# (Digital) Platform Data Power

Peter Young Won *

March 2026

**ABSTRACT**

How do digital platforms exploit their dual role as marketplace operators to strategically compete with third-party sellers? To answer this, I introduce the Data Power Index (DPI), a novel measure quantifying Amazon's accumulated informational advantage. Unlike traditional market power measures such as the HHI, the DPI reveals that tighter data environments compress seller margins and force exposed public firms to cut costs and accelerate innovation. Crucially, this informational advantage robustly predicts the timing and targets of Amazon's private-label entry. After Amazon's entry, incumbent sellers suffer sharp declines in profitability, while exposed firms face heightened risk. Ultimately, this data-driven encroachment creates an important welfare trade-off: aggregate consumer surplus expands, but it is systematically reallocated away from incumbent sellers.

JEL Classification: G14, G32, D83, L11, L13, L40

Keywords: Digital Platform, Data Power, Information Advantage, Private Label Entry, Competition, Consumer Welfare, Firm Risk

*Won: peter.young.won@eccles.utah.edu.

# Disclosure Statement

**Research Support:** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Conflict of Interest:** The author, Peter Y. Won, declares no financial relationships or other potential conflicts of interest that could inappropriately influence or bias the findings, interpretations, or conclusions of this study.

**Data Access and Responsibility:** As the sole author, I had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

# 1 Introduction

Over the past two decades, digital platforms have evolved from neutral intermediaries into active market participants, leveraging exclusive access to marketplace data to guide their own product entry strategies (Zhu and Liu, 2018; Martens, 2021). The economic implications of this dual role have recently culminated in landmark antitrust scrutiny, notably the Federal Trade Commission's (FTC) lawsuit against Amazon, which alleges that the platform exploits non-public seller data to unfairly advantage its own retail operations.[1] At the heart of this regulatory and academic debate is a profound economic trade-off. While superior platform information can resolve matching frictions and enhance consumer surplus through lower prices, the expropriation of incumbent data severely compresses third-party margins and threatens independent innovation. Yet, evaluating the net impact of this trade-off requires understanding exactly how platforms accumulate and deploy their informational advantage, a mechanism that remains largely unexplored in the empirical literature.

Evaluating this data-driven mechanism is empirically challenging because a platform's informational advantage is typically an unobservable state variable. While theoretical models conceptualize data as a productive capital asset that generates self-reinforcing market power (Arcidiacono et al., 2020; Bergemann et al., 2022), existing empirical studies predominantly document the consequences of platform integration without measuring the underlying data accumulation (Crawford et al., 2022; Bennati, 2024). To bridge this gap, this paper introduces a novel measure: the Data Power Index (DPI). Using Amazon as a canonical laboratory, the DPI quantifies the platform's accumulated informational precision derived from massive volumes of consumer interactions. By making this unobservable advantage measurable, I explicitly link data accumulation to the platform's entry decisions, the competitive responses of public firms, and the resulting redistribution of welfare.

My empirical analysis yields four main sets of findings that trace the lifecycle of this data-driven market power. First, I demonstrate that a stronger data environment inherently disciplines both marketplace sellers and public firms. In high-DPI categories, intensified price competition

[1] FTC v. Amazon.com, Inc. (2023). Complaint for Permanent Injunction, U.S. District Court for the Western District of Washington; see also Mattioli, D. (2020), "Amazon Scooped Up Data From Its Own Sellers to Launch Competing Products," *The Wall Street Journal*.

systematically compresses third-party margins, with a one-standard-deviation increase in the DPI lowering the average margin per product by 10.5%. To absorb this margin pressure, exposed U.S. public firms preemptively adopt leaner cost structures, significantly reducing operating expenses (SG&A) by 1.1 percentage points and accelerating the renewal of their knowledge base toward younger patent portfolios. Rather than competing head-on, these public firms attempt to out-innovate the platform's data advantage by shifting their R&D toward entirely new, unexploited technological frontiers. Crucially, these structural adaptations are not explained by traditional market concentration (HHI), confirming that the DPI captures a distinct, information-based channel of competition.

Second, I document that the DPI serves as a highly robust predictor of Amazon's strategic transition from a marketplace operator to a direct competitor. Even after controlling for average prices, margins, HHI, and platform-specific signals like review volume, the DPI remains highly significant in forecasting the timing and target category of Amazon's private-label entry. Economically, this reveals that big tech platforms do not merely chase historical profitability; instead, they act as informational arbitragers, deploying capital precisely into markets where their proprietary data eliminates demand uncertainty. This establishes accumulated data as the critical state variable governing platform integration.

Third, exploiting Amazon's staggered entry in a difference-in-differences framework, I causally identify the devastating consequences of this data-driven integration. Following Amazon's private-brand entry, incumbent third-party sellers experience a structural collapse, including an 11.9% drop in sales per product and a 17.4% decline in per-product profitability. Furthermore, this competitive shock heavily spills over into the broader corporate sector. U.S. public firms exposed to the entered categories face heightened financial uncertainty, manifesting as a 12% increase in total stock return volatility, a 10% rise in idiosyncratic risk, and a 9% increase in market beta. This indicates that platform entry is no longer just a localized product-market shock, but a systematic risk factor that fundamentally alters the cost of capital and risk profile of competing public firms.

Finally, translating these product-market and corporate-level shifts into aggregate welfare terms, I document a stark economic trade-off. While the consumer surplus generated by individual third-party stores contracts by 38.8% post-entry, the aggregate category-level consumer surplus significantly rises by 118%. This dynamic reveals the dual nature of informational monopoly: the plat-

form's data-driven entry enhances overall allocative efficiency through improved matching, but it acts as a mechanism of systematic wealth transfer. While the platform efficiently scales consumer access, it simultaneously acts as an algorithmic tollbooth, permanently extracting incumbent rents and deepening the distributional asymmetry within the digital economy.

Underpinning these findings is a novel, multi-dimensional dataset that overcomes traditional empirical limitations. By integrating over two decades of Amazon product reviews and ratings (2000–2022) with firm-level financial data from Compustat and intellectual property data from the USPTO, this study constructs a transparent, highly granular measure of data accumulation. I then leverage the staggered timing of Amazon's private-label entry across hundreds of product categories using a robust stacked difference-in-differences design, effectively isolating the causal impact of data-driven market power from pre-existing industry trends.

The contributions of this paper are threefold. First, this paper provides the first empirical quantification of the link between data accumulation and strategic entry in digital platform markets. Recent theoretical literature has elegantly modeled data as a networked, productive asset that generates self-reinforcing monopolies (Kirpalani and Philippon, 2020; Bergemann et al., 2022). However, empirical applications mapping these models to real-world platform behavior have severely lagged. By introducing the DPI as a theoretically grounded proxy for a platform's information advantage, this paper identifies information precision as the missing state variable in platform competition, offering the first empirical measure capable of mapping theoretical data capital into observable strategic entry.

Second, I bridge the gap between industrial organization literature and modern antitrust policy by jointly examining the consequences of information-based entry on prices, variety, and welfare. While recent empirical studies document the negative effects of Amazon's entry on third-party seller outcomes (Zhu and Liu, 2018; Crawford et al., 2022; Bennati, 2024), this paper complements and extends those findings by explicitly quantifying the redistribution of surplus across market sides. By demonstrating that informational power can simultaneously enhance allocative efficiency and deepen asymmetries in surplus extraction, this paper provides rigorous empirical evidence that addresses the exact distributional concerns raised by the FTC, placing economic measurement at the center of ongoing regulatory debates.

