \bar{\tau}^\upsilon, z^\omega \cdot \underline{\tau}^\omega + (1 - z^\omega) \cdot \bar{\tau}^\omega), \tag{51}$$

[27] The tax is equal to T units of labor.

[28] Note that the fact that the socially valuable information takes the form of signals about productivity is merely due to our modeling choices. As discussed in footnote 12, an isomorphic reformulation of our model replaces information about firm productivity with information about shocks to consumer preferences.

where $z^v, z^\omega \in [0,1]$ are policy parameters chosen by the planner. If given the choice, firms would never garble their own signals. Consequently, $(\bar{\tau}^v, \bar{\tau}^\omega)$ can be interpreted as the technological upper bounds on data-processing capabilities. While regulatory interventions cannot expand this technological frontier, they can limit firms' ability to exploit it. The following proposition demonstrates that a more targeted garbling policy is preferable when information about productivity and demand have conflicting effects on social surplus.

Proposition 8. *The planner never garbles information about productivity but garbles information about demand if rent extraction is socially destructive, i.e., in the case where* $\delta^S(\bar{\gamma}) < \delta^S(\underline{\gamma})$*, in which case* $z^\omega = 1$ *and* $T = 0$.

Intuitively, when information about demand is socially harmful, a policy of garbling signals about demand increases the macro-wedge, $\mathcal{E}$, to its upper bound, $\delta^S(\underline{\gamma})/\delta^R(\underline{\gamma})$. Once the planner sets $z^\omega = 1$, $\mathcal{E}$ becomes independent of information production, i.e., of $\mu^\star$. As seen in Equation (50), firms then internalize the social surplus generated by their information choices.

Broader Policy Lessons. The above data policies offer valuable insights into the ongoing debate about firms' access to and use of consumer data. A central concern among regulators is that unrestricted access to consumer data may enable firms to "misuse" personal information, potentially engaging in discriminatory practices. In response, the European Union's General Data Protection Regulation (GDPR) includes provisions aimed at curbing firms' ability to collect, store, and exploit consumer data in ways that may foster discriminatory behavior.

Although stylized, the interventions of Propositions 7 and 8 support the general aims of such regulatory measures, while also highlighting several caveats. First, as demonstrated in Proposition 7, the optimal regulation of data may involve *subsidizing* information production. This is because rent extraction is not always socially harmful. Indeed, it may incentivize firms to internalize a greater share of the social surplus generated by their choices, raising overall welfare. Second, as shown in Proposition 8, even when rent extraction is socially harmful, blanket restrictions on information production are an overly blunt instrument. In fact, regulatory measures that broadly limit access to consumer data are not optimal, and policies that selectively target socially harmful information sources are preferable.

Finally, while the preceding analysis focused on interventions targeting the economy's information structure—a logical approach given the central role of information frictions in equilibrium inefficiencies—a natural question arises: could more elaborate policies yield better outcomes? The answer is "yes". In Appendix C, we investigate richer second-best policies, which allow the social planner to target the externality more effectively by making data regulation also size dependent. Although these policies are firm-specific, making implementation more complex, we show that they exhibit several parallels to the above data-regulation policies.

An instructive case arises in the scenario in which rent extraction is severe, so that optimal data regulation reduces firms' incentives to produce information by imposing a tax (Proposition 7). We show that a more targeted policy—besides achieving this objective—also reduces the firm-size concentration. This is because, in this scenario, in the laissez-faire equilibrium, large firms are excessively large relative to the size that is warranted by their contribution to social surplus. This result contrasts sharply with the findings from the literature that explores the welfare costs of rising market power (e.g., Edmond *et al.*, 2023; Boar and Midrigan, 2024; Eeckhout *et al.*, 2024). A common theme in this work is that large firms—those that have higher market power and markups—are inefficiently small, precisely because they constrain production to raise prices. By contrast, in our rent-extracting economy, firms do not need to constrain production to extract surplus; they can do so through price discrimination instead. As a consequence, our results show that large firms are indeed *too large*.

6 Supporting Evidence and Quantification

We have demonstrated how advances in data-processing technologies affect the economy by enabling firms to optimize their (i) scale of operations, (ii) product choice, and (iii) pricing strategies. Our results show that improvements in data-processing can, in general, be either *beneficial* or *detrimental* to economic efficiency and welfare. In this section, we use our analysis to shed light on the quantitative implications that advances in data-processing have had on U.S. economy over the past two decades. Has the documented increase in firm informativeness, as presented in Section 2, contributed to gains in productivity and welfare? And if so, to what extent, and through which channels? To answer these questions, we proceed to calibrate and quantify our rent-extracting economy using our data and empirical findings from Section 2.

6.1 Model Validation and Parametrization

Before we use our model framework as a quantitative laboratory, we provide a first-pass validation of our theory. We document that the cross-sectional predictions of our model are both qualitatively and quantitatively consistent with salient features of US firm-level data.

Qualitative Validation: We qualitatively validate the cross-sectional predictions of our framework, discussed in detail in Section 4.1 and summarized in Corollary 3. Recall that the *baseline economy* and the *rent-extracting economy* are observationally equivalent when using data on firm-level revenues and input choices (Proposition 4 and footnote 22).

1) Information and Resource Allocation: First, a central role of information within our framework is to improve the allocation of inputs *within* a firm. Indeed, (revenue-based) measured tfp is, all else equal, higher and less volatile for more informed firms, as firms with more precise demand-side information are better able to tailor their products to consumer preferences, thereby increasing and stabilizing their effective demand. Consistent with this prediction, Panel (a) of Figure D.1 shows a pronounced negative relationship between a firm's squared revenue error and its measured tfp, even after controlling for firm age and sector. In both the data and the model, more accurate firms are, all else equal, more productive.[29] Figure D.3 further shows that more accurate firms also exhibit lower volatility in measured tfp.

Second, another key role information has within our framework is in reducing (ex-post) factor misallocation *across* firms. It does so by enabling firms to better align their input choices with realized productivity shocks. Panel (b) of Figure D.1 reports the average cross-sectional dispersion in marginal revenue products, with estimates shown separately for informed and uninformed firms. We define an "informed" firm as one that (i) falls below the median in the mean squared error of its one-year-ahead revenue expectations, and (ii) has at least three observations. Appendix D.1 provides further details on the computation. Across multiple measures of misallocation, informed firms consistently display lower cross-sectional dispersion, consistent with the idea that more accurate firms are better able to anticipate their productivity and demand conditions.[30] While it is likely that distortions unrelated to information—such as those documented by David and Venkateswaran (2019)—account for a significant share of the observed dispersion, the evidence in Panel (b) nevertheless suggests that accuracy improvements are systematically associated with a more efficient input allocation across firms.

2) Information and Firm-Size Distribution: Finally, as stressed above, information in our framework also has a close relationship to firm size. As highlighted by the empirical evidence in Section 2 (Figure 2), larger firms are, all else equal, more likely to be informed. At the same time, as discussed above, the production of information itself also leads firms to expand, by improving the efficiency of resource allocation. Our framework thus generates a tight, two-directional relationship between a firm's informativeness and its size. Table A.13 in the Appendix leverages the panel structure of our dataset to further show that more accurate firms at time t subsequently experience faster growth, in line with our model's predictions.

In summary, the main cross-sectional predictions of our theory, as formalized in Corollary 3, are qualitatively consistent with observed firm-level patterns. We next turn to evaluating the quantitative fit between model and data, with particular emphasis on the two-directional

[29] We estimate (revenue-based) firm-level tfp following Ottonello and Winberry (2020), modifying the approach to account for a finite elasticity of substitution. This estimation is consistent with an extension of our framework that incorporates capital in production (Section 6.3).

[30] The striking 30–40 percent difference in cross-sectional dispersion observed in Panel (b) of Figure D.1 is robust to alternative definitions of "informed" and "uninformed" firms (e.g., using the top quartile of the error distribution), alternative industry classifications (e.g., six-digit NAICS codes), and is consistent over time.

Table II: Parametrization: Model vs. Data (2002-2007)

	Data	Model
Mean of log-productivity of informed firms	0.033	0.037
Mean of log-productivity of uninformed firms	0.015	0.015
Variance of log-productivity of informed firms	0.058	0.056
Variance of log-productivity of uninformed firms	0.108	0.131
Unconditional variance of log-productivity	0.097	0.123
Root-mean-squared error of informed firms	0.036	0.036
Root-mean-squared error of uninformed firms	0.132	0.132
Share of information producing firms	0.100	0.100

Note: The table compares data moments from I/B/E/S-Compustat sample over the period 2002-2007 to those from the calibrated model. The table shows the mean and variance of log-productivity, in addition to the root-mean-squared-error of firms' one-year-ahead log-revenue forecasts. Firm productivity is estimated as in Ottonello and Winberry (2020). We define an informed firm as a firm that is in the bottom 10 percent of the mean-squared-error distribution over the initial period and for which we have at least 3 observations.

relationship between accuracy and firm size. This requires us to first parameterize the model.

Model Calibration: The aim of our parameterization is to ensure that the model quantitatively replicates key features of firm-level outcomes and captures the rich heterogeneity in expectations documented in the data. To this end, we set the elasticity of substitution between goods, θ, equal to 3, following e.g., Hsieh and Klenow (2009), normalize the labor endowment, N, to 1, and calibrate the remaining parameters internally.

As we discussed in Section 5.3 (and footnote 22), there is a continuum of rent-extracting economies—characterized by the distribution of H-types (i.e., $\bar{\gamma}$ and $\underline{\gamma}$) and the associated demand shifters (i.e., $\delta_H = \exp(\widehat{\delta})$ and $\delta_L = \exp(-\widehat{\delta})$)—that yield identical firm-level moments.[31] This continuum includes, as a special case, an economy without asymmetric information (i.e., with $0 = \underline{\gamma} < \bar{\gamma} = 1$), whose allocations coincide with those of our baseline model. For concreteness, we focus here on a rent-extracting economy where $\bar{\gamma} = 0 < 1 = \underline{\gamma}$. We then leverage the indeterminacy among different combinations of $(\widehat{\delta}, \underline{\gamma}, \bar{\gamma})$ to construct sharp bounds on the costs and benefits associated with advances in data-processing technologies. In Section 6.3, we further illustrate how product-level evidence can help resolve this indeterminacy.

We calibrate the variance of ex-ante log-productivity, τ_μ^{-1}, the variance of productivity shocks, τ_a^{-1}, and the symmetric demand shifter, $\widehat{\delta}$, to match the unconditional variance of log-productivity observed over the first five years of our sample (2002–2007). These parameters are also used to match the conditional variance of productivity for *informed* and *uninformed* firms over the same period. Consistent with evidence from Brynjolfsson and McElheran (2016)

[31] The assumption of symmetric taste shifters serves as a normalization, since the overall demand level for a firm's product does not affect our identification strategy or quantitative results.

Table III: Size and Accuracy Relationship

	log. squared errors	
	Data	Model
Size (labor)	-0.454^{***}	-0.480^{***}
	(0.052)	(0.001)
Time (years)	0.007	-0.089^{***}
	(0.007)	(0.002)

Note: Least-squares estimates of the relationship between squared normalized errors and firm size (quintiles of the initial employment distribution). We estimate this relationship both in the I/B/E/S-Compustat data and in the calibrated model (Section 2). The column labelled data further controls for sector fixed effects and age (Column (2) in Table I). Robust (clustered) standard errors in parentheses. *p<0.1,**p<0.05,***p<0.01.

on the adoption of data-driven decision-making, we conservatively assume that 10 percent of firms are informed at baseline.[32] The information cost parameter, χ, is calibrated to target this share. Lastly, we set the precision of firms' information, $\underline{\boldsymbol{\tau}}$ and $\bar{\boldsymbol{\tau}}$, to match the mean squared error of revenue expectations for informed and uninformed firms, respectively. Table II presents the match between model-implied and data moments, while Table D.1 in the Appendix summarizes the calibrated parameter values.

Quantitative Validation: A central implication of our theory is the tight relationship between firms' information and size. Above, we established the *qualitative* consistency of this relationship. We now use our calibrated model to also *quantitatively* assess our framework against the motivating evidence presented in Section 2. To do so, we simulate expectation errors over a 20-year period, linearly reducing the information cost parameter, χ, to replicate the 41 percent increase in average accuracy documented earlier. Concurrently, we linearly increase $\bar{\boldsymbol{\tau}}$ to reflect observed improvements in data processing and reductions in productivity volatility among information-producing firms. The number of firm-year observations is set to match that in the data. Table III and Figure D.3 illustrate how information choices translate into a monotonic relationship between the accuracy of firm expectations and firm size.[33]

Although our calibration does not explicitly target the size–accuracy relationship, the model nevertheless generates a relationship that closely mirrors that in the data. The regression coefficient on size, measured by a firm's employment quintile, is 0.480 in the model and 0.454 in the data (Table III). In both the model and the data, an increase in firm size by one quintile reduces the squared error by approximately 45–50 percent of its average value. Furthermore, the model replicates the overall slope of the size–accuracy relationship reason-

[32] We define an informed firm as one in the bottom 10% of the squared-error distribution.

[33] To replicate the patterns in the data, we estimate that the cost of information production, χ, declined by 42%, while the precisions of produced information, $\bar{\tau}^{\upsilon}$ and $\bar{\tau}^{\omega}$, increased by 10% and 2%, respectively.

ably well (Figure D.3), although the model somewhat overstates the improvement in accuracy attainable when moving from the fourth to the fifth quintile.

Crucially, our calibrated model also replicates key patterns in the evolution of data use and firm size over time. While precise estimates of firms' adoption of data-processing technologies remain difficult to obtain (Baley and Veldkamp, 2025), recent evidence from Brynjolfsson and McElheran (2024) estimates that approximately 73% of medium-to-large manufacturing firms systematically used data to inform their decision-making by 2021. Our model slightly understates this figure, predicting that 72% of firms were "information producers" by the end of 2022. Nonetheless, the magnitude and trajectory of this increase are remarkably similar, particularly given our model's stylized structure. Lastly, the model also captures the rise in the employment share of large firms over the sample period. In particular, the share of total employment accounted for by firms above the 80th percentile of the employment distribution increases by 15%-points in the model, compared to an increase of approximately 6–7%-points in the data. While this suggests the model somewhat overstates the overall shift, the direction and qualitative pattern are consistent with the empirical trend.

In summary, the results in this subsection have shown that our framework both *qualitatively* and *quantitatively* captures salient relationships between the accuracy of firms' expectations and firm-level outcomes. We conclude that our model offers a suitable laboratory to explore the consequences that the evolution in firms' informativeness has on the macroeconomy.

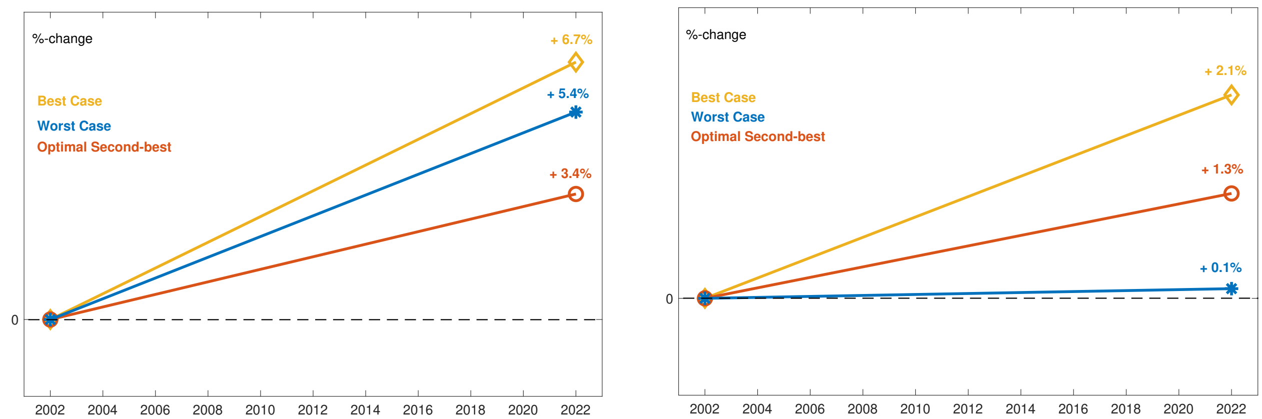
6.2 Quantification and Decomposition

We quantify the aggregate effects of the estimated declines in information costs, χ, and increases in information precision, $\bar{\boldsymbol{\tau}}$. Due to the indeterminacy arising from the severity of distortions associated with price discrimination, we compute these effects in two benchmark scenarios. Specifically, we search over all combinations of the parameters $(\widehat{\delta}, \underline{\gamma}, \bar{\gamma})$ that match the targeted model moments and identify the parameterizations that yield the highest and lowest welfare outcomes. We refer to these as the "best-case" and "worst-case" economies, respectively. Together, the two scenarios bound the overall effects of advances in data-processing technologies. The best-case economy corresponds to a rent-extracting environment without asymmetric information—i.e., $0 = \underline{\gamma} < \bar{\gamma} = 1$—which, as previously discussed, yields outcomes identical to those in our baseline economy. Figure 7 and Table IV summarize the results.

On balance, the estimated effects of improvements in data-processing technologies are substantial. Over the past two decades, total factor productivity (TFP) gains attributable to these improvements are estimated to be between 5.4% and 6.7% (Panel (a) of Figure 7). For comparison, TFP in the data has increased by approximately 15% over our sample period.[34]

[34]See, e.g., the FRED series on TFP (`https://fred.stlouisfed.org/series/RTFPNAUSA632NRUG`).

Figure 7: Effects of Improvements in Data-Processing Technologies



Panel (a): TFP   Panel (b): Welfare

Note: Panel (a) shows the estimated effects for different calibrated scenarios on TFP, $\mathcal{A}$. Panel (b) depicts the results for overall welfare, C. We showcase results for the "best-case" scenario, i.e., the model with $\underline{\gamma} = 0$ and $\bar{\gamma} = 1$, and for the "worst-case" scenario, which is estimated to feature $\underline{\gamma} = 0.71$ and $\bar{\gamma} = 0.87$. We also illustrate the results for the "worst-case" scenario under the optimal corrective policy described in Section 5.5.

Overall, our estimates account for nearly two-thirds of the potential TFP gains that could be realized through the complete removal of information frictions.[35] Although there are clear uncertainties associated with our first-pass estimates, taken at face value, our results simultaneously suggest that (i) improvements in firms' use of data have been a major contributor to the observed rise in TFP over the past two decades; and (ii) perhaps more concerningly, much of the potential productivity gains from improved data use may have already been realized.

Table IV decomposes the overall rise in TFP into its three constituent channels, illustrating a main strength of our framework.[36] The results document that the dominant contributor to TFP growth over the sample period is firms' improved ability to determine their optimal scale of operations—accounting for more than three-quarters of the total increase. In contrast, firms' enhanced capacity to tailor products to consumer preferences contributed a more modest 1.6%-points to TFP. Crucially, notice that, in the worst-case scenario, the gains from improved product design are almost entirely offset by firms' distorted incentives to produce information to extract rents from consumers. We estimate the resulting drag on TFP from this channel to be as much as –1.5%-points.

[35] We compute the later estimates by setting the cost information production to zero and increasing the accuracy towards infinity and one, for tfp and demand-side information, respectively. We then re-compute relevant variables in the economy without information frictions and compare our results. The increase in TFP from removing information frictions is 6.5% and 8.9% in the two cases, respectively.