Third, this paper contributes to the broader corporate finance literature connecting product

market competition to firm risk and innovation dynamics (Aghion et al., 2005; Kogan et al., 2017; Makaew and Maksimovic, 2020; Baker et al., 2023). Rather than focusing solely on traditional balance-sheet metrics or marketplace-bound sellers, I highlight how platform-induced margin compression fundamentally forces U.S. public firms to implement leaner cost structures and rotate their patent portfolios to survive. By tracing how algorithm-driven competition propagates from the platform level directly to corporate risk exposures and innovation decisions, this paper offers new evidence on how digital monopolies reshape the strategic behavior of firms across the entire economy.

The remainder of the paper is organized as follows. Section 2 describes the data, details the construction of the Data Power Index (DPI), and examines how the DPI relates to product market and firm dynamics. Section 4 lays out the empirical strategy for predicting Amazon private-label entry. Section 5 studies post-entry adjustments across product-market outcomes, consumer welfare, and corporate responses. Section 6 concludes.

# 2 Data

## 2.1 Amazon Platform Data

The primary dataset used in this study is the Amazon Review Dataset (Hou et al., 2024), which contains comprehensive product- and review-level information from May 1996 to September 2023. The dataset provides (i) product-level metadata, including product features, category, seller (store) identifiers, average number of reviews, average rating, and listing price, and (ii) consumer-level information, such as review text, rating, verified purchase indicator, review timestamp, and helpfulness votes. The dataset contains approximately 570 million reviews across 34 major product categories, offering rich and granular information that enables identification of activity from the consumer to the seller level.

On the Amazon platform, each product is uniquely identified by a "Parent ASIN." Variants of the same product, such as different colors or sizes, are assigned distinct "ASINs." Accordingly, Parent ASIN serves as the main identifier for product-level analysis in this study. On the consumer side, each review record includes the date and time of submission, the review content (if

provided), the rating, and whether the review is verified through an actual purchase. Even if no textual comment is provided, any numerical rating constitutes a review entry, ensuring coverage of all consumer feedback events.

The product-level metadata do not provide explicit information on the time span of a product's availability. If a product has never received a review, its metadata exist but its consumer review record is missing. Therefore, I approximate a product's lifecycle by using the timestamp of the first review as its market entry date and the timestamp of the last review as its exit date. Due to the data limitation, I acknowledge that this procedure introduces potential sample selection bias, as products without consumer reviews are under-represented.

To construct the dataset, product metadata were merged with consumer review data using Parent ASIN as the unique identifier. The merged data were then aggregated at the monthly frequency. For each product-month, I computed the mean or total number of product prices, consumer ratings, reviews, verified reviews, and etc. Because each product record identifies its associated store, I aggregate and construct store-level variables such as the number of active products, total and verified reviews, and average product price.

Another data limitation is that the direct information on the number of units sold is unavailable in the dataset. Hence, the number of verified reviews was used as a proxy for sales, following prior studies that link verified purchase reviews to actual transactions. The total number of reviews may be influenced by marketing campaigns, fake reviews, or non-purchasing activities. Therefore, focusing on verified reviews allows for a more accurate representation of realized sales, although I acknowledge that there is a gap between realized sales and our sales proxy.

Although fee-related data are not included in the review dataset, Amazon's fee structure plays an important role in determining seller profitability and strategic behavior. Amazon's fees can be broadly categorized into three components. First, the Referral Fee is a percentage of the sales price that varies by product category, typically ranging from 8% to 15% (e.g., around 8% for Electronics and up to 15% for Beauty or Clothing). Second, the Fulfillment Fee (FBA Fee) applies to sellers who use the Fulfillment by Amazon (FBA) service and depends on product weight and size tier. It includes storage and shipping charges, with additional long-term storage fees for inventory held over extended periods. Third, Subscription and Other Service Fees cover the monthly Professional Seller plan fee ($39.99), advertising (Sponsored Products), returns processing, and

packaging costs.

Using official documentation from Amazon Seller Central[2] , I compiled these fee components at the category-month level. Specifically, I computed the average Referral Fee rate and FBA utilization rate by category to capture variation in seller cost structures. This categorization allows us to construct measures of seller margin and margin rate in addition to sales proxies, thereby enabling a more complete analysis of product-market dynamics on the Amazon platform.[3]

Before constructing the final sample from the Amazon dataset, this study excludes several product categories that do not align with the characteristics of standard physical goods markets. Specifically, the categories "Books," "CDs and Vinyl," "Digital Music," "Gift Cards," "Kindle Store," "Magazine Subscriptions," "Movies and TV," and "Subscription Boxes" are omitted from the sample. This is because categories such as "Books," "Music," and "Movies" often list the author or artist as the store rather than the actual manufacturer or distributor, making them structurally distinct from typical retail products. In addition, categories such as "Gift Cards," "Kindle Store," and "Subscription Boxes" do not represent tangible manufactured goods and therefore differ fundamentally in their pricing, production, and market dynamics. To maintain consistency with the focus on product-based retail markets, these categories are excluded from the final analytical sample. After these data filtering process, I have 10,811,312 observations in our final sample.

## 2.2 Descriptive Statistics of Amazon Product Market

Table 1 presents summary statistics of key variables constructed at the store–month level using the Amazon dataset. The final sample consists of 156,831 unique stores. Overall, the distribution of sales activity across stores on Amazon is highly skewed. The average number of products per store is 4.19, but the standard deviation is large (9.69), and the 90th percentile reaches 10 products. This indicates a highly unequal structure in which most sellers manage only a few products, while a small number of large sellers operate extensive product portfolios. Consistent with this skewed size distribution, the average number of product entries and exits per store-month is also low around

[2] `https://sellercentral.amazon.com/help/hub/reference/external/G200336920`

[3] It should be noted that I cannot observe the true internal production or overhead costs of individual stores. Therefore, the "margin" discussed here refers strictly to a platform-level margin, calculated by deducting platform fees from the revenue. While incorporating actual store-level costs would be ideal, such data are practically unobservable. Thus, these figures must be interpreted purely as a platform-specific margin concept.

2.2 per month, while the dispersion is substantial. Even though the median is zero, meaning most sellers do not introduce or withdraw products in a typical month, the upper tail reaches up to five product entry or exit events. This further suggests that a small group of active, multi-product sellers drive most of the product turnover on the platform, whereas the majority of sellers operate with a stable and limited product set. Consumer review variables further illustrate disparities in seller activity. The average store receives 15.45 reviews per month. The average consumer rating is 2.6, lower than the median value of 3.58, reflecting a left-skewed distribution driven by a subset of negative reviews.

The average product price is about $42.4. The price distribution is wide, with the 10th percentile at $9.99 and the 90th percentile at $89.99, respectively. Because price information is available for roughly half of all products in the metadata, the number of observations for this variable is about 50% smaller than for other variables. Sales-related variables exhibit a similar degree of skewness. The average total monthly sales per store amount to $667.19, but the standard deviation is as high as $1,826.37, and the median is 119.50. For standardization, I construct sales per product, and it also shows substantial heterogeneity, with a mean of $80.10 and a 90th percentile of $184.43. These patterns collectively suggest a highly skewed marketplace where a small subset of products and sellers account for a disproportionate share of economic activity.