[36] This decomposition follows directly from Proposition 5.

Table IV: Decomposition of The Rise in TFP

Model	Overall (%)	Scale (pp)	Product (pp)	Pricing (pp)	$\underline{\gamma}$	$\bar{\gamma}$
Best case	6.7	5.1	1.6	0	0.00	1.00
Worst case	5.4	5.1	1.6	-1.5	0.71	0.87

Note: The table decomposes the rise in TFP in Figure 7 into its three constituent channels: (i) scale, (ii) product design, and (iii) pricing. The table does so for the "best case" and for the "worst case" scenarios.

Although the overall rise in TFP is reasonably similar across the two scenarios (6.7% vs 5.4%) , the estimated welfare consequences differ markedly. Panel (b) of Figure 7 shows that, in the best-case scenario, economy-wide welfare increased by approximately 2.1% as a result of advances in data-processing technologies. Instead, under the worst-case scenario, these potential welfare gains are almost entirely offset by firms' excessive information production. Indeed, the worst-case estimates suggest that the welfare benefits from advances in data processing over the past two decades have been negligible (0.1%). While the worst-case scenario may clearly overstate the adverse consequences of rent extraction (see, however, Section 6.3), our results nevertheless demonstrate that a substantial disconnect is quantitatively possible between TFP gains resulting from advances in data-processing technologies and their welfare consequences. In particular, once one accounts for the endogenous changes in firms' pricing behavior, the welfare effects of such technological changes may be significantly diminished—even in the presence of sizable productivity gains.

This disconnect raises the question of the potential welfare improvements that corrective policy can deliver. Table IV implies that, in the worst-case scenario, the economy operates in the parameter region where rent extraction is *severe* but *not socially destructive* (Definition 4). Consequently, as demonstrated in Section 5.5, there is a clear rationale for corrective data regulation that discourages information production by firms. Panel (b) of Figure 7 shows that the optimal second-best policy increases welfare gains from 0.1% to 1.3% (Section 5.5 and the formal characterization in Appendix C).[37] In addition to curbing information production (by around half), the policy reduces *firm-size concentration*: we estimate that it halves the rise in the employment share of large firms relative to the no-policy baseline. If the planner were instead restricted to implementing a simple information tax, as in Proposition 7, our estimates imply that welfare gains would increase from 0.1% to 1.1%—capturing the majority of the total welfare improvements achievable through policy.

We therefore conclude that limiting information production—particularly by large firms—

[37] The remainder of the lost welfare gains are due to misallocation of goods across consumer types, arising from discriminatory pricing strategies, which the second-best policy is unable to correct; see Section 5.5.

could recover approximately 50% of the potential welfare benefits from advances in data-processing technologies (i.e., 1.1–1.3 percentage points out of a total of 2.1%).

6.3 Quantitative Refinements

Capital and Variety Accumulation. A potential concern with the above estimates is that they may represent a *conservative* range. The economy considered so far features both a fixed supply of production factors and a fixed set of varieties, both of which limit the economy's reaction to improvements in data processing. To explore the implications in a more flexible environment, we extend the rent-extracting economy to include capital accumulation and endogenous variety entry. Appendix D.6 presents the extended model and focuses, for tractability, on the steady state. Figure 8 contrasts the results from advances in data-processing technologies to those highlighted in Figure 7, using the same calibration strategy.[38]

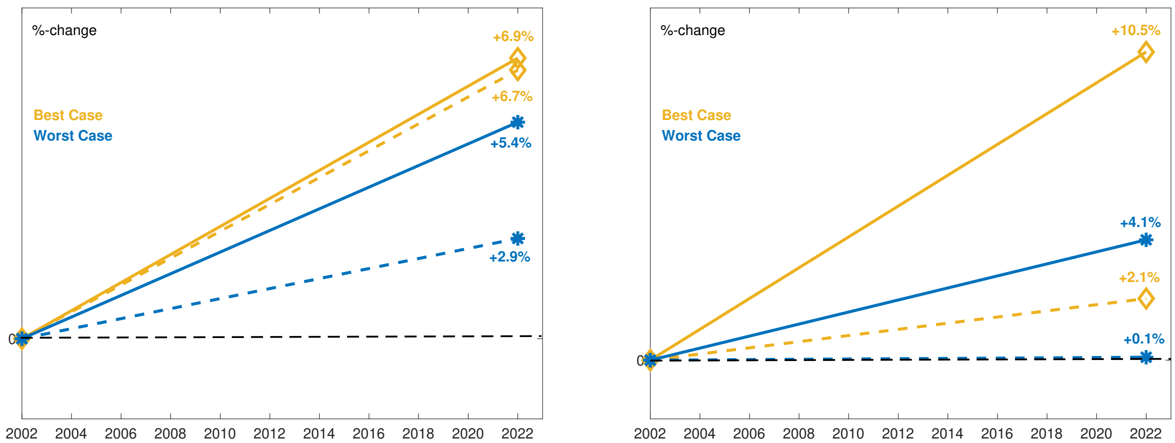
Relative to our earlier estimates, the gap between the best- and worst-case scenarios widens in the extended model. While the estimated TFP increase in the best-case scenario remains similar (6.9% compared to 6.7%), the TFP gain in the worst-case scenario is notably smaller—2.9% versus 5.4% previously. As shown in Table D.2 in the Appendix, this discrepancy is primarily driven by two factors: a modest increase in the TFP drag from rent extraction and a significant decline in the number of product varieties. The decline in varieties occur because rent extraction reduces the effective market size of firms, thereby discouraging entry and leading to the decumulation of varieties. Over time, this dynamic further depresses TFP. In this way, the reduction in variety magnifies the adverse consequences of rent extraction.

Finally, although the notion of welfare is more complex in this dynamic extension—particularly given that the transition to a new steady state may take considerable time—Figure 8 shows that the estimated increase in steady-state consumption is now larger. This reflects the economy's increased ability to accumulate production factors (i.e., capital), which amplifies the consumption response to improvements in data-processing technologies. Importantly, however, a sizable gap between the best- and worst-case scenarios persists, underscoring once again the potential welfare gains from effective corrective data regulation.

Re-calibrated Productivity Parameters. Our baseline estimates assess the effects of advances in data-processing technologies while holding constant the underlying process for firm-level productivity (Equation 4). However, due to the two-sided relationship between firm size and firm information in our framework, a potential confounder is that the productivity

[38] In the extended model, we estimate that the cost of information production, χ, has fallen by 33%, while the upper-accuracy parameters, $\bar{\tau}^{\upsilon}$ and $\bar{\tau}^{\omega}$, have increased by 10% and 2%, respectively. We discuss the calibration of the extended model in Appendix D.6.2. Notice that the fall in the information cost, χ, is smaller than before. This follows the increased "elasticity" of the extended model to increases in information.

Figure 8: Capital and Variety Accumulation



Panel (a): TFP

Panel (b): Welfare

Note: Panel (a) shows the estimated effects for different calibrated scenarios on TFP, $\mathcal{A}$. Solid lines indicate results for the augmented quantification, while dashed lines indicate those from Figure 7. Panel (b) shows the results for overall consumption/welfare, C. We showcase results for the "best-case" scenario, i.e., the model with $\underline{\gamma} = 0$ and $\bar{\gamma} = 1$, and for the "worst-case" scenario in each case.

process itself may have evolved over time—thereby inducing changes in information production and expectations accuracy even absent technological progress in data processing. To explore this possibility, Table D.3 in the Appendix presents results from an alternative calibration that, in addition to updating the information parameters, also re-calibrates the productivity process (τ_μ, τ_a).[39] While the estimated gains in TFP and welfare are somewhat larger under this extended calibration, the contribution of data-processing technologies remains similar to before: TFP increases are in the range of 5–7%, underscoring the robustness of our findings.

Estimates from the Retail Sector. Finally, our baseline estimates bound the potential costs and benefits associated with advances in data-processing technologies. A natural question that arises from our approach is: where does the U.S. economy likely fall within this estimated range? The answer depends on the severity of distortions introduced by price discrimination—as captured in our model by the extent of information asymmetries between firms and consumers. Measuring these distortions requires detailed product-level transaction data across a broad set of goods and sectors in the U.S. economy. Although such an exercise lies beyond the scope of this paper, we offer a preliminary attempt to address this question.

First, we adopt the view that price discrimination occurs primarily through quantity

[39]Specifically, we recalibrate (τ_μ, τ_a) at the end of the sample period to match the same firm-level moments targeted at the start of the sample. We use averages over the final three years as targets.

Table V: Non-linear Pricing in the Retail Sector

	Bornstein-Peter		*Model*	
	(1)	(2)	Best case	Worst case
Log quantity	-0.60	-0.39	0.0	-0.34
	(0.0001)	(0.0001)	(0.02)	(0.02)
Sample	All	Expensive	All	All
Fixed effects	✓	✓	✓	✓
Observations	88.3M	4.5M	826 x2	826 x2

Note: The table compares estimates using the Nielsen Scanner Data from Bornstein and Peter (2024) of the relationship between the log of the average payment for individual retail items, t_j^l/q_j^l, and the associated log of quantities, q_j^l, to those from our model, controlling for product-line and store (firm) fixed effects. The number of firms in the model equals that in the I/B/E/S-Compustat sample.

bundling, the formulation that we have adopted in the main text (see Appendix C for alternatives). Second, we draw on estimates from Bornstein and Peter (2024), who use Nielsen Retail Scanner Data from October 2017 to estimate the relationship between the log of average payments for individual retail items, t_j^l/q_j^l, and the corresponding log of quantities, q_j^l. Table V presents their estimates and compares them to those from our calibrated model.

While our estimates align with those observed for higher-end, more expensive products, the degree of pricing non-linearity in the data exceeds even that implied by the worst-case scenario in our rent-extracting economy.[40] This finding echoes the conclusion in Bornstein and Peter (2024), who shows that simple models of price discrimination struggle to account for the magnitude of observed non-linearities in retail pricing data. Of course, the aggregate economy encompasses many sectors beyond retail, and per-unit price variation may arise from factors other than heterogeneous consumer preferences. Nonetheless, the above estimates suggest that distortions stemming from price discrimination could be substantial—and that, all else equal, the economy may be closer to the worst-case, severe rent-extraction scenario.

In summary, in this section, we have employed the rent-extracting economy as a quantitative laboratory to bound the potential costs and benefits of advances in data-processing. Calibrating our model to firm-level developments, we estimate that such technological changes have led to a significant increase in TFP over the past two decades (c. 3.0-7.0%), primarily by helping firms optimize their scale of operations (c. leading to a 5pp rise in TFP). Yet, the

[40]In the best-case scenario, the estimated coefficient is zero. Under CES preferences, the equalization of marginal utilities across consumer types implies the equalization of average utilities, which is what the monopolist extracts from each consumer absent information asymmetries. By contrast, in the worst-case scenario, information asymmetries imply that the H-type has a lower marginal utility of consumption than the L-type (see Section 5.1). Coupled with the fact that the H-type consumer earns information rents, this implies that she pays a lower per-unit price, which is reflected in the estimated coefficient being less than zero.

welfare benefits of these changes appear modest—often in the range of only a couple of percent of steady-state consumption. The main reason for the potential lackluster welfare benefits is the excessive information production by firms, highlighting the importance of regulation to mitigate distortions caused by data-driven price discrimination.

7 Conclusion

Advances in data-processing technologies hold the potential to reshape many dimensions of economic life. This paper has focused on one such dimension: the ability of these technologies to enhance firms' information about economic fundamentals. Using micro data on managerial forecasts, we documented a systematic rise in the accuracy of U.S. firms expectations over the past two decades and showed that this increase is closely linked to shifts in the firm-size distribution. To assess the macroeconomic implications of this trend, we developed a quantitative-theoretical framework in which improved information enables firms to optimize their *scale*, *product choice*, and *pricing strategies*.

Consistent with the data, our model predicts that firms leveraging information more effectively allocate inputs more efficiently, design better products, achieve higher profitability, and grow faster and larger. We next quantified the aggregate consequences of these micro-level shifts. For plausible parameters, in line with the observed increase in accuracy, our estimates suggest that in the absence of this informational improvement, TFP and household welfare in 2022 would have been 5.3–6.7% and 0.1–2.1% lower, respectively. We decomposed these overall effects and found that the bulk of the TFP gains—approximately two-thirds to three-quarters—stem from firms' improved ability to determine their optimal scale, with product improvements contributing a smaller, though nontrivial, share.

That said, our analysis also reveals that much of the welfare gain from higher TFP may have been offset by excessive information production, as firms increasingly spend resources to engage in discriminatory pricing. This finding underscores that, without appropriate regulation, improvements in data-processing technologies may yield limited benefits. The design of an appropriate data regulation is central to ensuring that technological advances in data-processing translate into broad-based welfare improvements.

Our framework opens several avenues for future research. A particularly promising direction is to incorporate household-side responses, i.e., how consumers use data to navigate product space and how this behavior interacts with firms' strategic decisions. Another valuable extension would leverage product-level transaction data across a broader set of sectors to refine empirical estimates of the consumer-side costs associated with data-driven firm behavior. While this paper offers an initial step using firm-level data, much remains to be understood regarding the broader welfare implications of the data revolution.

A Motivating Evidence

A.1 Data Construction: I/B/E/S-Compustat

In our main analysis of firms' expectations, we use a combination of the I/B/E/S managerial guidance database and Compustat Fundamentals Annual. The combined sample for the I/B/E/S-Compustat merger covers the period 2002-2022 for 12,917 firm-years spanning 2,570 US firms. To construct our sample, we follow convention and discard utilities and financials, as well as any firm-years that have negative or non-existing values for revenue, employment, and/or the capital stock. We focus on revenue expectations, which comprise the lion's share of all forecasts provided by managers, and analyze one-year ahead annual expectations.[41] We only use "centered forecasts": that is, either point estimates or forecasts that are stated as a range. In the latter case, we use the mid-point of the range as the point estimate. We remove observations that are related to the top and bottom 1 percent of the error distribution.

Variable Definitions: We use the following variables from Compustat Fundamentals Annual: revenue (code: `sale`), profits (code: `ib`), capital (code: `ppent, ppegt`), investment (code: `capx`), assets (code: `at`), employment (code: `emp`), mergers and acquisitions (code: aqa), and industry classification (code: `naics` and `sic`). We measure a firm's debt as the total net value of liabilities (code: `dllt+dlc-che`) and the stock of acquired intangibles, adjusted for amortization and finanical goodwill as, following Chiavari and Goraya (2023) (code: `ITAN+AM-GDWL`). We deflate nominal variables where appropriate with US CPI (code: `CPIAUCLS` from FRED) and compute (revenue-based) total factor productivity as in Ottonello and Winberry (2020), adjusting for a finite constant degree of elasticity of substitution between goods. Finally, Compustat only has limited data on wages. We use total labor and related expenses as our measure of the overall wage bill (code: `xlr`). We link the Compustat data with the I/B/E/S database using the CRSP ID that is available for both. The annual expectations employed from the I/B/E/S managerial guidance database (code: `val1` and `val2`) are those that pertain to "centered forecasts" (code: `fdesc=1,2`) in millions or billions of USD. We mainly study expectations of future revenue (code: `measure=SAL`), although we also consider expectations of future profits (code: `measure=NET`) and capital expenditures (code: `measure=`cpx). We define a firm's error as the difference between the realized value of the variable from Compustat and the one-year-ahead expectation of the variable from I/B/E/S.

Descriptive Statistics: Table A.1 reports descriptive statistics for our merged data set.

[41] For an individual firm, we study the first forecast made in the year (Jan-April) that pertains to the firm's end-of-year financial results. Firms mainly report previous year's financial results in Q1 of the following year.

Table A.1: Descriptive statistics: I/B/E/S-Compustat

Variable Name	Obs.	Mean	Std.	Median
Revenue	12,917	3,762	11,518	769.49
Profits	12,917	279.30	1,403	28.98
Capital	12,825	1,052	4,692	122.79
Investment	12,910	184.13	857.47	46.70
Wages	845	1,841	3,974	353.89
Assets	12,917	5,720	22,152	1012.85
Employment	12,835	12.90	32.42	2.90
Revenue/capital	12,821	13.23	32.16	6.97
Expectation/capital	12,821	14.71	51.24	7.00
Expectation Log	12,821	1.94	1.13	1.95
Expectation Error	12,567	-0.14	1.91	0.01
Expectation Error Log	12,563	-0.01	0.13	0.00

Notes: The table reports descriptive statistics for the sample of 2,570 firms from 2002-2022 in the merged Compustat-I/B/E/S database. The units of the first seven rows are USD millions. The employment row is in '000-employees. The first eight rows capture, respectively, firm revenue, GAAP net-profits, book value of the capital stock, total value of capital expenditures, end-of-period total liabilities and assets, overall expenditures to labor and related expenses, and the total number of employees. The next three rows measure revenue scaled by a firm's tangible capital and the (log) of the year-ahead expectation. The final two rows are for the year-ahead error defined as realized future (log)-revenue minus (log of) the expectation. In the final two rows, observations have been removed that are in top and bottom 1 percent of the error distribution.

A.2 Data Construction: Duke-Richmond Fed CFO Survey

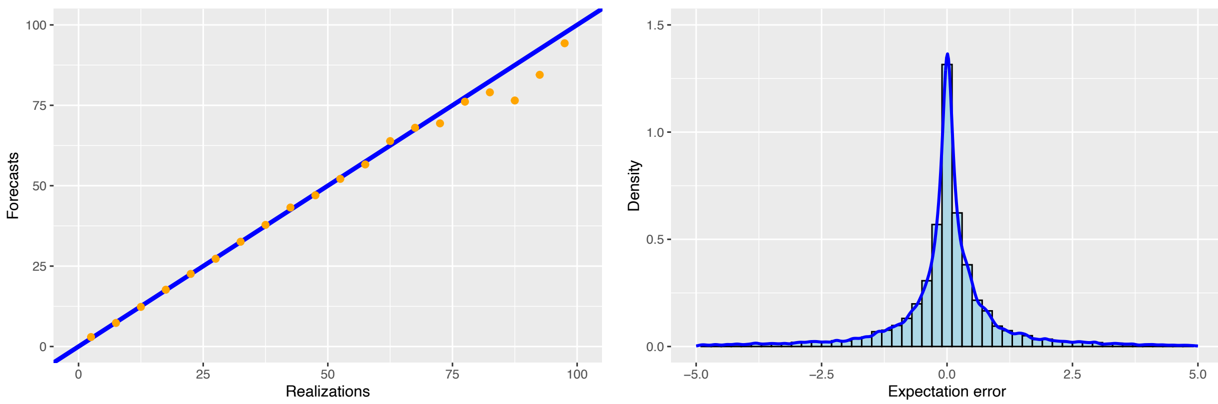
The CFO Survey is a quarterly survey of U.S. business leaders designed to elicit the financial outlook for their firms, the economic challenges they face, and their expectations about the broader U.S. economy. We exploit a combination of survey answers from The CFO Survey and data on economy-wide outcomes from FRED. The sample covers the period 2020-2022, the period for which data is available, for 3,470 firm-years spanning 826 U.S. firms. We remove expectations that are not one-year-ahead, as well as any firm-years that have non-existing profits. We throughout focus on annualized real GDP growth expectations (code: `GDPC1`). Firm size is measured by the number of domestic full-time employees. The size buckets used below correspond to quintiles of the 2020 size distribution: `size=1` (fewer than 6 employees); `size=2` (6-40 employees); `size=3` (40-130 employees); `size=4` (130-500 employees); and `size=5` ($>$ 500 employees). The familiarity with the concept of "Gross Domestic Product" (GDP) is measured numerically with a scale from 1-3.

A.3 Additional Data Comments

In this appendix, we provide a brief overview of the relationship between firms' revenue expectations and their input choices in the I/B/E/S-Compustat database. First, notice that firms' revenue errors are close to unbiased. Table A.1 shows that the mean error of firms' log revenue errors is, for example, -0.01 compared to an average value of log revenue of 1.94. Figure A.1 provides a full bin-scatter plot of firms' revenue expectations and realizations, and shows that the associated error distribution is also close to symmetric. These results are consistent with the evidence in, e.g., Chen *et al.* (2023), who show that Japanese firms' expectations about their own sales, in addition to their expectations about macroeconomic and sector-specific inflation rates, are close to unbiased. All else equal, firms in our sample do not appear to systematically skew their revenue expectations one way or another. Relatedly, Chen *et al.* (2024) explore how positive and negative E/P/S revisions respond to new information in the I/B/E/S-Compustat database, finding also a consistent pattern. Second, Table A.2 conducts an exercise akin to that in Tanaka *et al.* (2020). Panel (a) documents the relationship between the realized growth in inputs and the current-period (firm-specific) expectation of *future revenue*. We find that, all else equal, more optimistic firms employ and investment more, consistent with these firms being viewed as more optimistic. The estimated effect sizes are, furthermore, substantial: a 1 percent increase in expected revenue is associated with 0.14 percent increase in investment, for example. In Section 6.1 in the main text, we discuss how the *accuracy* of firms' expectations, in addition to their level, systematically affect firms' input choices, consistent with our model framework. Panel (b) in Table A.2 instead explores the relationship between the realized growth in various inputs and the *previous period's revenue error*. We find that firms that have been positively surprised—i.e., have higher revenue than previously expected, and hence a positive revenue error—subsequently employ and invest more, in line with these firms being genuinely surprised about the revenue realization. We conclude that the results in Table A.2 are consistent with the those in Tanaka *et al.* (2020), among others, who show that firms who are more optimistic about the future invest and employ more, and that positive profit (or revenue) surprises result in more inputs being employed and allocated subsequently.

Finally, we note that the arguments we provide in the main text do not require that firms' reported expectations strictly equal their (correct) mathematical expectation of future revenue. We do not require the complete absence of strategic or behavioral drivers of expectations. We only require that changes in reported expectations (and in their accuracy) in part reflect changes in information. The results in Tables A.1 and A.2 are consistent with this role of information. The results in Table A.5, which show that larger firms in the Duke-Richmond CFO survey report more accurate expectations of a variable (real GDP growth) over which they have no control, further bolster this case.

Figure A.1: Expectations, realizations, and errors



Panel (a): expectations and realizations

Panel (b): distribution of errors

Note: Data from I/B/E/S-Compustat. Panel (a): a bin-scatter plot of firms' one-year-ahead revenue expectations and their realizations. Panel (b): a histogram of the associated error distribution. Revenue errors are scaled by a firm's tangible capital stock and normalized by their mean value in the sample. Sample: 2002-2022.

Table A.2: Expectations and input choices

	Panel (a): outcomes and expectations		
	Employment (%)	Capital (%)	Investment (%)
Revenue expectation (%)	0.069***	0.222***	0.139*
	(0.026)	(0.085)	(0.078)
Firm age (quintile)	-2.072	-5.794	-0.014
	(1.627)	(4.198)	(1.725)
Observations	10,260	10,277	10,255
Firm FE	✓	✓	✓
Time FE	✓	✓	✓
F statistic	52.856***	137.772***	26.116***
	Panel (b): outcomes and errors		
	Employment (%)	Capital (%)	Investment (%)
Revenue error lagged (%)	0.188***	0.195	0.661**
	(0.053)	(0.266)	(0.312)
Firm age (quintile)	-2.492	-7.159	-0.888
	(1.772)	(4.733)	(2.079)
Observations	10,020	10,096	10,069
Firm FE	✓	✓	✓
Time FE	✓	✓	✓
F statistic	4.210**	3.228**	5.3786***

Notes: Panel least-squares estimates from the merged I/B/E/S-Compustat sample. Panel (a): estimate of the relationship between realized growth in employment (capital and investment) and the current-period firm-specific expectations of revenue growth. Panel (b): estimate of the relationship between realized growth in employment (capital and investment) and the one-period lagged revenue error. The table also controls for a firm's age, measured in quintiles of the overall age distribution. Expectation (errors) related to the top and bottom 1 percent of the error distribution have been removed. All estimates controls for time and firm fixed effects. Robust (clustered) standard errors in parentheses. Sample: 2002–2022.

A.4 Additional Estimates

Table A.3: Time evolution of accuracy and size

	Panel (a): revenue errors and time			
	Absolute error		Squared error	
	(1)	(2)	(3)	(4)
Time	-0.024^{***}	-0.015^{***}	-0.028^{***}	-0.024^{***}
	(0.003)	(0.002)	(0.007)	(0.005)
Constant	1.251^{***}	1.135^{***}	1.303^{***}	1.219^{***}
	(0.037)	(0.026)	(0.085)	(0.065)
Observations	12,567	12,563	12,567	12,563
Covid dummy	×	✓	×	✓
Residual std. error	1.835	1.278	4.302	3.148
F statistic	67.11^{***}	57.23^{***}	17.70^{***}	22.59^{***}
	Panel (b): size and time			
	50th perc.	70th perc.	80th perc.	90th perc.
	(1)	(2)	(3)	(4)
Time	0.011^{***}	0.012^{***}	0.009^{***}	0.005^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.567^{***}	0.374^{***}	0.215^{***}	0.111^{***}
	(0.015)	(0.016)	(0.007)	(0.005)
Observations	21	21	21	21
Residual std. error	0.029	0.032	0.021	0.014
F statistic	119.85^{***}	115.36^{***}	153.64^{***}	98.96^{***}

Notes: Panel least-squares estimates from the merged I/B/E/S-Compustat sample. Panel (a): estimate of the coefficient of the absolute value (squared value) of individual one-year ahead revenue errors on time. Revenue errors are scaled by a firm's tangible capital stock and normalized by the overall average absolute (squared) error in the sample. The top and bottom 1 percent of errors have been removed. Panel (b): estimate of the coefficient of the share of firms with employment greater than the xth percentile of firms in 2002 on time. Columns (1) and (3) are in levels, whereas Columns (2) and (4) pertain to the logs of variables. Robust standard errors in parentheses. Sample: 2002–2022.

Table A.4: Time evolution of accuracy: sector fixed effects

	Panel (a): sector fixed effects			
	Absolute error		Squared error	
	(1)	(2)	(3)	(4)
Time	-0.024^{***}	-0.016^{***}	-0.027^{***}	-0.026^{***}
	(0.003)	(0.002)	(0.007)	(0.005)
Constant	0.586^{***}	1.818^{**}	0.405^{***}	3.472
	(0.117)	(0.860)	(0.106)	(2.738)
Observations	12,567	12,563	12,567	12,563
Covid dummy	✓	✓	✓	✓
Sector FE	✓	✓	✓	✓
Residual std. error	1.814	1.269	4.282	3.141
F statistic	17.32^{***}	13.71^{***}	7.15^{***}	5.66^{***}
	Panel (b): sector×time fixed effects			
	Absolute error		Squared error	
	(1)	(2)	(3)	(4)
Time	-0.023^{***}	-0.011^{***}	-0.028^{***}	-0.020^{***}
	(0.003)	(0.002)	(0.007)	(0.005)
Constant	1.246^{***}	1.114^{***}	1.301^{***}	1.209^{***}
	(0.033)	(0.026)	(0.082)	(0.067)
Observations	12,567	12,563	12,567	12,563
Sector×time FE	✓	✓	✓	✓
Residual std. error	1.652	1.257	4.184	3.176
F statistic	79.09^{***}	29.37^{***}	18.52^{***}	15.54^{***}

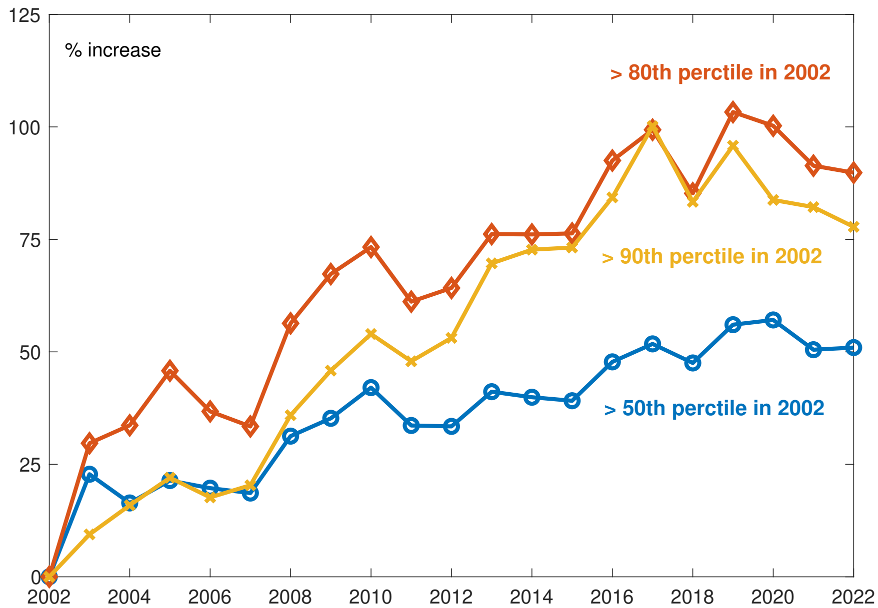
Notes: Panel least-squares estimates from the merged I/B/E/S-Compustat sample. Panel (a): estimate of the coefficient of the absolute value (squared value) of individual one-year ahead revenue errors on time after having partialled out for sector (NAICS-2) fixed effects and a COVID dummy. Revenue errors are scaled by a firm's tangible capital stock and normalized by the overall average absolute (squared) error in the sample. The top and bottom 1 percent of errors have been removed. Panel (b) instead partials out for sector×time fixed effects. Columns (1) and (3) are in levels, whereas Columns (2) and (4) pertain to the logs of variables. Robust standard errors in parentheses. Sample: 2002–2022.

Table A.5: Output expectations from the Duke CFO Survey

	Squared error		Absolute error	
	(1)	(2)	(3)	(4)
Firm size	-0.077^{**}	-0.060^{*}	-0.050^{***}	-0.045^{***}
	(0.028)	(0.024)	(0.012)	(0.010)
GDP familiarity		0.033		0.018
		(0.043)		(0.025)
Constant	1.639^{***}	1.676^{***}	1.648^{***}	1.681^{***}
	(0.115)	(0.090)	(0.051)	(0.042)
Observations	1,584	1,464	1,584	1,464
Sector FE	✓	✓	✓	✓
Time FE	✓	✓	✓	✓
Residual std. error	1.578	1.544	0.812	0.793
F Statistic	19.774^{***}	20.165^{***}	29.175^{***}	29.814^{***}

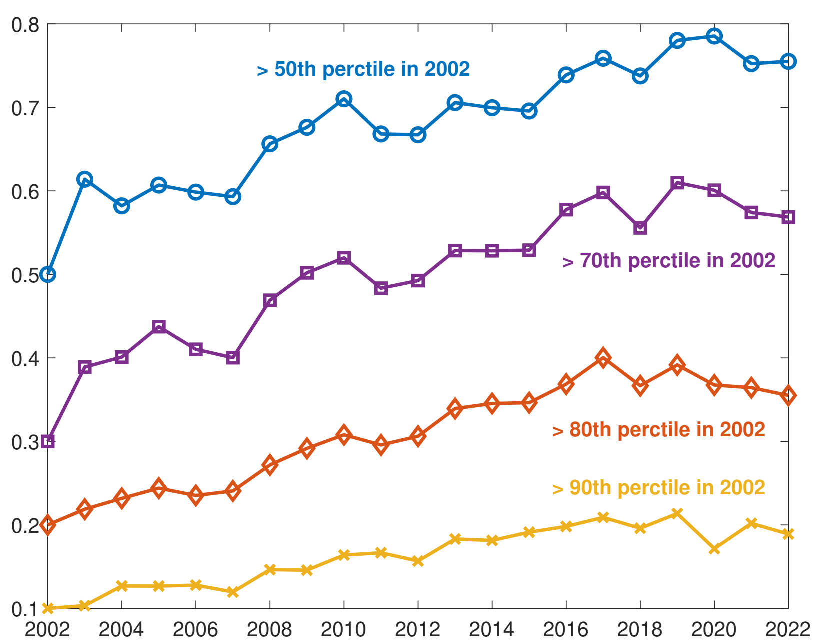
Notes: Estimates from the Duke CFO Survey. Column (1) shows estimates from a regression of the square of individual one-year-ahead real GDP growth errors on firm size (employment) and sector and time fixed effects. Firm size is measured discretely (values 1-5), depending on which quintile firm employment is in relative to the 2020-employment distribution. Column (2) controls for the familiarity of the respondent with the concept of GDP. Columns (3) and (4) consider the absolute value of individual errors. GDP errors are normalized by the overall average squared (absolute) error in the sample. The top and bottom 1 percent of errors have been removed. Robust clustered standard errors in parentheses. Sample: 2020-2022.

Figure A.2: Time evolution of firm size



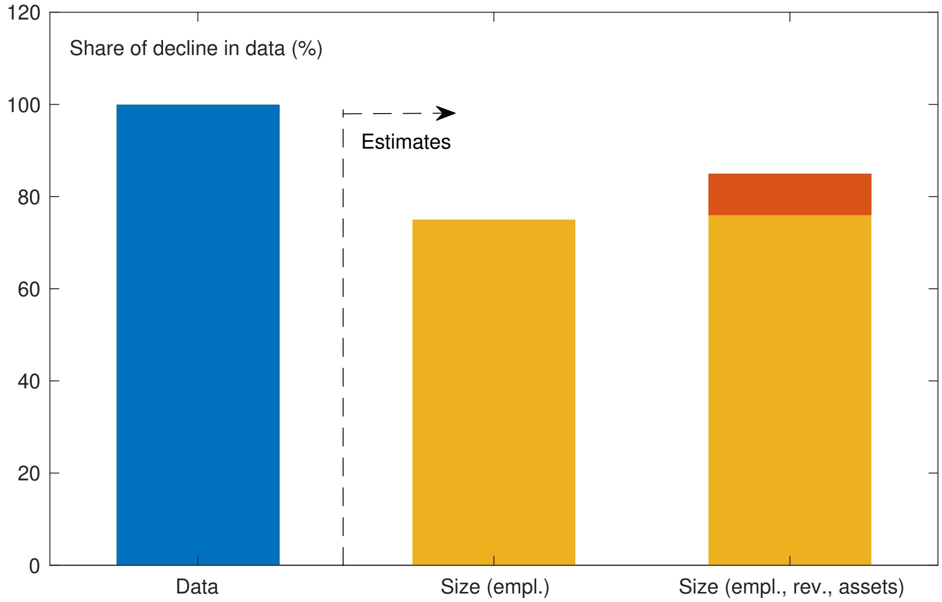
Note: Data from the I/B/E/S-Compustat sample. The panel shows the percentage increase from 2002 in the share of firms in a given year with employment exceeding the xth percentile of the 2002-employment distribution. The 50th, 80th, 90th percentile of the 2002-employment distribution correspond to around 1,000, 7,000, and 18,000 employees, respectively. Table A.3 in the Appendix shows the associated regression results.

Figure A.3: Time evolution of relative firm size



Note: Data from the I/B/E/S-Compustat sample. The figure shows the share of firms in a given year with employment exceeding the xth percentile of the 2002-employment distribution. The 50th, 80th, 90th percentile of the 2002-employment distribution correspond to around 1,000, 7,000, and 18,000 employees, respectively.

Figure A.4: Size and accuracy simulation



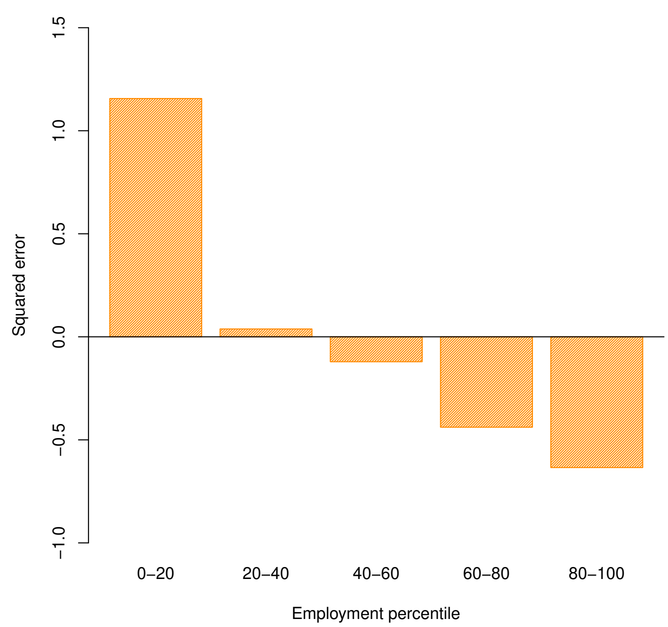
Note: The figure shows the overall rise in revenue accuracy in the I/B/E/S-Compustat sample (Figure 1) from its average value in 2002-2005 to 2022. The figure compares the decline to that implied from the change in the firm-size distribution, using the estimates in Table I Column 4 (with/without the inclusion of real firm assets and real firm revenue as further control variables). Firm real revenue (assets) are measured by the quintile the firm's revenue (assets) are in at time t relative to the 2002-revenue (asset) distribution. Firms revenues and assets are deflated by CPI-U from FRED.

Table A.6: Time evolution of asset size

	Panel (b): asset size and time*			
	50th perc.	70th perc.	80th perc.	90th perc.
	(1)	(2)	(3)	(4)
Time	0.013***	0.011***	0.007***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.408***	0.282***	0.164***	0.102***
	(0.016)	(0.011)	(0.005)	(0.005)
Observations	21	21	21	21
Residual std. error	0.031	0.026	0.016	0.016
F Statistic	126.291***	150.569***	145.456***	62.575***

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. The table estimates the coefficients of the share of firms with assets greater than the xth percentile of firms in 2002 on time. Assets are deflated by CPI-U from FRED. Robust standard errors in parentheses. The 50th, 70th, 80th, and 90th percentile of the 2002-asset distribution correspond to c. 288, 512, 992, and 3,728 million USD. Sample: 2002–2022.

Figure A.5: Revenue accuracy across the asset distribution



Note: The figure plots the difference between the average squared error of one-year-ahead log-revenue expectations from I/B/E/S-Compustat within size (asset) quintiles and the overall average taken across all size levels. Revenue errors are scaled by a firm's tangible capital stock and normalized by their mean value in the sample. Table A.7 reports the coefficient estimates, controlling for firm characteristics. Sample: 2002-2022.