Margins and platform fees also display substantial dispersion across sellers. The average monthly margin per store is about $530, but the standard deviation is more than double the average ($1,494.55), and the 90th percentile rises to $1,131.59. Average margin per product is similarly low on average, only $8.87 per product, yet highly dispersed, with a standard deviation of $77.53 and a 90th percentile of $23.71. This indicates that most products generate monthly profits around 11% on average. This variation reflects the coexistence of low-cost private-label sellers, high-cost niche sellers, and sellers facing intense price competition within their categories.

Platform fees also exhibit substantial variation. Total FBA fees average $56 per store-month, but the standard deviation reaches $1,224 due to the presence of high-volume sellers who rely heavily on Amazon's fulfillment services. Storage fees show similar dispersion, with an average of $0.87 and a 90th percentile of $1.19, suggesting that only a small fraction of sellers carry large or slow-moving inventories. Referral fees, the most ubiquitous fee category, average $139.90 but vary considerably across sellers (standard deviation of $1,238.55), reflecting differences in sales levels,

listing price, and category-specific referral fee rates. Summing across categories, the total monthly fees paid by a store average nearly $197, with the 90th percentile exceeding $306, indicating that fee burdens increase sharply for high-volume sellers.

Overall, the dispersion in margins and fees highlights that Amazon's marketplace supports a highly heterogeneous group of sellers with markedly different cost structures, business models, and competitive environments. These patterns further reinforce the implication that platform outcomes are dominated by a relatively small subset of large, active sellers while the majority of sellers participate at a much smaller scale.

*[Insert Table 1 Here]*

## 2.3 Linkage with US Firm-level Variables

In the next step, I link the processed Amazon dataset to firm-level information for publicly listed U.S. companies. This linkage allows us to examine how firms operating on digital platforms engage in product market strategies and how such platform-level activities relate to their corporate performance and innovation behavior. In particular, this integrated dataset enables the analysis of how firm-level characteristics shape, and are shaped by, their platform-based brand operations.

To establish this linkage, I implement a multi-stage fuzzy matching procedure connecting Amazon sellers (stores) to actual firms. In the first stage, each Amazon store name (brand name) is matched with the corresponding trademark name in the United States Patent and Trademark Office (USPTO) dataset. The USPTO dataset contains detailed information on all registered trademarks in the United States, including the trademark text and the name of the trademark owner. In the second stage, the owner name is used to identify the corresponding firm in the Compustat database, thereby linking brand-level activities on Amazon to firm-level financial and organizational data.

Because Amazon requires sellers participating in its Brand Registry to hold a registered (or pending) trademark under the USPTO or an equivalent intellectual property authority, this institutional condition provides a strong basis for assuming correspondence between Amazon stores and USPTO-registered trademarks.[4] In this paper, the matching procedure employs a text similarity approach based on term frequency–inverse document frequency (TF-IDF) and cosine similarity to

[4] `https://sell.amazon.com/blog/brand-registry-requirements`

compute the closeness between store names and trademark names (e.g., Paulsen et al., 2023).[5] In this version, I conservatively retain only the matches with a perfect similarity score (100%) for the baseline analysis, while manually reviewing and refining the linkage to improve coverage and accuracy over time.

Through this process, I identify 552 unique publicly listed U.S. firms that can be reliably matched to Amazon stores. This firm–platform integrated dataset provides a novel foundation for analyzing how digital platform participation interacts with firm-level competition, innovation, and performance outcomes.

## 2.4 Descriptive Statistics of Firm-linked Variables

Table 2 summarizes firm-quarter level characteristics for the 552 publicly listed U.S. firms that are successfully linked to Amazon's product market data. All financial variables are drawn from quarterly Compustat and expressed in millions of U.S. dollars. Overall, the linked firms represent a broad cross-section of the U.S. corporate sector, encompassing both large, diversified enterprises and smaller, specialized companies that actively operate on Amazon's marketplace.

Firm size and asset distributions are highly skewed, reflecting the heavy right tail typical of publicly traded firms. The mean market capitalization is approximately \$64.9 billion, whereas the median is only \$5.6 billion, suggesting that a few mega-cap firms dominate the distribution. The book-to-market ratio exhibits a mean of 0.46 and a median of 0.36, consistent with a predominance of growth-oriented firms among those engaging with Amazon. A similar pattern appears in total assets, which average \$52.7 billion with a median of \$5.8 billion.

Sales and profitability measures also display substantial dispersion. Average quarterly sales amount to roughly \$5.4 billion, but the median is below \$1 billion, again indicating wide heterogeneity in firm scale. Net income averages \$818 million per quarter, with a median of only about \$54 million, underscoring that most firms in the sample are much smaller and less profitable than the largest outliers. Research and development expenditures show a mean of \$850.84 million and a median of \$45.43 million, implying that only a subset of firms invest heavily in innovation while

[5] Specifically, I compute character-level n-gram (2–5) TF-IDF vectors for both datasets and measure pairwise cosine similarity between store and trademark names. For each trademark, I identify the store name with the highest similarity score as its best match.

the majority spend relatively little on R&D.

Lastly, the number of Amazon stores linked to each firm, reveals notable variation in firms' platform engagement. On average, each firm operates about 15 distinct stores on Amazon, with a median of 3 and a 90th percentile of 63. This dispersion reflects heterogeneous platform strategies: while many firms maintain a limited online presence under a small number of brand storefronts, a minority of firms leverage Amazon extensively through multiple brand or product-line stores.

Overall, these statistics indicate that firms integrated with Amazon's marketplace are not confined to a particular segment of the corporate distribution. Instead, they encompass both large multinationals and smaller niche players, providing a comprehensive view of how participation in digital platforms interacts with firm-level financial and strategic characteristics.

*[Insert Table 2 Here]*

# 3 What Drives Digital Platforms?

The competitiveness of digital platforms depends less on traditional production factors or economies of scale and more on how broadly and precisely they can collect and utilize data (Martens, 2021; Bergemann and Bonatti, 2024). In traditional product market, market power is typically measured by prices, profit margins, or market concentration (e.g., HHI). Yet such indicators is expected to fail to capture the distinctive nature of the platform economy, where informational asymmetry between the platform and other market participants is the fundamental source of competitive advantage (Eisenmann et al., 2011; Demary and Rusche, 2018).

In platform markets, the primary asset is the accumulated data from consumers and products. By observing the entire spectrum of user behaviors, platforms and sellers continuously learn the structure of market demand in real-time (Parker and Van Alstyne, 2005; Anderson and Bedre Defolie, 2024). In particular, platforms such as Amazon can infer consumer preferences with high granularity from massive digital traces, such as search queries, product reviews, star ratings, clicks, and conversion rates, whereas individual sellers can only observe the performance of their own products (Dinerstein et al., 2018; Crawford et al., 2022).[6] This structural information asymmetry

[6]Although sellers can access certain market-level metrics through programs like Amazon Brand Analytics or rely on third-party estimation tools (e.g., Helium 10, Jungle Scout), they do not have the ability to purchase raw, user-level

constitutes the foundation of the platform's informational dominance within the market (Hagiu et al., 2022; Zennyo, 2022; Bergemann and Bonatti, 2024).

This data-driven competitive structure fundamentally differs from the mechanisms observed in traditional product markets. For instance, when offline retailers such as Walmart entered local markets, their competitive advantage largely stemmed from price reductions and supply-chain efficiency gains that enhanced consumer welfare (Basker, 2007; Hausman and Leibtag, 2007). In contrast, digital platform, such as Amazon, transforms market structure through data accumulation, improving the precision of demand forecasting and matching efficiency (Etro, 2021; Gutierrez Gallardo, 2021).

Accordingly, while Walmart's advantage is grounded in logistics and price competition, Amazon's strength lies in its predictive power and learning rate derived from data. This process benefits consumers by offering richer information and more personalized choices, yet simultaneously exposes sellers to copycat risks and downward pressure on profitability (Lee, 2013; Arcidiacono et al., 2020; Long and Amaldoss, 2024). Consequently, data power in platform markets is not merely a source of efficiency but a transformative force that reshapes both market competition and welfare distribution (Kirpalani and Philippon, 2020; Hagiu et al., 2022; Bergemann and Bonatti, 2024).

Building on these insights, I construct a Data Power Index (DPI) to empirically measure the informational advantage of digital platforms. Using this measure, I analyze (i) price and product-structure dynamics within the platform, (ii) consumer engagement and welfare responses, (iii) strategic and financial adjustments among U.S. public firms exposed to high-DPI categories, and (iv) whether Amazon's own brand entry decisions are systematically predicted by prior increases in category-level DPI. Overall, this section conceptualizes data power as the central competitive engine of the digital platform economy, and the following section demonstrates how it can be empirically employed.

platform data. Consequently, a fundamental information asymmetry persists: the platform observes the complete, granular customer journey across the entire ecosystem, while sellers are restricted to aggregated indices and their own localized sales data.

## 3.1 Theoretical Motivation of Data Power Index (DPI)

The conceptual foundation of the *Data Power Index* (DPI) rests on the economics of information in digital platforms. For example, Bergemann and Bonatti (2024) characterize data power not merely as the accumulation of volume, but as an informational advantage, specifically, the precision gap between a platform's knowledge of consumer preferences and the generic market priors. To formalize this concept and derive a structural form for measurement, I adopt the two-sided directed search framework of Kirpalani and Philippon (2020), which decomposes platform value creation into two orthogonal components: matching technology and information technology.

Consider a platform inter-mediating between a mass of buyers $N_b$ and sellers $N_s$. The total number of matches formed is governed by the platform's matching technology, denoted by $\alpha$. This parameter captures the efficiency of the infrastructure reducing search frictions. The matching function is given by:

$$M(N_b, N_s) = \alpha N_b^{\gamma} N_s^{1-\gamma}, \tag{1}$$

where $\alpha$ represents the baseline matching efficiency, effectively measuring the quantity of trade facilitated by the platform.

However, the volume of matches is insufficient to characterize value in digital markets. The quality of these matches depends on the platform's information technology, denoted by $\delta$. Kirpalani and Philippon (2020) model $\delta$ as the precision of the signal $\sigma$ that the platform generates to steer consumers toward their preferred variety $u_i$. Formally, if a consumer's true ideal variety is $u$, the platform observes a signal $\sigma$ with precision $\delta$:

$$\Pr(\sigma = i \mid u_i = u) = \delta. \tag{2}$$

Assuming rational Bayesian updating, the parameter $\delta$ maps directly to the posterior probability of a successful match. That is, conditional on the platform recommending product $i$ ($\sigma = i$), the probability that the consumer actually derives utility from it is exactly $\delta$:

$$\Pr(u_i = u \mid \sigma = i) = \frac{\Pr(\sigma = i \mid u_i = u)\Pr(u_i = u)}{\sum_j \Pr(\sigma = i \mid u_i = j)\Pr(u_i = j)} = \delta. \tag{3}$$

Thus, $\delta$ serves as a sufficient statistic for the platform's ability to identify specific consumer tastes

among heterogeneous varieties, interpreted effectively as the probability of *ex-post* consumer satisfaction. In other words, $\delta$ implies the platform's ability to resolve uncertainty about heterogeneous consumer tastes, such that it captures the precision of identifying 'who likes what' rather than identifying 'what is best'.

The critical implication of this framework is that the value generated for participants is multiplicative in these two dimensions. The equilibrium value functions for sellers ($V_s$) and buyers ($V_b$) are derived as:

$$V_s(\delta, n) \propto \alpha \cdot \delta \cdot n^{-\gamma}, \tag{4}$$

$$V_b(\delta, n) \propto \alpha \cdot \delta \cdot n^{1-\gamma}. \tag{5}$$

Equations (4) and (5) demonstrate that data power cannot be captured by matching efficiency ($\alpha$) or information precision ($\delta$) in isolation. A platform with high traffic but poor targeting (high $\alpha$, low $\delta$) generates low surplus, as does a platform with perfect precision but no scale (low $\alpha$, high $\delta$).

Consequently, to empirically gauge the platform's data power at the category level, any valid index must mirror this functional form. I therefore define the Data Power Index (DPI) not as a linear combination, but as the interaction of observable proxies for these two latent parameters:

$$\text{Data Power}_{st} \equiv f(\hat{\alpha}_{st}) \times g(\hat{\delta}_{st}). \tag{6}$$

In the subsequent empirical specification, I construct $\hat{\alpha}_{st}$ to proxy for the transactional friction and volume (matching technology) and $\hat{\delta}_{st}$ to proxy for the resolution of uncertainty regarding consumer preferences (information technology).

## 3.2 Constructing Data Power Index (DPI)

Building on the theoretical decomposition of platform value derived in Section 3.1, this section details the construction of the Data Power Index (DPI). I first define the baseline measure used throughout the main analysis, followed by a discussion of potential measurement concerns and the alternative specifications constructed to address them.

### 3.2.1 The Baseline Measure

The primary measure, simply denoted as *DPI*, is designed to capture the fundamental accumulation of information through two orthogonal dimensions: the capacity of the matching infrastructure and the precision of consumer targeting (Kirpalani and Philippon, 2020). To reflect the theoretical complementarity between these dimensions, I aggregate them using a geometric mean:

$$DPI_{s,t} = \sqrt{\underbrace{\psi\left(\frac{1}{|J_{s,t}|}\sum_{j\in s}\log(1+N_{j,t})\right)}_{\text{Matching Efficiency }(\alpha)} \times \underbrace{\psi\left(\frac{1}{|J_{s,t}|}\sum_{j\in s}\left[\text{Avg.Rating}_{j,t}\times\text{Entropy(Rating)}_{j,t}\right]\right)}_{\text{Information Precision }(\delta)}} \quad (7)$$

where $j$ indexes individual products within category $s$ at time $t$, and $|J_{s,t}|$ is the total number of products in that category-month. For the matching capacity ($\alpha$), $N_{j,t}$ represents the number of reviews for product $j$, and the log-transformed volumes are averaged across all products. For the information precision ($\delta$), $\text{Avg.Rating}_{j,t}$ represents the product's average consumer rating, and $\text{Entropy(Rating)}_{j,t}$ denotes the Shannon entropy of the 1-to-5 star rating distribution for product $j$. The precision component aggregates this interaction across all products in the category. The function $\psi(\cdot)$ denotes the Min-Max normalization operator, which scales each component to the unit interval $[0,1]$ before taking the geometric root. The use of the geometric mean imposes strict complementarity: a platform cannot possess true data power if either its transaction volume or its matching precision is near zero.