Table A.7: Robustness of revenue expectations, firm size, and time relationship

	Absolute error		*Squared error*		*Squared error log*
	(1)	(2)	(3)	(4)	(5)
Firm size	-0.353^{***}	-0.293^{***}	-0.560^{***}	-0.450^{***}	
	(0.034)	(0.040)	(0.072)	(0.084)	
Firm assets					-0.243^{***}
					(0.056)
Time	-0.006		-0.003		
	(0.006)		(0.012)		
Firm age	0.004	0.023	0.046	0.058	0.022
	(0.016)	(0.024)	(0.035)	(0.052)	(0.031)
Rev. volatility		0.010^{***}		0.010^{**}	0.004
		(0.002		(0.002)	(0.011)
Observations	12,489	6,819	12,489	6,819	6,834
Time FE	×	✓	×	✓	✓
Sector FE	✓	✓	✓	✓	✓
F Statistic	10.460^{***}	7.704^{***}	5.494^{***}	5.043^{***}	2.083^{***}

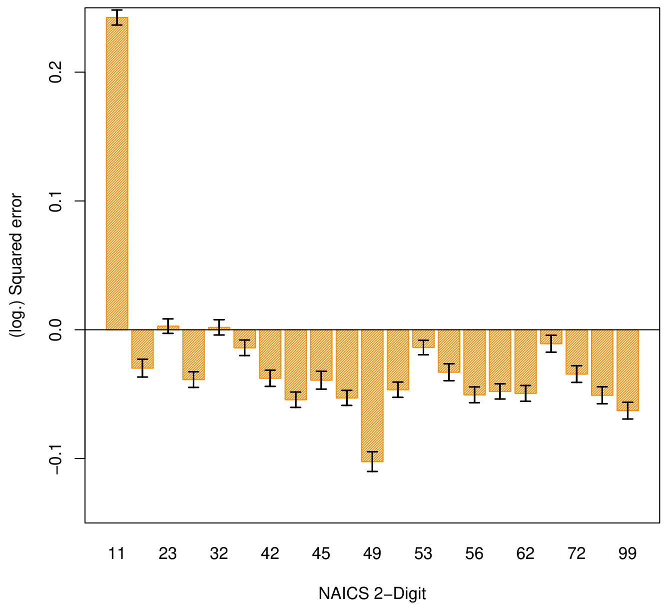
Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Column (1) shows estimates from a regression of the absolute value of individual one-year-ahead revenue errors on firm size (employment), controlling for time, firm age, and sector fixed effects (NAICS-4). Firm size is measured based on which quintile the firm's employment level is at time t relative to the 2002-employment distribution. Column (2) considers the same regression specification but includes time fixed effects, as well as the rolling four-year volatility of revenue. Columns (3) and (4) consider the same specifications studied in Columns (1) and (2), but instead use the squared value of individual errors as the dependent variable. Finally, Column (5) uses the squared value of individual one-year-ahead revenue errors and measures firm size based on which quintile the firm's asset level is at time t relative to the asset distribution. Revenue errors are scaled by a firm's tangible capital stock and normalized by the overall average absolute (squared) error. The top and bottom 1 percent of errors have been removed. Robust (clustered) standard errors in parentheses. Sample: 2002-2022.

Table A.8: Revenue expectations, firm size, and fixed effects

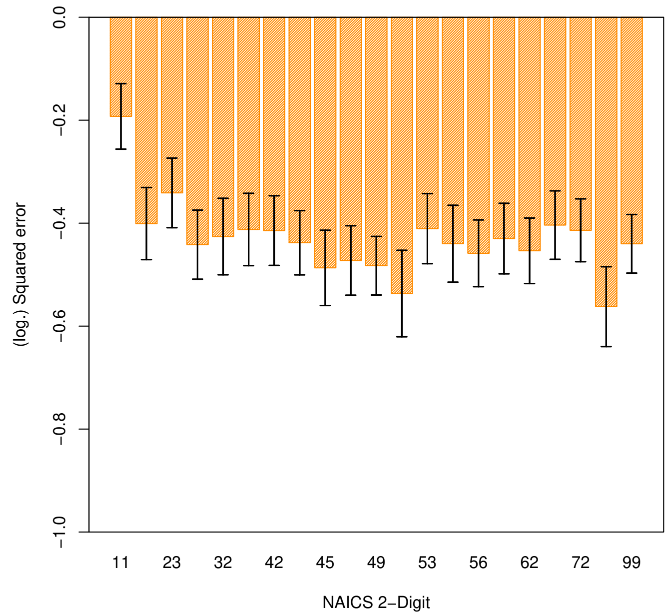
	Squared log-revenue error		
	(1)	(2)	(3)
Firm size	-0.343^{***}	-0.438^{***}	-0.316^{***}
	(0.068)	(0.054)	(0.064)
Firm age	0.054^{*}		0.023
	(0.030)		(0.029)
Log. revenue volatility	-0.006		-0.006
	(0.011)		(0.011)
Observations	6,809	12,488	6,809
Sector FE	✓	×	×
Time FE	✓	×	×
Time×Sector FE	×	✓	✓
F statistic	2.441^{***}	2.401^{***}	1.438^{***}

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Column (1) shows estimates of the squared value of one-year-ahead log-revenue errors on firm size (employment) and sector (NAICS-4) and time fixed effects. We also control for firm age and the individual four-year-rolling average of the volatility of revenue. Firm size is measured by the quintile the firm's employment is at time t relative to the 2002-employment distribution. Columns (2) and (3) adds time×sector (NAICS-2) fixed effects. Revenue errors are scaled by firm capital and normalized by the overall average absolute error. The top and bottom 1 percent of errors have been removed. Robust (clustered) standard errors in parentheses. Sample: 2002-2022.

Figure A.6: Sectoral heterogeneity and the time and size relationship



Panel (a): time relationship



Panel b: size relationship

Note: Panel least-squares estimates from the merged I/B/E/S-Compustat sample. Panel (a): estimate of the coefficient of the squared value of individual one-year-ahead (log-) revenue errors on time for different NAICS 2-digit sectors. Panel (b): estimate of the coefficient of the squared value of individual one-year-ahead (log-) revenue errors on firm size for different NAICS 2-digit sectors. Firm size is measured measured by the quintile the firm's employment is at time t relative to the 2002-employment distribution. Robust (clustered) standard errors in parentheses. Sample: 2002–2022.

Table A.9: Other variables (profits and capex)

	Panel (a): errors and time			
	Profits		Capex	
	Abs. error	Sqr. error	Abs. error	Sqr. error
Time	-0.039^{***}	-0.068^{***}	-0.012^{***}	0.001
	(0.010)	(0.024)	(0.003)	(0.011)
Constant	1.398^{***}	1.691^{***}	1.126^{***}	0.986^{***}
	(0.126)	(0.303)	(0.035)	(0.120)
Observations	2,487	2,487	1,839	1,839
Residual std. error	2.482	6.187	1.871	6.982
F statistic	15.27^{***}	7.385^{***}	12.91^{***}	0.011
	Panel (b): errors and size			
	Profits		Capex	
	Abs. error	Sqr. error	Abs. error	Sqr. error
Firm size	-0.361^{***}	-0.578^{***}	-0.074^{***}	-0.130^{***}
	(0.082)	(0.195)	(0.018)	(0.047)
Firm age	-0.029	0.047	-0.070^{***}	-0.117^{*}
	(0.080)	(0.147)	(0.016)	(0.061)
Constant	0.456	0.680	-0.113^{***}	-0.144
	(0.314)	(0.486)	(0.042)	(0.165)
Observations	2,487	2,487	1,839	1,839
Sector FE	✓	✓	✓	✓
Time FE	✓	✓	✓	✓
Residual std. error	2.347	6.209	1.740	6.744
F statistic	2.694^{***}	1.056	7.906^{***}	2.849^{***}

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Panel (a): estimate of the coefficient of the absolute value (squared value) of individual one-year ahead profit (capex) errors on time. Forecast errors are scaled by a firm's tangible capital stock and normalized by the overall average absolute (squared) error in the sample. The top and bottom 1 percent of errors have been removed. Panel (b): estimate of the coefficient of the absolute value (squared value) of individual one-year ahead errors on firm size, controlling for firm age, and time and sector (NAICS level 4) fixed effects. Firm size is measured as in Table I. Robust (clustered) standard errors in parentheses. Sample: 2004–2022 (profits) and 2002–2022 (capex).

Table A.10: Percentage increase in revenue

	Absolute error		*Squared error*		
	(1)	(2)	(3)	(4)	(5)
Firm size	-0.210^{***}	-0.177^{***}	-0.372^{***}	-0.308^{***}	
	(0.025)	(0.031)	(0.056)	(0.064)	
Time	0.0005		0.005		-0.014^{**}
	(0.004)		(0.008)		(0.007)
Firm age	-0.050^{***}	0.009	-0.095^{**}	0.026	
	(0.016)	(0.025)	(0.039)	(0.051)	
Rev. volatility (pct.)		0.004^{***}		0.005^{***}	
		(0.001)		(0.002)	
Observations	10,083	5,700	10,083	5,700	10,138
Time FE	×	✓	×	✓	×
Sector FE	✓	✓	✓	✓	×
F Statistic	5.589^{***}	4.648^{***}	3.491^{***}	3.163^{***}	6.747^{***}

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Column (1) shows estimates from a regression of the absolute value of individual one-year-ahead revenue growth errors (in pct.) on firm size (employment), controlling for time, firm age, and sector fixed effects (NAICS-4). Firm size is measured based on which quintile the firm's employment level is at time t relative to the 2002-employment distribution. Column (2) considers the same specification but includes time fixed effects, as well as the rolling four-year volatility of revenue. Columns (3) and (4) consider the same specifications studied in Columns (1) and (2), but instead use the squared value of individual errors as the dependent variable. Finally, Column (5) considers the raw correlation with time. Revenue errors are scaled by a firm's tangible capital stock and normalized by the overall average absolute (squared) error. The top and bottom 1 percent of forecast errors have been removed. Robust (clustered) standard errors in parentheses. Sample: 2002-2022.

Table A.11: Large acquisitions and forecast accuracy

	Squared error	Absolute error
	(1)	(2)
Large acquisitions	-0.140^{*}	-0.061^{*}
	(0.007)	(0.033)
Large acquisitions(-1)	-0.119^{*}	-0.082^{***}
	(0.066)	(0.031)
Large acquisitions(-2)	-0.106^{*}	-0.079^{***}
	(0.062)	(0.031)
Large acquisitions(-3)	0.040	-0.040
	(0.068)	(0.031)
Firm age	0.038	0.005
	(0.034)	(0.017)
Log revenue volatility	0.027^{**}	0.017^{***}
	(0.013)	(0.005)
Observations	5,108	5,108
Sector FE	✓	✓
Time FE	✓	✓
Residual std. error	2.324	1.048
F statistic	1.569^{***}	2.954^{***}

Notes: Panel least-squares estimates from the merged I/B/E/S-Compustat sample. The table estimates the coefficient of the squared value (absolute value) of individual one-year-ahead log-revenue errors on firm acquisitions, controlling for firm age, revenue volatility (rolling 4-year average), and time and sector (NAICS level 4) fixed effects. Errors are scaled by a firm's tangible capital stock and normalized by the overall average squared (absolute) error in the sample. The top and bottom 1 percent of errors have been removed. A large acquisitions is defined as one above 5 percent of a firm's assets, consistent with the definition in Ottonello and Winberry (2020). Robust (clustered) standard errors in parentheses. Sample: 2002–2022.

Table A.12: Accuracy and intangible capital

	Accuracy of expectations	
	Sqr. error	Abs. error
Firm acq. stock of intangibles	-0.070*	-0.038***
	(0.040)	(0.015)
Firm size	-0.405***	-0.211***
	(0.041)	(0.019)
Firm age	-0.051	-0.034***
	(0.034)	(0.013)
Observations	11,371	11,371
Sector FE	✓	✓
Time FE	✓	✓
F statistic	3.721***	6.499***

Notes: Panel least-squares estimates from I/B/E/S-Compustat. The table estimates the relationship between of the stock of acquired intangibles and the accuracy of firms' log-revenue expectations. The stock of acquired intangibles accounts adjusts amortization and take-outs financial goodwill (Compustat: `INTAN+AM-GDWL`), and the nominal stock is deflated. This is in accordance with Chiavari and Goraya (2023). Column (1) considers the squared value of individual errors, while Column (2) considers the absolute value. Errors are normalized by the overall average squared (absolute) error in the sample. The top and bottom 1 percent of errors have been removed. All estimates controls for time and sector (NAICS-4) fixed effects. Robust (clustered) standard errors in parentheses. Sample: 2002–2022.

Table A.13: Accuracy and growth in firm size

	Employment		*Revenue*	
	(1)	(2)	(3)	(4)
Informed firms	0.100***	0.110***	0.087**	0.081**
	(0.037)	(0.037)	(0.037)	(0.039)
Initial employment	0.067***	0.066***		
	(0.003)	(0.003)		
Initial revenue			1.175***	1.152***
			(0.001)	(0.001)
Observations	10,186	10,186	10,234	10,234
Sector FE	✓	×	✓	×
Time FE	✓	✓	✓	✓
Industry FE	×	✓	×	✓
Added controls	age	age	age	age
F statistic	1,681***	272.7***	1,573***	271.2***

Notes: Panel estimates from the merged I/B/E/S-Compustat sample. Columns (1) and (2) report estimates of a firm's subsequent employment (2007-2022) on whether a firm was "informed" or not in the initial period (2002-2007), the firm's initial age, as well as the firm's initial employment (2002). We further control for time and sector (NAICS-2) or industry (NAICS-4) fixed effects. Columns (3) and (4) report estimates which instead focus on a firm's revenue. Robust (clustered) standard errors in parentheses. Sample: 2002-2022

B Proofs and Derivations for Sections 3 and 4

Consumer Tastes. Here, we provide a simple microfoundation for the formulation of household preferences in (1). Suppose that instead the utility of household i is given by:

$$\mathcal{U}_i = C_i = \left[\int_0^1 \left(\delta_{ij}^{\mathrm{r}} \cdot c_{ij}^{\mathrm{r}} + \delta_{ij}^{\mathrm{b}} \cdot c_{ij}^{\mathrm{b}}\right)^{\frac{\theta-1}{\theta}} \cdot dj\right]^{\frac{\theta}{\theta-1}}, \tag{A1}$$

where c_{ij}^{r} and c_{ij}^{b} are the consumptions of the red- and blue-type variety j, respectively, and where δ_{ij}^{r} and δ_{ij}^{b} are taste shocks. These taste shocks are assumed to have a noisy mapping to the demand state ω_j. In particular:

$$\delta_{ij}^{\mathrm{r}} = \begin{cases} \delta_H & \text{if } \widetilde{\omega}_{ij}^{\mathrm{r}} = 1 \\ \delta_L & \text{otherwise} \end{cases} \text{ and } \delta_{ij}^{\mathrm{b}} = \begin{cases} \delta_H & \text{if } \widetilde{\omega}_{ij}^{\mathrm{b}} = 1 \\ \delta_L & \text{otherwise} \end{cases}, \tag{A2}$$

where $\widetilde{\omega}_{ij}^{\mathrm{r}}$ and $\widetilde{\omega}_{ij}^{\mathrm{b}}$ are random variables that take values in $\{0,1\}$ and, conditional on the demand state $\omega_j \in \{\mathrm{red}, \mathrm{blue}\}$, they are distributed independently across households and varieties as follows:

$$\mathbb{P}\left(\widetilde{\omega}_{ij}^{\mathrm{r}} = 1 | \omega_j = \mathrm{red}\right) = \bar{\gamma} > \underline{\gamma} = \mathbb{P}\left(\widetilde{\omega}_{ij}^{\mathrm{r}} = 1 | \omega_j = \mathrm{blue}\right), \tag{A3}$$

and

$$\mathbb{P}\left(\widetilde{\omega}_{ij}^{\mathrm{b}} = 1 | \omega_j = \mathrm{blue}\right) = \bar{\gamma} > \underline{\gamma} = \mathbb{P}\left(\widetilde{\omega}_{ij}^{\mathrm{b}} = 1 | \omega_j = \mathrm{red}\right). \tag{A4}$$

Given this formulation, if the demand state is ω_j and if the variety j supplied in the market in equilibrium is of type $x_j = \omega_j$, then for a share $\bar{\gamma}$ of households:

$$\delta_{ij}^{\mathrm{r}} \cdot c_{ij}^{\mathrm{r}} + \delta_{ij}^{\mathrm{b}} \cdot c_{ij}^{\mathrm{b}} = \delta_H \cdot c_{ij}, \tag{A5}$$

and for the remaining share:

$$\delta_{ij}^{\mathrm{r}} \cdot c_{ij}^{\mathrm{r}} + \delta_{ij}^{\mathrm{b}} \cdot c_{ij}^{\mathrm{b}} = \delta_L \cdot c_{ij}, \tag{A6}$$

where c_{ij} are the total units of the variety consumed by household i. If instead the variety j available in the market is of type $x_j \neq \omega_j$, then for a share $\underline{\gamma}$ of households:

$$\delta_{ij}^{\mathrm{r}} \cdot c_{ij}^{\mathrm{r}} + \delta_{ij}^{\mathrm{b}} \cdot c_{ij}^{\mathrm{b}} = \delta_H \cdot c_{ij}, \tag{A7}$$

and for the remaining share:

$$\delta_{ij}^{\mathrm{r}} \cdot c_{ij}^{\mathrm{r}} + \delta_{ij}^{\mathrm{b}} \cdot c_{ij}^{\mathrm{b}} = \delta_L \cdot c_{ij}, \tag{A8}$$

where c_{ij} are the units of the variety consumed by household i.

Therefore, given the types of their varieties chosen by firms in equilibrium, this formulation of household preferences coincides with that in Equation (1) in the main text. □

Proof of Proposition 1. Using household demand in (9) and market clearing for variety j, for given choices of inputs and information, (x_j, n_j, ι_j), firm j's expected profits equal:

$$\mathbb{E}\left[\pi_j|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] = \mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] \cdot n_j^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - w \cdot (n_j + \chi \cdot \iota_j), \tag{A9}$$

where δ_j is given by Equation (9), and where due to independence of the shocks υ_j and ω_j:

$$\mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] = \mathbb{E}\left[\delta_j^{\frac{\theta-1}{\theta}}|s_j^{\omega}, \tau_j^{\omega}\right] \cdot \mathbb{E}\left[A_j^{\frac{\theta-1}{\theta}}|\mu_j, s_j^{\upsilon}, \tau_j^{\upsilon}\right]. \tag{A10}$$

Using the properties of γ_j in Equation (2), we have that:

$$\mathbb{E}\left[\delta_j^{\frac{\theta-1}{\theta}}|s_j^{\omega}, \tau_j^{\omega}\right] = \mathbb{P}\left(\omega_j = x_j|s_j^{\omega}, \tau_j^{\omega}\right) \cdot \delta\left(\bar{\gamma}\right)^{\frac{\theta-1}{\theta}} + \left(1 - \mathbb{P}\left(\omega_j = x_j|s_j^{\omega}, \tau_j^{\omega}\right)\right) \cdot \delta\left(\underline{\gamma}\right)^{\frac{\theta-1}{\theta}}. \tag{A11}$$

Now, because the demand shifter of the "representative consumer" is strictly higher when the firm customizes the variety to its tastes, i.e., $\delta\left(\bar{\gamma}\right) > \delta\left(\underline{\gamma}\right)$, and because signals are weakly informative, firm j optimally chooses to set:

$$x_j = s_j^{\omega} \in \{\text{red}, \text{blue}\} \quad \Longrightarrow \quad \mathbb{P}\left(\omega_j = x_j|s_j^{\omega}, \tau_j^{\omega}\right) = \tau_j^{\omega}. \tag{A12}$$

Maximizing expected profits in Equation (A9) with respect to n_j now yields:

$$n_j = \mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^{\theta} \cdot \Omega, \tag{A13}$$

as stated in the proposition. Lastly, plugging back the optimal choices of x_j and n_j into the expression for expected profits in Equation (A9), we get:

$$\mathbb{E}\left[\pi_j|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] = \left(\frac{1}{\theta-1} \cdot n_j - \chi \cdot \iota_j\right) \cdot w. \tag{A14}$$

Thus, at *Stage 1*, the firm produces information if and only if:

$$\frac{1}{\theta-1} \cdot \left(\mathbb{E}\left[n_j|\mu_j, \bar{\boldsymbol{\tau}}_j\right] - \mathbb{E}\left[n_j|\mu_j, \underline{\boldsymbol{\tau}}_j\right]\right) \geq \chi. \tag{A15}$$

Finally, to arrive at the expression for the expected employment in the text, observe that:

$$\mathbb{E}\left[\delta_j^{\frac{\theta-1}{\theta}}|s_j^{\omega},\tau_j^{\omega}\right]=\tau_j^{\omega}\cdot\delta\left(\bar{\gamma}\right)^{\frac{\theta-1}{\theta}}+\left(1-\tau_j^{\omega}\right)\cdot\delta\left(\underline{\gamma}\right)^{\frac{\theta-1}{\theta}}, \tag{A16}$$

which is independent of the realization of s_j^{ω}, and:

$$\mathbb{E}\left[A_j^{\frac{\theta-1}{\theta}}|\mu_j,s_j^{\upsilon},\tau_j^{\upsilon}\right]=\exp^{\frac{\theta-1}{\theta}\cdot\mu_j}\cdot\exp^{\frac{\theta-1}{\theta}\cdot\frac{\tau_j^{\upsilon}}{\tau_j^{\upsilon}+\tau_a}\cdot s_j^{\upsilon}+\frac{1}{2}\cdot\left(\frac{\theta-1}{\theta}\right)^2\cdot\frac{1}{\tau_j^{\upsilon}+\tau_a}}, \tag{A17}$$

which implies that:

$$\mathbb{E}\left[\mathbb{E}\left[A_j^{\frac{\theta-1}{\theta}}|\mu_j,s_j^{\upsilon},\tau_j^{\upsilon}\right]^{\theta}|\mu_j,\tau_j^{\upsilon}\right]=\exp^{(\theta-1)\cdot\mu_j}\cdot\exp^{\frac{1}{2}\cdot\frac{(\theta-1)^2}{\theta}\cdot\frac{\tau_a+\theta\cdot\tau_j^{\upsilon}}{\tau_a+\tau_j^{\upsilon}}\cdot\frac{1}{\tau_a}}. \tag{A18}$$

Therefore, the expected employment of firm j, conditional on its information choice is:

$$\begin{aligned}\mathbb{E}\left[n_j|\mu_j,\boldsymbol{\tau}_j\right]&=\left[\tau_j^{\omega}\cdot\delta\left(\bar{\gamma}\right)^{\frac{\theta-1}{\theta}}+\left(1-\tau_j^{\omega}\right)\cdot\delta\left(\underline{\gamma}\right)^{\frac{\theta-1}{\theta}}\right]^{\theta}\cdot\exp^{(\theta-1)\cdot\mu_j}\cdot\exp^{\frac{1}{2}\cdot\frac{(\theta-1)^2}{\theta}\cdot\frac{\tau_a+\theta\cdot\tau_j^{\upsilon}}{\tau_a+\tau_j^{\upsilon}}\cdot\frac{1}{\tau_a}}\cdot\Omega\\&=\exp^{(\theta-1)\cdot\mu_j}\cdot g\left(\boldsymbol{\tau}_j\right)^{\theta-1}\cdot\Omega,\end{aligned} \tag{A19}$$

where $g(\cdot)$ is as defined in Equation (15). □

Proof of Corollary 1. The proof follows immediately from the definition of tfp_j in Equation (19), the fact that firm optimality implies that, conditional on $(\mu_j,\boldsymbol{\tau}_j)$, the demand shifter δ_j (as defined by Equation 9) equals $\delta(\bar{\gamma})$ with probability τ_j^{ω} and $\delta(\underline{\gamma})$ with probability $1-\tau_j^{\omega}$, combined with the fact that $\log(A_j)|\mu_j\sim\mathcal{N}(\mu_j,\tau_a^{-1})$. □

Proof of Corollary 2. From the optimality condition for n_j in Equation (12) and the definition of mrp_j in Equation (22), we have:

$$\mathrm{mrp}_j=\log\left(p_j\cdot y_j\right)-\log\left(\mathbb{E}\left[p_j\cdot y_j|\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right]\right)-\log\left(\frac{\theta}{\theta-1}\cdot w\right). \tag{A20}$$

It therefore follows immediately that:

$$\mathbb{VAR}\left[\mathrm{mrp}_j|\mu_j,\boldsymbol{\tau}_j\right]=\mathbb{VAR}\left[\mathrm{error}_j|\mu_j,\boldsymbol{\tau}_j\right]. \tag{A21}$$

Next, using household demand and market clearing for variety j, we have that firm j's revenues are given by the expression, as also used in the proof of Proposition 1:

$$p_j\cdot y_j=\left(\delta_j\cdot A_j\right)^{\frac{\theta-1}{\theta}}\cdot n_j^{\frac{\theta-1}{\theta}}\cdot C^{\frac{1}{\theta}}. \tag{A22}$$

Thus, we have that:

$$
\begin{aligned}
\text{error}_j &= \frac{\theta-1}{\theta}\cdot\left(\log\left(\delta_j\cdot A_j\right)-\mathbb{E}\left[\log\left(\delta_j\cdot A_j\right)|\mu_j,\boldsymbol{s}_j,\boldsymbol{\tau}_j\right]\right)\\
&= \frac{\theta-1}{\theta}\cdot\left(\log\left(\delta_j\right)-\mathbb{E}\left[\log\left(\delta_j\right)|\tau_j^{\omega}\right]\right)+\frac{\theta-1}{\theta}\cdot\left(\log\left(A_j\right)-\mathbb{E}\left[\log\left(A_j\right)|\mu_j,s_j^{\upsilon},\tau_j^{\upsilon}\right]\right)\\
&= \frac{\theta-1}{\theta}\cdot\left(\log\left(\delta_j\right)-\mathbb{E}\left[\log\left(\delta_j\right)|\tau_j^{\omega}\right]\right)+\frac{\theta-1}{\theta}\cdot\left(\frac{\tau_j^{a}}{\tau_j^{\upsilon}+\tau_a}\cdot\upsilon_j-\frac{\tau_j^{\upsilon}}{\tau_j^{\upsilon}+\tau_a}\cdot\varepsilon_j\right).
\end{aligned}
\tag{A23}
$$

Therefore:

$$
\mathbb{VAR}\left[\text{mrp}_j|\mu_j,\boldsymbol{\tau}_j\right]=\left(\frac{\theta-1}{\theta}\right)^2\cdot\mathbb{VAR}\left[\log\left(\delta_j\right)|\tau_j^{\upsilon}\right]+\left(\frac{\theta-1}{\theta}\right)^2\cdot\frac{1}{\tau_j^{\upsilon}+\tau_a} \tag{A24}
$$

$$
=\mathbb{VAR}\left[\text{tfp}_j|\mu_j,\tau_j^{\upsilon}\right]+\left(\frac{\theta-1}{\theta}\right)^2\cdot\left(\frac{1}{\tau_j^{\upsilon}+\tau_a}-\frac{1}{\tau_a}\right). \tag{A25}
$$

□

Proof of Corollary 3. The expression for the marginal-type $\bar{\mu}$ that is just indifferent to producing information follows from the optimality condition for information production in Equation (17) and the expression for a firm's expected employment in Equation (18). Next, consider two firms, j and j', with mean productivities $\mu_j < \bar{\mu} < \mu_{j'}$, so that firm j' produces information while firm j does not.

First, the statement that firm j' on average has a higher and less dispersed tfp_j follows from Corollary 1 and the fact that $\mu_{j'} > \mu_j$ and $\tau_{j'}^{\omega} > \tau_j^{\omega}$.

Second, the statement that firm j' has less dispersed mrp_j follows from Corollary 2 and the fact that $\mu_{j'} > \mu_j$ and $\tau_{j'}^{\upsilon} > \tau_j^{\upsilon}$, combined with the fact that it has less dispersed tfp_j.

Finally, that $\mathbb{E}[n_{j'}|\mu_{j'},\boldsymbol{\tau}_{j'}] > \mathbb{E}[n_j|\mu_j,\boldsymbol{\tau}_j]$ follows from the expression for a firm's expected size in Equation (18), the definition of the information shifter in Equation (15), combined with the fact that $\mu_{j'} > \mu_j$ and $\boldsymbol{\tau}_{j'} > \boldsymbol{\tau}_j$. Clearly, it then also follows that (i) $\mathbb{E}[n_{j'}|\mu_{j'},\boldsymbol{\tau}_{j'}]+\chi\cdot\iota_{j'} > \mathbb{E}[n_j|\mu_j,\boldsymbol{\tau}_j]+\chi\cdot\iota_j$, since $\iota_{j'} = 1 > 0 = \iota_j$; and (ii) $\mathbb{E}[p_{j'}\cdot y_{j'}|\mu_{j'},\boldsymbol{\tau}_{j'}] > \mathbb{E}[p_j\cdot y_j|\mu_j,\boldsymbol{\tau}_j]$ and $\mathbb{E}[p_{j'}\cdot y_{j'}-w\cdot n_{j'}|\mu_{j'},\boldsymbol{\tau}_{j'}] > \mathbb{E}[p_j\cdot y_j-w\cdot n_j|\mu_j,\boldsymbol{\tau}_j]$, since both are proportional to a firm's expected employment (Equation 18). Lastly, it follows that expected profits are higher for j', since firm j' has chosen to produce information. □

Proof of Lemma 1. Equation (28) in the text follows by using the definition of aggregate TFP, $\mathcal{A}$, and the definition of aggregate employment in goods production, $\mathcal{N}$, combined with the optimality conditions in Proposition 1. Moreover:

$$\int_0^1 \mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^{\theta} \cdot dj = \int_0^1 \exp^{(\theta-1)\cdot\mu_j} \cdot g\left(\boldsymbol{\tau}_j\right)^{\theta-1} \cdot dj$$
$$= g\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1} \cdot \int_{-\infty}^{\bar{\mu}} \exp^{(\theta-1)\cdot\mu} \cdot d\Phi\left(-\mu \cdot \sqrt{\tau_\mu}\right) + g\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} \cdot \int_{\bar{\mu}}^{\infty} \exp^{(\theta-1)\cdot\mu} \cdot d\Phi\left(-\mu \cdot \sqrt{\tau_\mu}\right), \quad \text{(A26)}$$

where, using the definition $\xi\left(\bar{\mu}\right) \equiv \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu} + \frac{\theta-1}{\sqrt{\tau_\mu}}\right)$, we have:

$$\int_{\bar{\mu}}^{\infty} \exp^{(\theta-1)\cdot\mu} \cdot d\Phi\left(-\mu \cdot \sqrt{\tau_\mu}\right) = \int_{\bar{\mu}}^{\infty} \frac{1}{\sqrt{2 \cdot \pi \cdot \tau_\mu}} \cdot \exp^{-\mu^2 \cdot \frac{1}{2} \cdot \tau_\mu + (\theta-1)\cdot\mu} \cdot d\mu \quad = \exp^{\frac{1}{2} \cdot \frac{(\theta-1)^2}{\tau_\mu}} \cdot \xi\left(\bar{\mu}\right),$$

and, similarly:

$$\int_{\bar{\mu}}^{\infty} \exp^{(\theta-1)\cdot\mu} \cdot d\Phi\left(-\mu \cdot \sqrt{\tau_\mu}\right) = \exp^{\frac{1}{2} \cdot \frac{(\theta-1)^2}{\tau_\mu}} \cdot \left(1 - \xi\left(\bar{\mu}\right)\right). \quad \text{(A27)}$$

The result then follows by replacing these expressions into Equation (28). □

Proof of Proposition 3. Combining Equations (26) and (31), together with the expression for aggregate TFP in Lemma 1, we have that the equilibrium $\mu^\star$ satisfies:

$$\exp^{(\theta-1)\cdot\mu^\star} = \exp^{\frac{1}{2} \cdot \frac{(\theta-1)^2}{\tau_\mu}} \cdot (\theta - 1) \cdot \chi \cdot \frac{\frac{g(\underline{\boldsymbol{\tau}})^{\theta-1}}{g(\bar{\boldsymbol{\tau}})^{\theta-1} - g(\underline{\boldsymbol{\tau}})^{\theta-1}} + \xi\left(\mu^\star\right)}{N - \chi \cdot \Phi\left(-\mu^\star \cdot \sqrt{\tau_\mu}\right)}. \quad \text{(A28)}$$

The left-hand side of Equation (A28) is monotonically increasing in $\mu^\star$, whereas the right-hand side is monotonically decreasing in $\mu^\star$. As a result, either a decline in χ or a rise in $\bar{\boldsymbol{\tau}}$, both of which imply a decline in the right-hand side, lead to a decline in $\mu^\star$ and thus a rise in the share of firms that produce information. From Lemma 1, it then follows that aggregate TFP rises as well; in the case of a rise in $\bar{\boldsymbol{\tau}}$ it rises both *directly* as $g\left(\bar{\boldsymbol{\tau}}\right)$ increases and *indirectly* through the fall in $\mu^\star$. As for aggregate consumption and welfare, it is straightforward to show that Equation (A28) implies, after a few derivations, that:

$$\frac{d}{d\bar{\mu}} \cdot \left(\mathcal{A}\left(\bar{\mu}, g\right) \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)\right]\right) |_{\bar{\mu}=\mu^\star} = 0. \quad \text{(A29)}$$

Hence, by an envelope argument, we need only consider the direct effect of changes in χ or $\bar{\boldsymbol{\tau}}$ on $C^\star$. But, these are clearly positive, since, for a given $\mu^\star$, a lower χ reduces costs of information production, whereas a higher $\bar{\boldsymbol{\tau}}$ raises aggregate TFP. □

C Proofs and Derivations for Section 5

Optimal Trading Mechanism. We provide detailed derivations for the optimal trading mechanism of Section 5.1. Since at *Stage* 3, the firm's supply of variety j is inelastic and given by $Q_j = A_j \cdot n_j$, the firm's objective is to maximize the revenues collected from the consumers (Equation 34) subject to incentive compatibility (Equation 35), individual rationality (Equation 36), and feasibility (Equation 37) constraints.[42]

We conjecture (and then verify) that the individual rationality (IR) constraint of the H-type and the incentive compatibility (IC) constraint of the L-type are slack. Given this, the firm's problem becomes:

$$\max_{\left(t_j^H, q_j^H\right), \left(t_j^L, q_j^L\right) \text{ with } q_j^L, q_j^H \geq 0} \quad \gamma_j \cdot t_j^H + (1-\gamma_j) \cdot t_j^L \tag{A30}$$

s.t.

$$\left(\delta_H \cdot q_j^H\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - t_j^H \geq \left(\delta_H \cdot q_j^L\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - t_j^L, \tag{A31}$$

$$\left(\delta_L \cdot q_j^L\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - t_j^L \geq 0, \tag{A32}$$

$$\gamma_j \cdot q_j^H + (1-\gamma_j) \cdot q_j^L \leq Q_j. \tag{A33}$$

By inspection, it is optimal for the firm to set t_j^L and t_j^H so that both the IC-constraint of the H-type and the IR-constraint of the L-type bind. Moreover, it is clear that the firm will allocate all of its goods to the consumers, so that the feasibility constraint binds. As a result, the problem reduces to:

$$\max_{q_j^L, q_j^H \geq 0} \ \gamma_j \cdot \delta_H^{\frac{\theta-1}{\theta}} \cdot {q_j^H}^{\frac{\theta-1}{\theta}} + \left(\delta_L^{\frac{\theta-1}{\theta}} - \gamma_j \cdot \delta_H^{\frac{\theta-1}{\theta}}\right) \cdot {q_j^L}^{\frac{\theta-1}{\theta}} \tag{A34}$$

s.t.

$$\gamma_j \cdot q_j^H + (1-\gamma_j) \cdot q_j^L = Q_j. \tag{A35}$$

Clearly, if $\gamma_j = 0$, then the firm only faces L-type consumers and, thus, it is optimal to set $q_j^L = Q_j$. In this case, the firm is indifferent to any q_j^H, and for convenience we suppose that $q_j^H \geq q_j^L$. If $\gamma_j \geq (\delta_L/\delta_H)^{\frac{\theta-1}{\theta}}$, it is optimal for the firm to exclude the L-type consumer and set $q_j^H = \gamma_j^{-1} \cdot Q_j \geq 0 = q_j^L$. Finally, for $\gamma_j \in (0, (\delta_L/\delta_H)^{\frac{\theta-1}{\theta}})$, the allocation of goods is

[42] Notice that we here use Q_j for the overall quantity produced of variety j instead of y_j as before. This is better link our notation to the individual quantities, q_j^l for $l = \{H, L\}$, proposed in the mechanism.

interior and given by:

$$q_j^L = \frac{(\psi_j \cdot \delta_L)^{\theta-1}}{\gamma_j \cdot \delta_H^{\theta-1} + (1-\gamma_j) \cdot (\psi_j \cdot \delta_L)^{\theta-1}} \cdot Q_j \text{ and } q_j^H = \frac{\delta_H^{\theta-1}}{\gamma_j \cdot \delta_H^{\theta-1} + (1-\gamma_j) \cdot (\psi_j \cdot \delta_L)^{\theta-1}} \cdot Q_j, \tag{A36}$$

where ψ_j is given in Equation (40) in the main text. The expressions for the shares α_j^L and α_j^H in Equations (39) and (38) then follow immediately.

Denote the optimal allocation of the type-l consumer by $\left(t_j^{l\star}, q_j^{l\star}\right)$, and let us verify the conjecture that the IC-constraint of the L-type and the IR-constraint of the H-type are satisfied at the optimum. The latter follows by combining the binding IC-constraint of the H-type and the binding IR-constraint of the L-type:

$$t_j^{H\star} = \left(\delta_H \cdot q_j^{H\star}\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - \left(\delta_H^{\frac{\theta-1}{\theta}} - \delta_L^{\frac{\theta-1}{\theta}}\right) \cdot \left(q_j^{L\star}\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} \leq \left(\delta_H \cdot q_j^{H\star}\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}}. \tag{A37}$$

The former follows from the fact that $q_j^{L\star} \leq q_j^{H\star}$ and, again, by combining the binding IC-constraint of the H-type and the binding IR-constraint of the L-type:

$$\begin{aligned}\left(\delta_L \cdot q_j^{H\star}\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - t_j^{H\star} &= \left(\delta_L \cdot q_j^{H\star}\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - \left[\left(\delta_H \cdot q_j^{H\star}\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - \left(\delta_H^{\frac{\theta-1}{\theta}} - \delta_L^{\frac{\theta-1}{\theta}}\right) \cdot q_j^{L\star\frac{\theta-1}{\theta}}\right] \\ &= \left(\delta_L^{\frac{\theta-1}{\theta}} - \delta_H^{\frac{\theta-1}{\theta}}\right) \cdot \left[\left(q_j^{H\star}\right)^{\frac{\theta-1}{\theta}} - \left(q_j^{L\star}\right)^{\frac{\theta-1}{\theta}}\right] \cdot C^{\frac{1}{\theta}} \leq 0, \end{aligned} \tag{A38}$$

and the L-type earns zero surplus at the optimal allocation.