The justification for these proxies is grounded in the structural decomposition of platform value (Kirpalani and Philippon, 2020). First, the average product-level review volume proxies for the matching capacity ($\alpha$). Ideally, the most precise measure of a platform's matching efficiency would rely on the exact algorithms and granular user inputs, such as individual browsing histories, search queries, and click-stream data, that Amazon uses to match buyers and sellers. However, these algorithmic mechanics are closely guarded trade secrets, rendering the direct observation of the matching process practically impossible (OECD, 2022). Given this structural data limitation, the most viable empirical strategy is to measure the outcome of the matching technology rather than its underlying process.

While actual product-level transaction volumes are the definitive outcome of successful matching, they remain unobservable without proprietary internal data. Consequently, the closest empirical proxy for realized matches is consumer feedback. Because consumers can only leave a rating after completing a transaction, this feedback serves as a verifiable footprint of a successful match. Indeed, a well-established literature affirms that review volume acts as a robust proxy for actual sales and transactional liquidity (Chevalier and Mayzlin, 2006).

In the Kirpalani and Philippon (2020) framework, $\alpha$ represents the efficiency with which the platform converts buyer and seller traffic into successful transactions. By utilizing consumer feedback volume, we capture this exact dynamic. Furthermore, by calculating the log-volume at the product level before averaging, this specification prevents the metric from being overwhelmingly skewed by a single superstar product, ensuring that $\alpha$ captures broad and robust transactional liquidity across the category's catalog. A higher average review volume implies that the platform has successfully minimized the friction of bringing buyers and sellers together, effectively expanding the quantity of trade (Anderson and Bedre Defolie, 2024).

Second, to proxy for information precision ($\delta$), I utilize the product-level interaction of average consumer satisfaction and the dispersion of consumer opinions. Unlike matching capacity, information precision is theoretically tied to consumers' *ex-post* satisfaction, making the average rating a necessary empirical baseline to capture vertical match quality. However, using the average rating in isolation is structurally insufficient for measuring a platform's data power. While a high average rating successfully captures general product acceptability, indicating that a product is broadly appealing to the average consumer, it fails to reveal the nuances of idiosyncratic tastes, such as whether a product serves a specific niche segment with heterogeneous preferences.

Theoretical frameworks, including Kirpalani and Philippon (2020), emphasize that a platform's true informational advantage lies not in identifying universally good products, but in steering heterogeneous consumers toward their unique ideal varieties. Consequently, measuring data power requires capturing the extent to which a platform recognizes and satisfies this diversity of preferences. The inclusion of $\text{Entropy}(\text{Rating})_{j,t}$ exactly captures this capacity to resolve heterogeneity. Following the framework of Kirpalani and Philippon (2020), I interpret $\delta$ as the probability of successfully steering consumers toward their idiosyncratic ideal varieties. In a data-poor environment (low $\delta$), the platform relies on aggregate priors, maximizing match probability by promoting

generic products with broad, uncontroversial appeal. This strategic concentration results in homogenized feedback patterns, where reviews are uniformly concentrated at the maximum rating (e.g., universally 5-stars), yielding a low rating entropy.

Conversely, an informational advantage enables the platform to implement precision steering, matching specific consumers to niche, polarizing products that would otherwise face high search frictions. This granular matching resolves horizontal differentiation, mechanically increasing $\text{Entropy}(\text{Rating})_{j,t}$ as the product attracts a diverse set of idiosyncratic opinions. Indeed, Clemons et al. (2006) emphasize that in information-rich environments, such "hyper-differentiated" products, those eliciting polarizing, high-variance feedback, generate immense value precisely because they cater to passionate niche segments rather than generic preferences. However, opinion dispersion alone is insufficient; it is only a valuable signal of data power if the matches ultimately satisfy the targeted consumers.

This structural requirement necessitates the multiplicative functional form between average ratings and rating entropy. As mathematically demonstrated by Sun (2012), the informational value of rating variance is strictly contingent on the mean. A product with a low average rating and high entropy typically indicates poor or inconsistent quality. In contrast, the interaction of a high average rating and high entropy uniquely identifies a successful niche product: one that delivers exceptional utility to a specific consumer segment despite inherent, polarizing taste heterogeneity. Therefore, by multiplying these two metrics, the $\delta$ component acts as a stringent filter, mathematically rewarding the platform only when it successfully simultaneously resolves horizontal taste heterogeneity (high entropy) and ensures high vertical match quality (high average rating).

### 3.2.2 Addressing Measurement Concerns: Alternative Specifications

While the baseline *DPI* captures the structural dimensions of data power, one might raise concerns regarding the quality, authenticity, and depth of the underlying signals. For instance, products exhibiting both high average ratings and high opinion dispersion could arguably be inflated by unverified purchases, coordinated fake reviews, or superficial feedback rather than reflecting genuine, precise matching. To mitigate these concerns, I construct two alternative versions ($DPI^{VQ}$ and $DPI^{DQ}$) that explicitly account for signal quality, trust, and complexity by evaluating these metrics at the granular, product level.

First, to address the noise introduced by unverified purchases, $DPI^{VQ}$ (Verified Quality) incorporates the share of verified purchases directly into the precision component:

$$DPI^{VQ}_{s,t} = \sqrt{\underbrace{\psi(\alpha_{s,t})}_{\text{Matching}} \times \underbrace{\psi\left(\frac{1}{|J_{s,t}|}\sum_{j\in s}\left[\text{Avg.Rating}_{j,t} \times \text{Entropy(Rating)}_{j,t} \times \text{Valid}_{j,t}\right]\right)}_{\text{Augmented } \delta}} \tag{8}$$

where $\alpha_{s,t} = \frac{1}{|J_{s,t}|}\sum_{j\in s}\log(1+N_{j,t})$ represents the baseline volume capacity, $j$ indexes individual products within category $s$, $|J_{s,t}|$ is the total number of products, and $\text{Valid}_{j,t}$ denotes the share of verified purchases. By integrating $\text{Valid}_{j,t}$, this specification heavily penalizes categories where signals are distorted by fake reviews.

Another concern is that star ratings and their entropies are relatively coarse signals that may fail to capture the full nuance of consumer learning. In markets with complex products, the true data power lies in the rich, contextual details provided by users. To capture this depth of engagement, $DPI^{DQ}$ (Deep Quality) introduces a composite richness weight, defined as $\Omega_{j,t} = \text{Valid}_{j,t} \times \text{Img}_{j,t} \times (1+\text{Len}_{j,t})$:

$$DPI^{DQ}_{s,t} = \sqrt{\psi(\alpha_{s,t}) \times \psi\left(\frac{1}{|J_{s,t}|}\sum_{j\in s}\left[\text{Avg.Rating}_{j,t} \times \text{Entropy(Rating)}_{j,t} \times \Omega_{j,t}\right]\right)} \tag{9}$$

where $\Omega_{j,t}$ explicitly penalizes unverified reviews while rewarding visual evidence ($\text{Img}_{j,t}$, the average percentage of reviews with images) and detailed textual feedback ($\text{Len}_{j,t}$, the log-transformed average word count). A high $DPI^{DQ}$ thus identifies categories where the platform learns not just from a binary transaction outcome, but from the immense complexity and depth of the consumer feedback.

Despite these differences in construction, all three specifications share the same structural form derived from the Kirpalani and Philippon (2020) framework. As shown in the subsequent empirical results, the main findings remain robust across $DPI$, $DPI^{VQ}$, and $DPI^{DQ}$.

## 3.3 Stylized Facts of Data Power Index (DPI)

The Data Power Index (DPI) quantifies the platform's informational capacity and learning precision over time. Figure 1 plots the time-series evolution of the DPI from 2000 to 2022, split by Amazon's private label entry status. The red line represents the average DPI for categories that Amazon brands eventually entered, demonstrating a steady and significant increase in informational richness over time. In contrast, the blue line tracks the categories that remained untouched by Amazon brands, which exhibit a gradual decline in their DPI. This distinct divergence highlights that categories targeted by Amazon possess a structurally different trajectory of data accumulation compared to non-targeted ones, motivating subsequent analysis on how the platform leverages this informational advantage.

*[Insert Figure 1 Here]*

To further validate the empirical relevance of the DPI and explore its relationship with platform noise, this paper examines a specific exogenous shock, the COVID-19 pandemic. Figure 2 plots the monthly evolution of the overall DPI alongside the share of nonsense brands between December 2018 and December 2020. This case study yields two critical insights into the platform's informational dynamics.

First, the DPI exhibits a distinct spike corresponding to the onset of the COVID-19 pandemic. As global lockdown induces an unprecedented shock to offline retail, digital platforms experienced an explosive influx of both consumers and third-party sellers. The sharp increase in the DPI during this period demonstrates that this index effectively captures the sudden surge in platform-level data generation and informational richness driven by real-world macroeconomic phenomena.

Second, the figure highlights a dynamic between the DPI and the proliferation of nonsense brands.[7] While the share of nonsense brands shows a steady upward trend over the sample period, it fundamentally moves in the opposite direction of the DPI. Conceptually, nonsense brands, which frequently rely on counterfeit products, algorithmic naming conventions, and fake reviews

[7]To systematically identify these entities, this paper develops a hybrid detection algorithm. The algorithm first screens out established, legitimate entities using a comprehensive global brand whitelist (e.g., Fortune 500 and Forbes). For the remaining store names, it computes a nonsense score by evaluating linguistic anomalies, such as high consonant density, lack of vowels, unnatural character repetitions, and the use of rare letters, supplemented by an algorithmic gibberish detector (i.e., the *nostril* package) and non-English character penalization.

to sustain their business models, do not contribute valuable consumer signals. Instead, they act as informational pollution or noise that degrades the platform's learning efficiency. Consistent with this visual evidence, pairwise correlations confirm a meaningful negative relationship between the nonsense brand exposure and our DPI measures, with correlation coefficients ranging from $-0.08$ to $-0.19$ across the baseline and adjusted DPI versions (untabulated). This negative correlation supports the premise that a higher penetration of opportunistic, low-quality sellers structurally impairs the platform's data power by diluting the quality of actionable consumer feedback.

*[Insert Figure 2 Here]*

Next, Table 3 summarizes the cross-sectional properties of the three DPI measures. Panel A reports descriptive statistics based on 6,723 category–month observations from 2000 to 2022. The average values of the baseline $DPI$, $DPI^{VQ}$, and $DPI^{DQ}$ are 0.52, 0.51, and 0.49, respectively, with standard deviations ranging from 0.23 to 0.28. The 10th percentiles lie between 0.13 and 0.17, whereas the 90th percentiles range from 0.80 to 0.92, indicating substantial dispersion in the platform's informational environment across categories. This heterogeneity suggests that while some product categories accumulate rich and varied consumer feedback, providing Amazon with stronger informational advantages, others remain relatively data-sparse and exhibit lower learning efficiency.

Panel B reports the correlation matrix among the DPI measures. The three DPI versions are strongly positively correlated with one another, with pairwise correlations ranging from 0.75 to 0.91. This high correlation indicates that while $DPI^{VQ}$ and $DPI^{DQ}$ incorporate stricter adjustments for data credibility and dynamic learning, they capture closely related dimensions of the underlying fundamental data environment. These results confirm that the refined DPI measures reflect a robust and consistent variation in the platform's information accumulation.

*[Insert Table 3 Here]*

Overall, DPI exhibits a long-run, monotonic increase, consistent with the cumulative strengthening of Amazon's data-based learning capacity. Second, the hierarchical structure across DPI versions implies that incorporating stricter informational filters yields lower absolute values but greater interpretive precision. Third, the wide dispersion across categories highlights persistent

heterogeneity in the platform's data environment. Finally, the evolving trends suggest that Amazon's information infrastructure systematically matures over time.

## 3.4 How does DPI relate to Amazon Product Market Dynamic?

In this section, I examine how the Data Power Index (DPI) maps into store outcomes in Amazon's product market. The key idea is that a higher DPI today reflects a richer and more precise information environment at the category level, which allows the platform to match consumers to products more accurately tomorrow. If Amazon uses this informational advantage when ranking, recommending, or promoting products, then categories with higher DPI should exhibit systematically different store-level outcomes, such as prices, sales, margins, and product scope, in subsequent periods.

Empirically, I therefore focus on the lagged DPI rather than its contemporaneous value. Using a contemporaneous measure $\text{DPI}_{c,t}$ would raise timing and simultaneity concerns, because the index is partly constructed from reviews and transactions that are themselves outcomes of store performance in month $t$. In contrast, $\text{DPI}_{c,t-1}$ summarizes the information environment that the platform has already accumulated before month $t$ and can leverage when shaping demand allocation and competitive interactions in that month. Lagging DPI in this way helps mitigate look-ahead bias and mechanical feedback from $Y_{i,c,t}$ into the construction of the index, and it aligns more closely with the interpretation of DPI as a predetermined measure of data-based informational power.

To differentiate this information-based view of platform power to traditional notions of product market structure, I also compare DPI against the standard Herfindahl–Hirschman Index (HHI) at the category level. Comparing the explanatory power of DPI and HHI allows me to assess whether variation in store outcomes, conditional on fixed effects, is better accounted for by data-driven informational advantages or by conventional concentration-based measures of market power.

Formally, I estimate the following specification:

$$Y_{i,c,t} = \alpha + \beta \, \text{DPI}_{c,t-1} + X'_{i,c,t-1}\theta + \delta_i + \gamma_t + \varepsilon_{i,c,t}, \tag{10}$$

where $Y_{i,c,t}$ denotes a store $i$ outcome in category $c$ and month $t$ (e.g., logarithm value of store's average price, total sales per product, average margin per product, number of existing products,

and rolling three month of net new products). The independent variable, $\text{DPI}_{c,t-1}$, is the lagged category level DPI constructed in equation (7) to (8). The vector $X_{i,c,t-1}$ collects time varying controls such as the number of verified reviews and HHI, and I include store fixed effects $\delta_i$ and year–month fixed effects $\gamma_t$. Standard errors are clustered at the store–month level. I use *DPI* as our baseline index in the main text and show that all results are robust to alternative constructions $DPI^{DQ}$ and $DPI^{VQ}$ in the Online Appendix.