We now derive the expression for the L-type consumer's surplus earned from a given allocation $\left(t_j^l, q_j^l\right)$ offered by the mechanism, which we had assumed in the above derivations. To this end, note that the utility that consumer i gains from consuming q_{ij} units of variety j, when her overall consumption is C_i, is given by:

$$\widetilde{u}_{ij}\left(q_{ij}\right) = \frac{\theta}{\theta-1} \cdot \left(\delta_{ij} \cdot q_{ij}\right)^{\frac{\theta-1}{\theta}} \cdot C_i^{\frac{1}{\theta}}. \tag{A39}$$

Suppose that λ_i is the marginal value of income to the consumer, expressed in units of the aggregate consumption bundle, which we assume is the numeraire. Then, the surplus of the consumer in units of the numeraire is $\frac{\widetilde{u}_{ij}(q_{ij})}{\lambda_i}$. Since households are ex-ante identical, in any symmetric equilibrium it must be that $\lambda_i = \lambda$ and $C_i = C$ for all i. In equilibrium, therefore, aggregation implies that $\lambda = \frac{\theta}{\theta-1}$. Because in equilibrium all agents have correct expectations about λ and C, they understand that consumer i's willingness to pay for q_{ij} units of its variety is $\left(\delta_{ij} \cdot q_{ij}\right)^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}}$, as assumed above.

Now, in the above, we assumed that the representative household's marginal value of income, λ, is well defined. For this to be the case, we need to be able to assign a utility value to an additional unit of income in the hands of the household. However, if firms were only offering the bundles $\left\{\left(t_j^{l\star}, q_j^{l\star}\right)\right\}$, which are all accepted in equilibrium, the household would have no opportunities to spend this additional income. To address this issue, we follow an approach similar to Bornstein and Peter (2024), and suppose that firms post additional latent allocations in their menus, which are not accepted in equilibrium but ensure that the consumer faces linear prices off-equilibrium. Thus, suppose that, in addition to the bundles $\left\{\left(t_j^{l\star}, q_j^{l\star}\right)\right\}$, each firm j offers to sell to the consumer any quantity $q > q_j^H$ for a total payment of:

$$t_j(q) = t_j^{H\star} + \kappa_j^\star \cdot \left(q - q_j^{H\star}\right) \text{ with } \kappa_j^\star = \delta_H^{\frac{\theta-1}{\theta}} \cdot \left(q_j^{H\star}\right)^{-\frac{1}{\theta}} \cdot C^{\frac{1}{\theta}}. \tag{A40}$$

By construction, no bundle with $q > q_j^{H\star}$ is picked up in equilibrium by the household with a marginal value of income $\lambda = \frac{\theta}{\theta-1} = (\kappa_j^\star)^{-1} \cdot \tilde{u}_{Hj}(q_j^{H\star})$, where the right-hand side is the marginal gain in utility to the H-type household per unit of the numeraire that is spent on obtaining an additional unit above $q_j^{H\star}$. Moreover, the marginal value of income to the household is now well defined, since she can now spend the additional income (if she had any) to obtain units above $q_j^{H\star}$ for the varieties for which she is an H-type. □

Alternative Trading Environments. The trading mechanism that we considered in Section 5 resembles that of a centralized market for each variety j, in which price discrimination occurs through quantity bundles. Here, we provide two alternative specifications, which generalize our results to markets where goods trade in a decentralized fashion, or where goods are indivisible and discrimination occurs through quality differentiation.

Bargaining in Decentralized Markets. Consider now a decentralized marketplace, in which each firm j meets each consumer on an "island" and engages in bargaining. In particular, as in Section 5.1, the firm proposes menus to the consumer; however, due to a "search friction" the firm is unable to ship the goods across islands.[43] This implies that the feasibility constraint in (37) must now replaced by $q_j^l \leq Q_j$.

Following in the same footsteps of the derivation of the optimal mechanism of Section 5.1 (see above), we can show that the solution to this mechanism design problem is as follows:

- If $\gamma_j \geq (\delta_L/\delta_H)^{\frac{\theta-1}{\theta}}$, then $\alpha_j^H = \gamma_j^{-1} \cdot Q_j > 0 = \alpha_j^L$;
- If $\gamma_j < (\delta_L/\delta_H)^{\frac{\theta-1}{\theta}}$, then $\alpha_j^H = \alpha_j^L = 1$;

where α_j^l denotes the share of its goods that firm j allocates to type-l consumer. Thus, we see that firm j either excludes the L-type completely or trades with both consumer types the

[43] If the firm could ship goods freely across islands, then the allocations of the optimal mechanism would coincide with those of the mechanism in Section 5.1.

entire quantity on equal terms. Analogously to Section 5.2, we can then derive the revenue- and surplus-based demand shifters also in this setting:

$$\delta_j^R = \begin{cases} \delta_L \\ \gamma_j^{\frac{\theta}{\theta-1}} \cdot \delta_H \end{cases} \qquad \delta_j^S = \begin{cases} (\gamma_j \cdot \delta_H^{\frac{\theta-1}{\theta}} + (1-\gamma_j) \cdot \delta_L^{\frac{\theta-1}{\theta}})^{\frac{\theta}{\theta-1}} & \text{if } \gamma_j < (\delta_L/\delta_H)^{\frac{\theta-1}{\theta}} \\ \gamma_j^{\frac{\theta}{\theta-1}} \cdot \delta_H & \text{otherwise} \end{cases} . \tag{A41}$$

The properties of these two demand shifters, which are central to all of the positive and normative results in Section 5, are just as stated in Lemma 2. Namely, (i) the revenue-based demand shifter, δ_j^R, is increasing in γ_j (strictly so for $\gamma_j \geq (\delta_L/\delta_H)^{\frac{\theta-1}{\theta}}$); (ii) the surplus-based demand shifter, δ_j^R is non-monotonic in γ_j (in this setting, this feature always arises); and (iii) the ratio δ_j^S/δ_j^R increases at first (i.e., in the region $\gamma_j < (\delta_L/\delta_H)^{\frac{\theta-1}{\theta}}$) and then decreases in γ_j.

Quality Differentiation. Consider a simple reformulation of our setting that captures trade in indivisible goods. More concretely, let us suppose that each consumer demands a single unit of each variety j, but where c_{ij} in the consumer's preferences in (1) captures the quality of the good instead. In turn, suppose that at *Stage 3*, each firm j can use its units of effective labor, $A_j \cdot n_j$, to create goods of high quality, q_j^H, and low quality, q_j^L, and sell them to consumers at prices t_j^H and t_j^L, respectively. It is then straightforward to show that the allocations of this economy are identical to that of our rent-extracting economy studied in Section 5. If, instead, the firm must choose the quantity of its variety at *Stage 2*, i.e., $q_j = A_j \cdot n_j$, then it is straightforward to show that the allocations of this economy are similar to those under the *Bargaining-in-the-Decentralized-Markets* extension discussed above: variety j of quality q_j is either equally allocated to both consumer types or to the H-type consumer only. □

Proof of Lemma 2. See text. □

Proof of Proposition 4. The result follows immediately from the observation that the ex-post profits of firm j in the rent-extracting economy are given by Equation (41). □

Proof of Proposition 5. Given the equilibrium $\mu^\star$, aggregate TFP equals:

$$\begin{aligned}\mathcal{A} = C \cdot \mathcal{N}^{-1} &= \frac{\left(\int_0^1 \left[\gamma_j \cdot \delta_H^{\frac{\theta-1}{\theta}} \cdot \alpha_j^{H\frac{\theta-1}{\theta}} + (1-\gamma_j) \cdot \delta_L^{\frac{\theta-1}{\theta}} \cdot \alpha_j^{L\frac{\theta-1}{\theta}}\right] \cdot (A_j \cdot n_j)^{\frac{\theta-1}{\theta}} \cdot dj\right)^{\frac{\theta}{\theta-1}}}{\int_0^1 n_j \cdot dj} \\ &= \frac{\left(\int_0^1 \left(\delta_j^S \cdot A_j\right)^{\frac{\theta-1}{\theta}} \cdot n_j^{\frac{\theta-1}{\theta}} \cdot dj\right)^{\frac{\theta}{\theta-1}}}{\int_0^1 n_j \cdot dj},\end{aligned}$$

where the third equality follows from the definition of δ_j^S in Equation (44). Thus,

$$\mathcal{A} = \frac{\left(\int_0^1 \mathbb{E}\left[\delta_j^{S\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] \cdot \mathbb{E}\left[A_j^{\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] \cdot n_j^{\frac{\theta-1}{\theta}} \cdot dj\right)^{\frac{\theta}{\theta-1}}}{\int_0^1 n_j \cdot dj}$$
$$= \frac{\left(\int_0^1 \mathbb{E}\left[\delta_j^{S\frac{\theta-1}{\theta}}|\tau_j^\omega\right] \cdot \mathbb{E}\left[A_j^{\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] \cdot n_j^{\frac{\theta-1}{\theta}} \cdot dj\right)^{\frac{\theta}{\theta-1}}}{\int_0^1 n_j \cdot dj}, \tag{A42}$$

where the first equality follows from the observation that δ_j^S is independent of A_j, conditional on firm j's information set $(\mu_j, \boldsymbol{\tau}_j, \boldsymbol{s}_j)$, and the second equality follows from the fact that optimal product choice $x_j = s_j^\omega$ implies $\mathbb{E}\left[\delta_j^{S\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] = \mathbb{E}\left[\delta_j^{S\frac{\theta-1}{\theta}}|\tau_j^\omega\right]$. Using the definition of the gap, $\Delta(\boldsymbol{\tau})$, between profit and social surplus maximization in Equation (46) and the optimal labor choice in Proposition 4, we find that:

$$\mathcal{A} = \frac{\left(\int_0^1 \Delta\left(\boldsymbol{\tau}_j\right)^{\frac{\theta-1}{\theta}} \cdot \mathbb{E}\left[\left(\delta_j^R \cdot A_j\right)^{\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^\theta \cdot dj\right)^{\frac{\theta}{\theta-1}}}{\int_0^1 \mathbb{E}\left[\left(\delta_j^R \cdot A_j\right)^{\frac{\theta-1}{\theta}}|\mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right]^\theta \cdot dj}$$
$$= \exp^{\frac{\theta-1}{2}\cdot\frac{1}{\tau_\mu}} \cdot \frac{\left[\Delta\left(\underline{\boldsymbol{\tau}}\right)^{\frac{\theta-1}{\theta}} \cdot g^R\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1} \cdot \left(1-\xi\left(\mu^\star\right)\right) + \Delta\left(\bar{\boldsymbol{\tau}}\right)^{\frac{\theta-1}{\theta}} \cdot g^R\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} \cdot \xi\left(\mu^\star\right)\right]^{\frac{\theta}{\theta-1}}}{g^R\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1} \cdot \left(1-\xi\left(\mu^\star\right)\right) + g^R\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} \cdot \xi\left(\mu^\star\right)}$$
$$= \mathcal{E}\left(\mu^\star\right) \cdot \mathcal{A}(\mu^\star, g^R), \tag{A43}$$

where the two equalities follow from the definition of the information shifter $g^R(\cdot)$ in Equation (45), the expressions for $\xi(\mu^\star)$ and $\mathcal{A}(\mu^\star, g)$ in Lemma 1, as well as the definition of the macro-wedge, $\mathcal{E}$, in Equation (49). Given the expression for aggregate TFP, the expression for aggregate consumption then follows immediately. □

Proof of Corollary 4. See text. □

Proof of Proposition 6. Consider the problem of the social planner as described in Section 5.5. To solve this problem, we will begin with the conjecture that, given a quantity $Q_j = A_j \cdot n_j$ produced by firm j, the planner can allocate consumption across the two types of consumers freely, subject to the feasibility constraint:

$$\gamma_j \cdot c_{Hj} + (1-\gamma_j) \cdot c_{Lj} \leq Q_j, \tag{A44}$$

i.e., for now we ignore the fact that the planner also faces incentive compatibility and participation constraints. We will later show that the resulting allocations can be implemented as an outcome of a mechanism that satisfies these constraints, just as the privately described trading mechanisms in Section 5.1.

Given this conjecture, the planner will equalize marginal utilities across types:

$$\delta_H^{\frac{\theta-1}{\theta}} \cdot c_{Hj}^{-\frac{1}{\theta}} = \delta_L^{\frac{\theta-1}{\theta}} \cdot c_{Lj}^{-\frac{1}{\theta}} \implies c_{lj} = \widetilde{\alpha}_j^l(\gamma_j) \cdot Q_j, \tag{A45}$$

where $\widetilde{\alpha}_j^l(\gamma_j)$ is the share of the good allocated to type-l consumer, and it is the same as the share $\alpha_j^l(\gamma_j)$ given in Equations (38) and (39), except that the micro-wedge, ψ_j, equals one.

Let λ denote the marginal value of a unit of labor to the social planner, i.e., the multiplier on the aggregate resource constraint $\int_0^1 (n_j + \chi \cdot \iota_j) \cdot dj \leq N$. Given the above consumption allocation, it follows that the ex-post surplus (in terms of utils) produced by firm j that chooses (x_j, n_j, ι_j) can be expressed as:

$$u_j(x_j, n_j, \iota_j, \upsilon_j, \omega_j) = \frac{\theta}{\theta-1} \cdot (\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} \cdot n_j^{\frac{\theta-1}{\theta}} \cdot C^{\frac{1}{\theta}} - \lambda \cdot (n_j + \chi \cdot \iota_j) \tag{A46}$$

where $\delta_j = \left(\gamma_j \cdot \delta_H^{\theta-1} + (1-\gamma_j) \cdot \delta_L^{\theta-1}\right)^{\frac{1}{\theta-1}}$ is the demand shifter defined in Equation (9). Thus, at *Stage 2*, the social planner's choices (x_j, n_j, ι_j) are similar to those in Proposition 1:

$$x_j = s_j^{\omega} \text{ and } n_j = \mathbb{E}\left[(\delta_j \cdot A_j)^{\frac{\theta-1}{\theta}} | \mu_j, \boldsymbol{s}_j, \boldsymbol{\tau}_j\right] \cdot \widetilde{\Omega}, \tag{A47}$$

except that the effective "market size" faced by the planner is now $\widetilde{\Omega} \equiv \frac{C}{\lambda^{-\theta}}$, and:

$$\iota_j = \begin{cases} 1 & \text{if } \frac{1}{\theta-1} \cdot (\mathbb{E}[n_j|\mu_j, \bar{\boldsymbol{\tau}}] - \mathbb{E}[n_j|\mu_j, \underline{\boldsymbol{\tau}}]) \geq \chi \\ 0 & \text{otherwise} \end{cases}, \tag{A48}$$

with expected employment of firm j given by:

$$\mathbb{E}[n_j|\mu_j, \boldsymbol{\tau}_j] = \exp^{(\theta-1)\cdot\mu_j} \cdot g(\boldsymbol{\tau}_j)^{\theta-1} \cdot \widetilde{\Omega}. \tag{A49}$$

Thus, at the planner's allocation, firm j produces information if and only if

$$\mu_j \geq \bar{\mu}\left(\widetilde{\Omega}\right) = \frac{1}{\theta-1} \cdot \log\left[\frac{(\theta-1)\cdot\chi}{\left(g(\bar{\boldsymbol{\tau}})^{\theta-1} - g(\underline{\boldsymbol{\tau}})^{\theta-1}\right) \cdot \widetilde{\Omega}}\right], \tag{A50}$$

which note is the same schedule as that given in Equation (26). Since the planner does not leave any labor unused, if we take as given the marginal type $\bar{\mu}$ that produces information, the aggregate resource constraint yields the expression for the "market size" faced by the planner:

$$\widetilde{\Omega} = \widetilde{\Omega}\left(\bar{\mu}\right) = \frac{\mathcal{A}\left(\bar{\mu}, g\right) \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_{\mu}}\right)\right]}{\mathcal{A}\left(\bar{\mu}, g\right)^{\theta}}, \tag{A51}$$

where $\mathcal{A}(\cdot)$ is as given by Lemma 1. This schedule is also the same as the one in Equation (31). As a result, the allocations of the social planner coincide with those of our baseline economy: the marginal type that produces information is given by the intersection of the above two schedules:

$$\mu^{\star} = \bar{\mu}\left(\widetilde{\Omega}\left(\mu^{\star}\right)\right), \tag{A52}$$

and the aggregate TFP and welfare are in turn given by:

$$\mathcal{A}^{\star} = \mathcal{A}\left(\mu^{\star}, g\right) \text{ and } C^{\star} = \mathcal{A}^{\star} \cdot \left[N - \chi \cdot \Phi\left(-\mu^{\star} \cdot \sqrt{\tau_{\mu}}\right)\right]. \tag{A53}$$

As there are increasing returns to information production, we must nevertheless still verify that $\mu^{\star} = \arg\max_{\bar{\mu}} \mathcal{A}\left(\mu^{\star}, g\right) \cdot \left[N - \chi \cdot \Phi\left(-\mu^{\star}\right)\right]$, i.e., that the amount of information production chosen above actually maximizes welfare. But, this follows from the observation that:

$$\begin{aligned} &\frac{d}{d\bar{\mu}}\left(\mathcal{A}\left(\bar{\mu}, g\right) \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_{\mu}}\right)\right]\right) \\ &\propto (\theta - 1) \cdot \chi \cdot \frac{1}{g\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} - g\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1}} \cdot \frac{\mathcal{A}\left(\bar{\mu}, g\right)^{\theta-1}}{N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_{\mu}}\right)} - \exp^{(\theta-1)\cdot\bar{\mu}}, \end{aligned} \tag{A54}$$

where the right-hand side is positive (negative) whenever $\bar{\mu} < \mu^{\star}$ ($\bar{\mu} > \mu^{\star}$).

Lastly, we verify the conjecture that the allocations of the planner are incentive compatible and individually rational. To this end, consider the following menu provided to consumers of each variety j by the social planner: $\mathcal{M} = \left\{\left(t_j, q_j\right)\right\}_{q_j \geq 0}$ where q_j is the quantity of the good transferred to the consumer and t_j is the transfer of the numeraire good C from the consumer to the planner—which is then rebated lump sum to the consumer. The transfer schedule in turn satisfies:

$$t_j = t\left(q_j\right) = \left[\delta_H^{\frac{\theta-1}{\theta}} \cdot q_j^{H\star - \frac{1}{\theta}} \cdot C^{\star\frac{1}{\theta}}\right] \cdot q_j, \tag{A55}$$

where the superscript $\star$ indicates the allocations of the social planner that we obtained above (e.g., $q_j^{H\star} = \widetilde{\alpha}_j^H\left(\gamma_j\right) \cdot A_j \cdot n_j^{\star}$). Note that, since the marginal utilities across consumer types are equalized, it must also be that $t_j = \left[\delta_L^{\frac{\theta-1}{\theta}} \cdot q_j^{L\star - \frac{1}{\theta}} \cdot C^{\star\frac{1}{\theta}}\right] \cdot q_j$. Given this menu, and the conjectured equilibrium allocations, a consumer of type-l indeed optimally chooses $q_j = q_j^{l\star}$ since it equalizes her marginal utility to her marginal cost. Therefore, by revealed preference, the planner's allocations are incentive compatible and individually rational.