Table 4 reports the panel regression estimates linking the lagged category–month DPI to store outcomes in Amazon's marketplace. The specifications include the lagged category-level Herfindahl–Hirschman Index (HHI) as a standard proxy for product market concentration, along with lagged average ratings and verified reviews as additional controls, as well as store and year–month fixed effects. The results indicate a consistent expansion on the product–portfolio margin followed by a pro–competitive adjustment on the price–quantity margin and subsequent margin compression. Moreover, they show that DPI has substantially stronger explanatory power than HHI once both are included.

First, columns (1) and (2) examine the extensive margin of product scope, demonstrating how higher data power improves matching precision and attracts product entry. Column (1) shows that the number of existing products carried by a store increases with DPI, significant at the 1% level, with an estimated coefficient of 0.29. A one–standard–deviation improvement in DPI (0.28) is associated with an increase of 0.08 products, which is about 1.9% of the mean number of existing products (4.231). Column (2) similarly indicates that rolling three–month net new products rise with DPI, significant at the 1% level, with an estimated coefficient of 0.47. The implied effect of a one–standard–deviation increase in DPI is 0.13 additional net new products over three months, or about 144% of its mean value of 0.0910, highlighting the sensitivity of entry margins to the data environment. In contrast, HHI has a near-zero and insignificant coefficient in both specifications. These findings are consistent with the long-tail view that lower discovery and search costs expand variety by improving the match between niche demand and supply (Brynjolfsson et al., 2011).

Second, this influx of products and improved data precision intensify market competition, which subsequently drives down prices. Column (3) shows that a higher DPI in the previous month is associated with a statistically significant decline in next month's log average price at the 1% level, with an estimated coefficient of $-0.06$. This implies that a one–standard–deviation

increase in lagged DPI corresponds to a $-0.017$ change in log price. By contrast, the coefficient on lagged HHI is essentially zero and statistically insignificant, indicating that variation in prices, conditional on store and time fixed effects, is driven by differences in data precision rather than by static concentration. This pattern is consistent with the view that richer and more precise data reduce search and matching frictions, intensify price competition, and compress markups (Bergemann and Bonatti, 2024), in line with evidence that lower search costs in online markets strengthen price competition and reduce average prices even as some dispersion may persist (Baye et al., 2006).

Third, column (4) shows that alongside lower prices, lagged DPI is positively related to sales per product, significant at the 1% level, with an estimated coefficient of 27.25. A one–standard–deviation increase in DPI thus raises sales per product by 7.63 units, which amounts to about 9.5% of the mean sales per product (80.54). HHI enters with a much smaller coefficient of 0.39 and is statistically insignificant, implying that variation in sales is overwhelmingly associated with the informational environment captured by DPI. Together with lower prices, this pattern suggests that improvements in data precision increase visibility and relevance in search and recommendation, raise conversion, and reallocate demand toward better-matched products (Chevalier and Mayzlin, 2006).

Finally, column (5) reveals the ultimate impact on seller profitability. The average margin per product decreases with DPI, significant at the 1% level, with an estimated coefficient of $-3.35$. A one–standard–deviation increase in DPI lowers the average margin by $-0.94$ units, corresponding to a reduction of roughly 10.5% relative to its mean (8.919). In contrast, the coefficient on HHI is small (0.04) and statistically insignificant. This negative margin response captures the fundamental trade-off of platform-driven demand: while sellers experience higher sales volume, the intensified competition and the platform's ability to extract rents (e.g., via fees and targeted steering) place tighter constraints on extracting margins from consumers (Bergemann and Bonatti, 2024). The fact that DPI, but not HHI, strongly predicts this margin compression highlights that informational power rather than static concentration drives this effect.

Overall, a one–month improvement in category–level data precision is highly associated with broader product assortments and faster net entry, which subsequently lead to lower prices, higher unit sales, but compressed margins at the store level. Importantly, these effects remain large and

highly significant even after controlling for category–level HHI, while the coefficients on HHI are generally small and often economically negligible. Appendix Table IA.2 shows that when DPI is removed and lagged HHI is used as the sole market–structure variable, HHI is statistically indistinguishable from zero for prices, margins, product scope, and net entry, and even for sales per product its effect is economically tiny compared with the DPI estimates in Table 4. Appendix Table IA.3 further shows that higher DPI actually predicts more competitive market structures, in that an increase in lagged DPI is followed by lower HHI and higher entropy of seller sales in the next month.

What matters is not only how fragmented the seller side is, but how effectively the platform can use accumulated data to steer demand. DPI captures this information–based source of market power, and the evidence indicates that such data power has first–order implications for product variety, prices, and quantities that are largely orthogonal to what can be inferred from conventional product–market measures such as HHI or entropy alone.

*[Insert Table 4 Here]*

## 3.5 How Does Firm in Amazon React to Data Power Index (DPI)?

I have shown above that category–month DPI is tightly linked to marketplace dynamics on Amazon, such that higher DPI is associated with lower prices, higher quantities, margin compression, and richer product assortments, while standard market power measure (i.e., HHI) does not. A natural next question is whether firms that sell on Amazon adjust their corporate policies when the platform's data environment strengthens. This question matters because it bridges two strands of literature. First, the traditional product market literature predicts that stronger competition induces cost discipline and operational efficiency, as firms respond to tighter markups and greater demand uncertainty (Syverson, 2011). It also highlights a competition–innovation link in which more intense rivalry can shift firms toward expanding their innovation output and pivoting toward younger knowledge bases for faster renewal (Aghion et al., 2005). Second, the digital–platform literature emphasizes that data accumulation improves targeting and matching, reallocates surplus toward consumers, and thereby puts sustained pressure on seller markups (Bergemann and Bonatti, 2024). When search and discovery costs fall, product variety expands and experimentation becomes more

attractive (Brynjolfsson et al., 2011).

In this paper, the key question is how these familiar product market forces manifest in a digital platform environment, and whether they are primarily captured by traditional concentration measures such as HHI or by a platform–specific data power measure such as DPI. Accordingly, these perspectives imply that a rise in DPI should translate into observable adjustments in firm revenues, operating costs, and both the scale and age structure of their innovation portfolios. Comparing responses to DPI and HHI can reveal that digital platforms are disciplined by competition much like traditional markets, but through a data–driven channel that standard market power measures fail to capture.

To test these predictions, I estimate an analog of equation (10) at the firm–quarter frequency, replacing store and year–month fixed effects with firm and year–quarter fixed effects. Accordingly, the independent variable is the category–level DPI, lagged one quarter to mitigate simultaneity and to capture low–frequency corporate responses. Tables 5 and 6 report results for (i) firm sales and operating outcomes (revenue, sales, cost of goods sold, operating expense, and SG&A expense), and (ii) innovation outcomes (the cumulative number of active patents, and the average age of the patent portfolio based on active holdings and application dates). Except for the patent age variables (which are measured in years), all other dependent variables are normalized by total assets.