We have thus shown that the social planner's allocations coincide with those of our baseline economy. Since the equilibrium of the rent-extracting economy generically differs from the equilibrium of our baseline economy (except in the case where $0 = \underline{\gamma} < \bar{\gamma} = 1$), we conclude that it must generically be inefficient. $\square$

Proof of Proposition 7. Following in the steps of Proposition 6, we know that the marginal type $\mu^\star$ that is just indifferent to producing information in the laissez-faire equilibrium of the rent-extracting economy is the unique maximizer of:

$$\mathcal{A}(\bar{\mu}, g^R) \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)\right]. \tag{A56}$$

It therefore follows that the optimal information tax, which selects $\bar{\mu}$ to maximize:

$$\mathcal{E}\left(\bar{\mu}\right) \cdot \mathcal{A}\left(\bar{\mu}, g^R\right) \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)\right] \tag{A57}$$

is positive (negative) if $\mathcal{E}'\left(\cdot\right) > 0$ (< 0), since $\mathcal{E}'\left(\cdot\right) \gtreqless 0$ iff $\delta^S\left(\bar{\gamma}\right)/\delta^R\left(\bar{\gamma}\right) \lesseqgtr \delta^S\left(\underline{\gamma}\right)/\delta^R\left(\underline{\gamma}\right)$. $\square$

Proof of Proposition 8. From Proposition 2, the aggregate TFP and consumption in the rent extracting economy are:

$$\mathcal{A}^\star = \mathcal{E}\left(\mu^\star\right) \cdot \mathcal{A}\left(\mu^\star, g^R\right) \text{ and } C^\star = \mathcal{A}^\star \cdot \left[N - \chi \cdot \Phi\left(-\mu^\star \cdot \sqrt{\tau_\mu}\right)\right],$$

where $\mu^\star$ is the equilibrium marginal type as characterized by Proposition 1.

Let us first suppose that we are in the parametric case where rent extraction is socially destructive. Given any $\bar{\mu}$ and any garbling policy about productivity information, z^υ, both $\mathcal{E}\left(\bar{\mu}\right)$ and $\mathcal{A}\left(\bar{\mu}, g^R\right)$ are increasing in z^ω. At the maximum when $z^\omega = 1$, $\mathcal{E}\left(\bar{\mu}\right)$ becomes independent of $\bar{\mu}$ and recall $\mathcal{E}\left(\cdot\right)$ also does not depend on z^υ. Thus, $z^\omega = 1$ is optimal for the planner and, given $z^\omega = 1$, an optimal garbling policy also sets $z^\upsilon = 0$, since $\mathcal{A}\left(\bar{\mu}, g^R\right)|_{z^\upsilon, z^\omega = 1}$ is decreasing in z^υ for a given $\bar{\mu}$.[44] Lastly, given the optimal garbling policy, $z^\omega = 1$ and $z^\upsilon = 0$, by the same arguments as in the proof of Proposition 7, the optimal information tax T must be zero, since the equilibrium selects the marginal type that is the maximizer of $\mathcal{A}\left(\bar{\mu}, g^R\right)|_{z^\omega=1, z^\upsilon=0} \cdot \left[N - \chi \cdot \Phi\left(-\bar{\mu} \cdot \sqrt{\tau_\mu}\right)\right]$, which maximizes social welfare, since under the optimal garbling policy, $\mathcal{E}\left(\cdot\right)$ is independent of $\bar{\mu}$.

Next, suppose that we are in the parametric case where rent extraction is not socially destructive. Arguments analogous to those above imply that, for any given $\bar{\mu}$ and z^ω, $\mathcal{E}\left(\bar{\mu}\right) \cdot \mathcal{A}\left(\bar{\mu}, g^R\right)$ is decreasing z^υ. Therefore, the social planner again sets $z^\upsilon = 0$. $\square$

[44] The conditioning on (z^υ, z^ω) in $\mathcal{A}\left(\bar{\mu}, g^R\right)|_{z^\omega, z^\upsilon}$ simply indicates that we are evaluating TFP with the information shifter, in which the precision of produced information is $(\underline{\tau}^\upsilon + z^\upsilon \cdot (\bar{\tau}^\upsilon - \underline{\tau}^\upsilon), \underline{\tau}^\omega + z^\omega \cdot (\bar{\tau}^\omega - \underline{\tau}^\omega))$.

Second-Best Policies. In Section 5.5, we focused on interventions that solely target the economy's information structure. In this section, we consider a broader set of interventions, which are still constrained to take the firms' pricing strategies as given.

As we have already shown, conditional on the allocations induced by privately optimal mechanisms—i.e., the shares α_j^l in Equations (38) and (39)—the only deviation from social surplus maximization occurs because firms make their information and input choices based on the revenue-based demand shifter, δ_j^R, rather than the social-surplus based shifter, δ_j^S. A policy correcting this misalignment is therefore sufficient to achieve "second-best efficiency."

To this end, consider firm j's ex-post profits, augmented by a revenue subsidy S_j:

$$\pi_j = (1+S_j)\cdot\left(\delta_j^R\cdot A_j\right)^{\frac{\theta-1}{\theta}}\cdot n_j^{\frac{\theta-1}{\theta}}\cdot C^{\frac{1}{\theta}} - w\cdot n_j - w\cdot\chi\cdot\iota_j. \tag{A58}$$

By setting the subsidy to:

$$S_j = S(\gamma_j) = \left[\frac{\delta^S(\gamma_j)}{\delta^R(\gamma_j)}\right]^{\frac{\theta-1}{\theta}} - 1, \tag{A59}$$

the social planner ensures that firm j maximizes social surplus when making its information and input choices, (x_j, n_j, ι_j) (Equation 43). Among interventions that maintain the privately optimal trading mechanisms, the subsidy scheme $\{S_j\}_j$, financed through lump-sum taxes on households achieves the highest social welfare. Although this "second-best" policy is more complex and requires policymakers to understand the composition of consumers faced by each firm, it exhibits parallels with the optimal data regulation discussed in Section 5.5.

First, consider the case where rent extraction is severe but not socially destructive, i.e., $\delta^S(\bar{\gamma})/\delta^R(\bar{\gamma}) < \delta^S(\underline{\gamma})/\delta^R(\underline{\gamma})$ but $\delta^S(\bar{\gamma}) > \delta^S(\underline{\gamma})$. Proposition 7 demonstrates that a planner constrained to data regulation imposes a positive information tax. The subsidy in Equation (A59) extends this logic further. Since the economy's factor supply is fixed, these subsidies—being smaller for information-producing firms—implicitly tax information production. Yet, their proportionality to firm size ensures that firms' employment choices also reflect their contribution to social surplus. In particular, since $\delta^S(\bar{\gamma})/\delta^R(\bar{\gamma}) < \delta^S(\underline{\gamma})/\delta^R(\underline{\gamma})$, the social return to information (and to employment) within informed firms is lower than its private counterpart. Hence, it is straightforward to show that both information production and the relative size of information-producing vs non-producing firms declines after the intervention.

Second, consider the case where information production about demand is socially destructive, i.e., $\delta^S(\bar{\gamma}) < \delta^S(\underline{\gamma})$. Proposition 8 demonstrates that a planner limited to data regulation would garble firms' information about consumer tastes, reducing the share of H-type consumers faced by firms and thereby the likelihood of socially destructive rent extraction. The subsidy in Equation (A59) achieves a similar effect but, instead, by discouraging firms from customizing their products to consumer tastes. That is, firms, recognizing that post-subsidy profits are higher when facing a lower share of H-type consumers, choose to produce the variety types that, paradoxically, are *less likely* to match consumer preferences. $\square$

D Model Validation and Quantification

D.1 Sectoral Misallocation

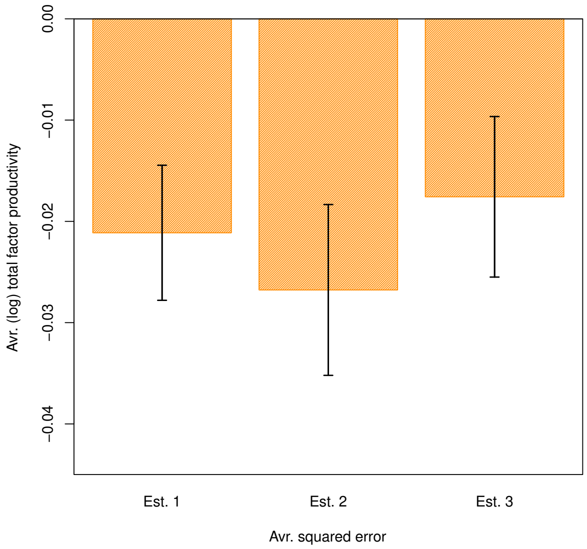
We define sectors by their four-digit NAICS industry classification. Building on the framework developed by Hsieh and Klenow (2009) and Gopinath *et al.* (2017), and consistent with our model setup, we compute our measures assuming a Cobb-Douglas production technology and monopolistic competition with CES demand. The profit-maximizing choice of an input for firm $i = \{1, 2, ..., N_s\}$ in sector $s = \{1, 2, ..., S\}$ at time $t = \{1, 2, ...\}$, thus, equates its marginal revenue product with its (sector-specific) cost. We assume the presence of two factors of productions, capital and labor. As a baseline and for comparability with our model below, we set the labor share α equal to 2/3, corresponding to the average labor share in the U.S. Our measure of misallocation is, nevertheless, not affected by the assumption that α is common across sectors, as these measures exploit within-sector variation of firm-level outcomes. Following the above terminology, we define *revenue-based total factor productivity* (TFPR) as revenue divided by output net of firm total factor productivity (TFP).[45] The *marginal revenue product of labor and capital* (MRPN and MPRK, respectively) are, by contrast, defined as revenue divided by labor and capital employed by the firm, respectively. We take revenue, labor, and capital stock measures from the I/B/E/S-Compustat sample (Appendix A.1). Panel (a) in Figure D.1 reports the average cross-sectional dispersion in mrpn, mrpk, and tfpr.[46] We report these estimates separately for accurate and inaccurate firms. We define "an informed" firm as one that (i) is below the median in terms of the mean-squared-error of one-year-ahead log-revenue expectations; and (ii) one for which we have at least three observations.

[45]The production technology used by firm j is $y_{j,s,t} = A_{j,s,t} \cdot n^{\alpha}_{j,s,t} \cdot k^{1-\alpha}_{j,s,t}$, $\alpha \in (0, 1)$, where $y_{j,s,t}$ is firm output, $l_{j,s,t}$ and $k_{j,s,t}$ the amount of labor and capital employed, respectively, and $A_{j,s,t}$ is firm total factor productivity. Let $p_{j,s,t}$ be the firm-specific product price. Then *revenue-based total factor productivity* is defined as: $\text{TFP}R_{j,s,t} \equiv \frac{p_{j,s,t} \cdot y_{j,s,t}}{n^{\alpha}_{j,s,t} \cdot k^{1-\alpha}_{j,s,t}}$. The marginal revenue product of labor and capital are, by contrast: $\text{MRPN}_{j,s,t} \equiv \kappa_L \cdot \frac{p_{j,s,t} \cdot y_{j,s,t}}{n_{j,s,t}}$ and $\text{MRPK}_{j,s,t} \equiv \kappa_k \cdot \frac{p_{j,s,t} \cdot y_{j,s,t}}{k_{j,s,t}}$, where $\kappa_L \in \mathbb{R}_+$ and $\kappa_K \in \mathbb{R}_+$ are common constants that depend on the elasticity of substitution and the labor share and capital share, respectively.

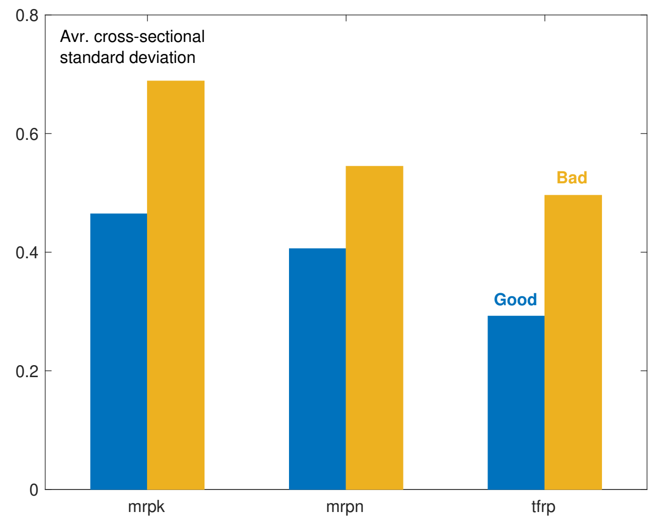
[46]We compute cross-sectional dispersion measures in two steps. First, we compute the standard deviation across firms j in a given sector s and year t. Second, for each year, we measure dispersion for the whole economy as the weighted average of dispersions across sectors. We give each sector a time-invariant weight equal to its average share in overall employment.

D.2 Qualitative Validation

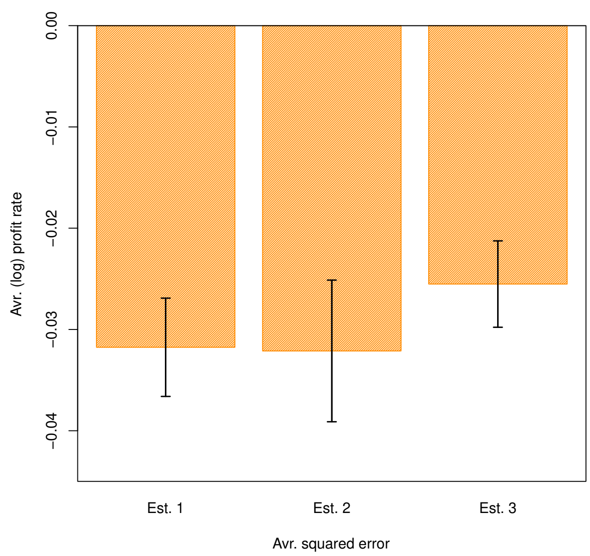
Figure D.1: Outcomes and accuracy



Panel (a): tfp-accuracy relationship



Panel (b): misallocation

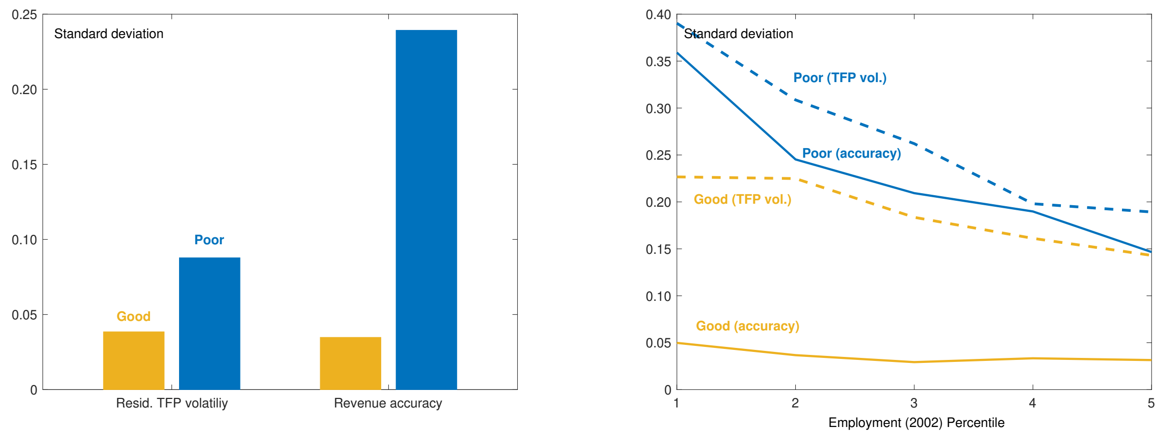


Panel (c): profit-accuracy relationship

Note: Panel (a) plots the coefficient from a regression of average firm (log-) total factor productivity on the mean-squared-error of one-year-ahead (log-)revenue forecasts (Est. 1) for the I/B/E/S-Compustat sample. The second column controls for firm age and sector fixed effects (Est. 2), while the third column trims TFP outlier observations at the 1 percent level (Est. 3). Panel (b) shows the average cross-sectional dispersion (standard deviation) in $\text{mrpn} \equiv \log(\text{MRPN})$, mrpk, and tfpr. Sectors are defined by their four-digit NAICS industry classification, and are weighted by their average share in overall employment. We define a "Good" firm as one that (i) is below the median in terms of the mean-squared-error of log-revenue expectations; and (ii) one for which we have at least three observations. We compute $\text{tfpr}_j = 1/3 \cdot \text{mrpk}_j + 2/3 \cdot \text{mrpn}_j$. Panel (c) plots the results from analogous estimates to those in Panel (a) of the average (log-)profit rate for an individual firm on its mean-squared-error and controls. Whisker intervals correspond to one-standard deviation robust (clustered) standard errors. Sample: 2002-2022.

D.3 Auxiliary Model Validation

Figure D.2: Forecast accuracy and productivity volatility



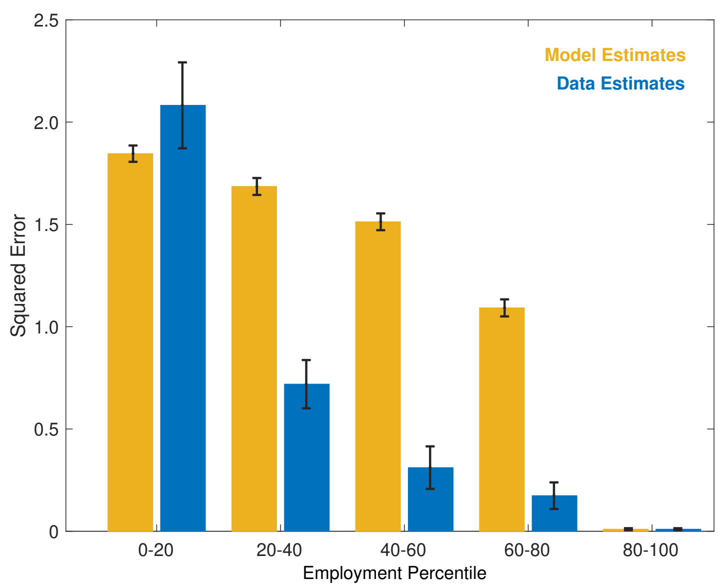
Panel (a): accuracy relationship

Panel (b): size quintiles

Notes: Panel (a) shows the difference between a "Good" and "Poor(ly)" informed firm's residualized (log-)tfp volatility, controlling for size, time, and and sector (NAICS-4) fixed effects. We define a "Good" firm as one that is below the 20th percentile of the mean-squared error distribution of one-year-ahead (log-) revenue forecasts. A "Poor" firm is one that is above the 80th percentile. Panel (b) shows the difference within each size quintile. Sample: 2002–2022 I/B/E/S-Compustat merger, described in Appendix A.1.

D.4 Quantitative Validation

Figure D.3: Accuracy Across the Size Distribution



Note: The figure shows the estimated relationship between squared normalized log. errors and firm size (quintiles of the initial employment distribution). We estimate this relationship both in the data and in the tmodel (Section 2). The bars labeled data control for sector and time fixed effects (Column 3 in Table I). Whisker intervals correspond to one-standard deviation robust (clustered) standard errors. Sample: 2002-2022.

D.5 Model Parametrization

Table D.1: Model parameters

Parameters	Value
Externally calibrated parameters:	
Frisch elasticity (ψ)	0.00
Labor endowment (N)	1.00
Internally calibrated parameters:	
Variance of ex-ante log-productivity (τ_μ^{-1})	1/100
Variance of log-productivity shock (τ_v^{-1})	1/8
Symmetric household demand shifter ($\hat{\delta}$)	0.034
Accuracy of low productivity information ($\underline{\tau}_\nu$)	17.0
Accuracy of high productivity information ($\bar{\tau}_\nu$)	620.0
Accuracy of low demand information ($\underline{\tau}_\omega$)	0.50
Accuracy of high demand information ($\bar{\tau}_\omega$)	0.86
Share of $H-$types after information production ($\bar{\gamma}$)	1.00
Share of $H-$types before information production ($\underline{\gamma}$)	0.00
Fixed cost of information (χ)	0.14
Share of information firms ($1-\Phi(z^\star)$)	0.10

D.6 The Extended Model

We augment the rent-extracting economy with capital and variety accumulation. To do so, we make four changes to the model. First, we assume household preferences are defined over consumption streams over time, $\{C_t^i\}$, and can be characterized by the utility function:

$$\mathcal{U}^i = \sum_{t=0}^{\infty} \beta^t \cdot \left[\log\left(C_t^i\right)\right], \tag{A60}$$

where $\beta \in (0,1)$ and C_t^i is the consumption aggregator defined by Equation (9) in every period. Second, we assume that each variety j at time t is produced as follows: the monopolistic firm j chooses the type x_{jt} and quantity y_{jt} of its variety to produce in accordance with:

$$y_{jt} = A_{jt} \cdot k_{jt}^{\alpha} \cdot n_{jt}^{1-\alpha}, \tag{A61}$$

where $\alpha \in (0,1)$ and k_{jt} denotes firm j's use of the capital input at time t. Capital depreciates at rate $\rho \in (0,1)$ and can be rented from households in a competitive rental market at rate r_t. Third, in this two-factor economy, we suppose that information production requires the allocation of χ units of the numeraire good to it. Finally, we assume that creation of varieties requires f units of the numeraire, but that varieties become obsolete at rate $\eta \in (0,1)$ in every period. The rest of the setup is the same as before. Namely, information production yields a firm signals about productivity and demand, which are now also assumed to be iid over time. We also assume that consumer preference shocks are iid over time. Investments into capital, varieties, and information take one-period to materialize.

D.6.1 Steady State Equilibrium

We focus on the steady state equilibrium of this economy.

Firm-level choices. At *Stage 3*, the solution to the optimal mechanism is the same as before, implying that, when making its choices (x_j, n_j, k_j, ι_j), a firm faces the revenue-based demand shifter δ_j^R. At *Stage 2*, the firms will choose to set $x_j = s_j^{\omega}$, and labor and capital as follows:

$$n_j = \frac{1-\alpha}{w} \cdot \frac{w^{1-\alpha} \cdot r^{\alpha}}{(1-\alpha)^{1-\alpha} \cdot \alpha^{\alpha}} \cdot \exp^{(\theta-1)\cdot\mu_j} \cdot g^R\left(\boldsymbol{\tau}_j\right)^{\theta-1} \cdot \Omega, \tag{A62}$$

and

$$k_j = \frac{\alpha}{r} \cdot \frac{w^{1-\alpha} \cdot r^{\alpha}}{(1-\alpha)^{1-\alpha} \cdot \alpha^{\alpha}} \cdot \exp^{(\theta-1)\cdot\mu_j} \cdot g^R\left(\boldsymbol{\tau}_j\right)^{\theta-1} \cdot \Omega, \tag{A63}$$

where the market size faced by firms is now defined as:

$$\Omega \equiv C \cdot \left(\frac{\theta}{\theta-1} \cdot \frac{w^{1-\alpha} \cdot r^{\alpha}}{(1-\alpha)^{1-\alpha} \cdot \alpha^{\alpha}}\right)^{-\theta}, \tag{A64}$$

while $g^R(\cdot)$ is the revenue-based information shifter defined in the main text. At *Stage 1*, firm j's expected profits for a given information choice can be expressed as:

$$\mathbb{E}\left[\pi_j | \mu_j, \boldsymbol{\tau}_j\right] = \frac{1}{\theta-1} \cdot \exp^{(\theta-1)\cdot\mu_j} \cdot g^R\left(\boldsymbol{\tau}_j\right)^{\theta-1} \cdot \Omega \cdot \frac{w^{1-\alpha} \cdot r^{\alpha}}{(1-\alpha)^{1-\alpha} \cdot \alpha^{\alpha}}, \tag{A65}$$

implying that a firm produces information if and only if

$$\beta \cdot \frac{1}{\theta-1} \cdot \exp^{(\theta-1)\cdot\mu_j} \cdot \left(g^R\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} - g^R\left(\underline{\tau}\right)^{\theta-1}\right) \cdot \Omega \geq \frac{\chi}{\frac{w^{1-\alpha} \cdot r^{\alpha}}{(1-\alpha)^{1-\alpha} \cdot \alpha^{\alpha}}}. \tag{A66}$$

Aggregate Implications. The aggregate TFP of this two-factor economy is given by:

$$\mathcal{A}^{\star} = \mathcal{E}\left(\mu^{\star}\right) \cdot \mathcal{A}\left(\mu^{\star}, g^{R}, M\right) \tag{A67}$$

where:

$$\mathcal{A}\left(\mu^{\star}, g^{R}, M\right) = M^{\frac{1}{\theta-1}} \cdot \exp^{\frac{\theta-1}{2} \cdot \frac{1}{\tau_{\mu}}} \cdot \left(\xi\left(\mu^{\star}\right) \cdot g^{R}\left(\bar{\boldsymbol{\tau}}\right)^{\theta-1} + \left(1-\xi\left(\mu^{\star}\right)\right) \cdot g^{R}\left(\underline{\boldsymbol{\tau}}\right)^{\theta-1}\right)^{\frac{1}{\theta-1}}, \tag{A68}$$

and where $\xi(\cdot)$ is given by Lemma 2. Thus, due to a "love-of-variety effect", the mass of varieties, M, also enters into the economy's TFP. To ensure that consumption growth is zero, the rate of return on capital must satisfy:

$$r = \beta^{-1} - 1 + \delta. \tag{A69}$$

The wage rate must in turn be such that the labor market clears:

$$N = \mathcal{A}\left(\mu^{\star}, g^{R}, M\right)^{\theta-1} \cdot \Omega \cdot \left(\frac{r}{w} \cdot \frac{1-\alpha}{\alpha}\right)^{\alpha}. \tag{A70}$$

Aggregation of individual capital holdings implies that the aggregate capital stock, K, equals:

$$K = \mathcal{A}\left(\mu^{\star}, g^{R}, M\right)^{\theta-1} \cdot \Omega \cdot \left(\frac{r}{w} \cdot \frac{1-\alpha}{\alpha}\right)^{\alpha-1}, \tag{A71}$$

while aggregate consumption is:

$$C = \mathcal{A}^{\star} \cdot K^{\alpha} \cdot N^{1-\alpha} - \delta \cdot K - M \cdot \left[\chi \cdot \Phi\left(-\mu^{\star} \cdot \sqrt{\tau_{\mu}}\right) + \eta \cdot f\right]. \tag{A72}$$

We note that C captures both the costs of information production, and of capital and variety creation, respectively. Finally, in equilibrium, as all potential entrant varieties are ex-ante identical, it follows that expected profits (including entry costs) must equal to zero:

$$\chi \cdot \Phi\left(-\mu^{\star} \cdot \sqrt{\tau_{\mu}}\right) + \left(1 - \beta \cdot (1-\eta)\right) \cdot f = \frac{1}{\theta-1} \cdot \frac{\mathcal{A}\left(\mu^{\star}, g^{R}, M\right)^{\theta-1}}{M} \cdot \beta \cdot \Omega \cdot \frac{w^{1-\alpha} \cdot r^{\alpha}}{(1-\alpha)^{1-\alpha} \cdot \alpha^{\alpha}}. \tag{A73}$$

Equations (A62) through (A73) fully characterize the steady state of this extended economy.

D.6.2 Extended Calibration

We set the fixed cost of entry, f, so that the mass of firms in the steady-state of the model equals one (i.e., the mass in the rent-extracting and baseline economy). We set the discount factor (β), the capital depreciation rate (ρ), the capital share (α), and the exit rate (η) to standard values used in the literature (0.96, 0.02, 0.33, 0.02, respectively). The rest of the parameters are calibrated in the same manner as in the rent-extracting and baseline economy.

D.6.3 Extension and Decomposition

Table D.2: Extended model: decomposition of the rise in TFP

Model	Overall (%)	Scale	Product	Pricing	Variety	$\underline{\gamma}$	$\bar{\gamma}$
Best case	6.9	5.0	2.3	0.0	-0.4	0.00	1.00
Worst case	2.7	5.0	2.3	-1.9	-2.6	0.70	0.85

Note: The table decomposes the rise in TFP in Figure 8 into its four constituent channels: (i) scale, (ii) product design, (iii) pricing, and (iv) variety generation using Equation (A67). The table does so for the "best-case" and for the "worst-case" economy identified as in the main text.

D.7 Alternative tfp Process

Table D.3: Alternative calibration

Parameters	TFP (%)	Welfare (%)
All parameters	(11.1, 13.1)	(7.1, 5.5)
Productivity only	(4.2, 4.3)	(4.5, 4.6)
Information only	(5.5, 6.9)	(2.2, 0.8)

Note: The table shows the effects of (i) changing both information and productivity parameters, as described in the main text (i.e., "all" parameters); (ii) changing only productivity parameters (i.e., τ_μ and τ_a); (iii) changing only information parameters. The table shows results both for the best- and worst-case scenarios.

References

Abis, S. and Veldkamp, L. (2024). The changing economics of knowledge production. *The Review of Financial Studies*, **37** (1), 89–118.

Adams, J. J., Fang, M., Liu, Z. and Wang, Y. (2024). The rise of ai pricing: Trends, driving forces, and implications for firm performance.

Ali, S. N., Lewis, G. and Vasserman, S. (2020). Voluntary disclosure and personalized pricing. *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 537–538.

Angeletos, G.-M., Huo, Z. and Sastry, K. A. (2021). Imperfect macroeconomic expectations: Evidence and theory. *NBER Macroeconomics Annual*, **35** (1), 1–86.

— and Lian, C. (2016). Incomplete information in macroeconomics: Accommodating frictions in coordination. *Handbook of Macroeconomics*, **2**, 1065–1240.

— and Pavan, A. (2007). Efficient Use of Information and Social Value of Information. *Econometrica*, **75** (4), 1103–1142.

Arias, A., Hansen, G. D. and Ohanian, L. E. (2007). Why have business cycle fluctuations become less volatile? *Economic theory*, **32**, 43–58.

Asriyan, V., Fuchs, W. and Green, B. (2017). Information spillovers in asset markets with correlated values. *American Economic Review*, **107** (7), 2007–2040.

—, — and Lorecchio, C. (2025). Multilateral bargaining with information spillovers. In *Technical Report, Working Paper*.

Autor, D., Dorn, D., Katz, L. F., Patterson, C. and Van Reenen, J. (2020). The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics*, **135** (2), 645–709.

Babina, T., Fedyk, A., He, A. and Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, **151**, 103745.

Baker, S. R., Johnson, S. G. and Kueng, L. (2020). Financial returns to household inventory management. *National Bureau of Economic Research*.

Baley, I. and Veldkamp, L. (2025). *The Data Economy: Tools and Applications*. Princeton University Press.

Begenau, J., Farboodi, M. and Veldkamp, L. (2018). Big data in finance and the growth of large firms. *Journal of Monetary Economics*, **97**, 71–87.

Blanchard, O. J., L'Huillier, J.-P. and Lorenzoni, G. (2013). News, noise, and fluctuations: An empirical exploration. *American Economic Review*, **103** (7), 3045–70.

Bloom, N., Brynjolfsson, E., Foster, L., Jarmin, R., Patnaik, M., Saporta-Eksten, I. and Van Reenen, J. (2019). What drives differences in management practices? *American Economic Review*, **109** (5).

—, Floetotto, M., Jaimovich, N., Saporta-Eksten, I. and Terry, S. J. (2018). Really uncertain business cycles. *Econometrica*, **86** (3), 1031–1065.

Boar, C. and Midrigan, V. (2024). Markups and inequality. *Review of Economic Studies*, p. rdae103.

Bornstein, G. and Peter, A. (2024). Nonlinear pricing and misallocation.

Brynjolfsson, E., McAfee, A., Sorell, M. and Zhu, F. (2008). Scale without mass: business process replication and industry dynamics. *Harvard Business School Technology & Operations Mgt. Unit Research Paper*, (07-016).

— and McElheran, K. (2016). The rapid adoption of data-driven decision-making. *American Economic Review*, **106** (5), 133–139.

— and — (2024). The rapid adoption of data-driven decision-making. *Working Paper*.

Chahrour, R. and Jurado, K. (2018). News or noise? the missing link. *American Economic Review*, **108** (7), 1702–36.

Chen, C., Hattori, T. and Luo, Y. (2023). Information rigidity and elastic attention: Evidence from japan.

Chen, H., Li, X., Pei, G. and Xin, Q. (2024). Heterogeneous overreaction in expectation formation: Evidence and theory. *Journal of Economic Theory*, p. 105839.

Chiavari, A. and Goraya, S. (2023). The rise of intangible capital and the macroeconomic implicationsn. *mimeo*.

Coyle, D. and Hampton, L. (2024). 21st century progress in computing. *Telecommunications Policy*, **48** (1), 102649.

Crémer, J. and McLean, R. P. (1988). Full extraction of the surplus in bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society*, pp. 1247–1257.

David, J. M., Hopenhayn, H. A. and Venkateswaran, V. (2016). Information, misallocation, and aggregate productivity. *The Quarterly Journal of Economics*, **131** (2), 943–1005.

— and Venkateswaran, V. (2019). The sources of capital misallocation. *American Economic Review*, **109** (7), 2531–2567.

De Loecker, J., Eeckhout, J. and Unger, G. (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, **135** (2), 561–644.

Dixit, A. K. and Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *The American economic review*, **67** (3), 297–308.

Edmond, C., Midrigan, V. and Xu, D. Y. (2023). How costly are markups? *Journal of Political Economy*, **131** (7), 000–000.

Eeckhout, J., Barcelona, U., Fu, C., Li, W. and Weng, X. (2024). Optimal taxation and market power.

— and Veldkamp, L. (2023). Data and markups: a macro-finance perspective. *UPF Working Paper*.

European Commission Report, E. (2020). A european strategy for data. *Strategy report*.

Farboodi, M., Haghpanah, N. and Shourideh, A. (2025). Good data and bad data: The welfare effects of price discrimination. *arXiv preprint arXiv:2502.03641*.

— and Veldkamp, L. (2020). Long-run growth of financial data technology. *American Economic Review*, **110** (8), 2485–2523.

— and — (2024). *A model of the data economy*. Tech. rep., National Bureau of Economic Research Cambridge, MA, USA.

Feng, Y., Zhao, Y., Zheng, H., Li, Z. and Tan, J. (2020). Data-driven product design toward intelligent manufacturing: A review. *International Journal of Advanced Robotic Systems*, **17** (2).

Gill, S. S., Wu, H., Patros, P., Ottaviani, C., Arora, P., Pujol, V. C., Haunschild, D., Parlikad, A. K., Cetinkaya, O., Lutfiyya, H. *et al.* (2024). Modern computing: Vision and challenges. *Telematics and Informatics Reports*, p. 100116.

Goldfarb, A., Greenstein, S. M. and Tucker, C. E. (2015). *Economic analysis of the digital economy*. University of Chicago Press.

Gopinath, G., Kalemli-Özcan, Ş., Karabarbounis, L. and Villegas-Sanchez, C. (2017). Capital allocation and productivity in south europe. *The Quarterly Journal of Economics*, **132** (4), 1915–1967.

Gorton, G. and Ordonez, G. (2014). Collateral crises. *American Economic Review*, **104** (2), 343–378.

Graham, J., Meyer, B., Parker, N. and Waddel, S. (2023). The cfo survey. *Duke University mimeo*.

Hsieh, C.-T. and Klenow, P. J. (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics*, **124** (4), 1403–1448.

Jin, W. and McElheran, K. (2024). Economies before scale: It strategy and performance dynamics of young us businesses. *Management Science*.

Kehoe, P. J., Larsen, B. J. and Pastorino, E. (2018). Dynamic competition in the era of big data. In *Technical Report, Working Paper*, Stanford University Stanford, CA.

Kwon, S. Y., Ma, Y. and Zimmermann, K. (2023). 100 years of rising corporate concentration. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2023-20).

Lorenzoni, G. (2009). A theory of demand shocks. *American economic review*, **99** (5), 2050–2084.

Lucas, R. E. (1977). Understanding business cycles. *Essential readings in economics*, pp. 306–327.

Lucas, R. E. J. (1972). Expectations and the neutrality of money. *Journal of Economic Theory*, **4** (2), 103–124.

Maćkowiak, B. and Wiederholt, M. (2009). Optimal sticky prices under rational inattention. *The American Economic Review*, **99** (3), 769–803.

Malherbe, F. (2012). Market discipline and securitization.

Mankiw, N. G. and Reis, R. (2002). Sticky information versus sticky prices: a proposal to replace the nesw keynesian phillips curve. *The Quarterly Journal of Economics*, **117** (4), 1295–1328.

Mckinsey and Company (2023). Frontiers of the data economy?

Nevo, A. and Wong, A. (2019). The elasticity of substitution between time and market goods: Evidence from the great recession. *International Economic Review*, **60** (1), 25–51.

Nordhaus, W. D. (2008). Two centuries of productivity growth in computing. *The Journal of Economic History*, **67** (1), 128–159.

O'Neill, L. (2023). 10 companies that are using big data. *ICAS Working Paper*.

Ordonez, G. (2009). *Larger crises, slower recoveries: the asymmetric effects of financial frictions*. Citeseer.

Ottonello, P. and Winberry, T. (2020). Financial heterogeneity and the investment channel of monetary policy. *Econometrica*, **88** (6), 2473–2502.

Senga, T. (2018). A new look at uncertainty shocks: Imperfect information and misallocation. *Working paper*.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, **50** (3), 665–690.

Tanaka, M., Bloom, N., David, J. M. and Koga, M. (2020). Firm performance and macro forecast accuracy. *Journal of Monetary Economics*, **114**, 26–41.

Varian, H. R. (1989). Chapter 10 price discrimination. *Handbook of Industrial Organization*, vol. 1, Elsevier, pp. 597–654.

Veldkamp, L. and Chung, C. (2024). Data and the aggregate economy. *Journal of Economic Literature*, **62** (2), 458–484.

Woodford, M. (2002). *Imperfect Common Knowledge and the Effects of Monetary Policy*. Nber working papers, Department of Economics, Columbia University.

Zolas, N., Kroff, Z., Brynjolfsson, E., McElheran, K., Beede, D. N., Buffington, C., Goldschlag, N., Foster, L. and Dinlersoz, E. (2021). Advanced technologies adoption and use by us firms: Evidence from the annual business survey. *NBER Working Paper*.