### 3.5.1 Sales and Operating Efficiency

Table 5 summarizes how sales, revenues, and operating efficiency respond to variation in DPI. In column (1), I find that lagged DPI is negatively associated with firm revenue, statistically significant at the 1% level. Economically, a one–standard–deviation increase in DPI (0.28) is associated with a decline in normalized revenue of about 4.4% relative to its mean (0.219). Column (2) shows a nearly identical effect for sales, with quarterly sales falling by roughly 4.4% relative to their mean. This decline in top-line metrics is consistent with the product-market evidence in Table 4, where intensified platform competition drives down prices and compresses per-product margins.

To absorb these tighter margins, columns (3) through (5) show that firms aggressively adjust their cost structures. Specifically, I document large and statistically significant declines in the cost of goods sold (column 3), operating expense (column 4), and SG&A expense (column 5). Accordingly, these results suggest that firms respond to the platform's strengthening data environment by

cutting variable and overhead costs to protect profitability.

Importantly, these patterns are specific to DPI. Once I control for the standard product–market concentration measure, HHI, the DPI coefficients remain large and statistically significant, whereas the HHI coefficients are economically minuscule and statistically insignificant across all specifications. In other words, traditional concentration does not explain firms' sales and operating adjustments, but the platform's data power does. Overall, the evidence is consistent with theoretical models and empirical work showing that stronger product market competition compresses markups and prompts managers to cut costs and improve productivity (Syverson, 2011; Bergemann and Bonatti, 2024), while these forces operate through a digital platform data channel captured by DPI rather than by HHI.

*[Insert Table 5 Here]*

### 3.5.2 Innovation Outcomes

Table 6 relates DPI to firms' innovation outcomes and patent portfolio age profiles. In the innovation literature, the age distribution of a firm's patent portfolio is an empirical proxy for the recency and renewal of its knowledge base, capturing shifts toward exploration versus exploitation and the obsolescence of incumbent technologies (Hall et al., 2005; Aghion et al., 2005; Ma, 2025). In column (1), I first find a positive association between lagged DPI and the cumulative number of active patents, statistically significant at the 10% level, suggesting a general expansion in innovation output.

Next, columns (2) and (3) reveal a structural shift in the age of these innovation portfolios. In column (2), I document a negative association between lagged DPI and the average age of active patents, statistically significant at the 1% level. This indicates that a one–standard–deviation increase in DPI (0.28) implies an age reduction of about 0.60 years, or roughly 4.7% of its mean (12.69). In column (3), using portfolio age measured by application date, the coefficient is again negative and statistically significant at the 5% level. The implied reduction is about 0.40 years, or around 2.8% of its mean (13.86). The coefficients on HHI in all columns are small and statistically indistinguishable from zero, suggesting that shifts in the scale and age profile of innovation portfolios are not explained by changes in standard concentration measures once DPI is taken into

account.

Therefore, the results point to a rotation toward younger intellectual property and a quicker renewal cycle. In practical terms, a richer data environment reduces search frictions and raises the expected payoff to launching new varieties and learning from early demand signals. To capitalize on this, firms refresh their portfolios rather than relying on older assets. Given the product market evidence of expanding variety, the younger age profile suggests that firms not only list more products but also feed that expansion with newer technologies, consistent with theories that stronger competition and lower search costs increase the return to exploration and experimentation (Aghion et al., 2005; Brynjolfsson et al., 2011).

*[Insert Table 6 Here]*

In sum, higher DPI does not merely reorganize transactions on the marketplace, but it is associated with economically meaningful firm–level adjustments. Firms adopt leaner cost structures by aggressively reducing operating and SG&A expenses, and tilt their patent portfolios toward younger assets. These relationships remain strong even after controlling for traditional product–market concentration, and the corresponding HHI coefficients are small and statistically insignificant. Thus, the adjustments documented in these results are not driven by standard concentration but by the platform's data power. These responses align with the dual predictions of the competition literature and the data–rich platform literature, highlighting that DPI captures a structural feature of the platform that shapes corporate financial and innovation choices in ways that HHI does not.

# 4 Does Data Power Index (PDI) Predict Amazon's Brand Entry?

Over the last two decades, digital platforms have evolved from intermediaries into direct market participants. In many online marketplaces, the platform launches its own private label products and competes side by side with third-party sellers. Amazon, for example, sells Amazon Basics, Pinzon, and Solimo alongside millions of third-party listings, and Alibaba's Tmall operates proprietary lines such as Tmall Supermarket. This platform-as-seller model raises central questions

for competition, market structure, and seller incentives, because the platform can influence search rankings, advertising placement, and access to consumer data. Amazon provides a data-rich setting for these questions. Since the mid-2000s, the company has introduced dozens of private labels across categories such as Electronics, Health & Household, and Grocery. These staggered entries create useful variation for studying competitive effects, and recent antitrust actions, including the 2023 U.S. Federal Trade Commission (FTC) lawsuit against Amazon, have further elevated the policy relevance of this topic. [8]

If platform competitiveness is driven less by traditional inputs and economies of scale and more by the breadth and precision with which data can be collected and used (Martens, 2021; Bergemann and Bonatti, 2024), then price, margins, and concentration measures such as HHI may be insufficient statistics for platform markets, where informational asymmetries and access to consumers are pivotal (Eisenmann et al., 2011; Demary and Rusche, 2018). I therefore ask whether Amazon's private label entry is explained by traditional category-level variables, and if not, whether a data-centric measure such as the Data Power Index (DPI) adds predictive power. Evidence that DPI forecasts entry, conditional on standard controls, would suggest that DPI captures a platform-specific mechanism that traditional product-market variables do not.

The theoretical framework in Kirpalani and Philippon (2020) is instructive. They formalize how more precise information increases the platform's outside option, that is, a copycat threat, and reduces the seller's outside option through a consumer access externality. In their model, an information precision parameter directly shifts the bargaining power balance between the platform and sellers. Interpreted in our setting, Amazon has incentives to select categories where data power is high, then to combine that informational advantage with platform infrastructure, including search, ads, logistics, and reviews. A natural empirical prediction is that categories with higher DPI should exhibit higher probabilities of Amazon brand entry.

I begin with Appendix Table IA.1, which summarizes the timing and scope of major Amazon's private label entry by category. For each treated category, I report the month and year of the first entry and the main brands launched. I construct these entries by cross-validating multiple public sources. The variation in both timing and categories provides the basis for event-time analysis and

[8] https://www.ftc.gov/news-events/news/press-releases/2023/09/ftc-sues-amazon-illegally-maintaining-monopoly-power

for cross-sectional prediction.

As a preliminary evidence, Figure 3 plots the mean number of products for Amazon-affiliated and non-Amazon brands from 2000 to 2022 at the category–year level. The Amazon series increases notably after 2015, while the non-Amazon series changes only modestly. This pattern indicates that product variety expands primarily within Amazon's own brands. The acceleration after 2015 is consistent with the maturation of internal data and infrastructure, including review accumulation, ad targeting, and fulfillment efficiencies.

*[Insert Figure 3 Here]*

Figure 4 reports an event-time difference plot of DPI, defined as the mean difference between treated categories which experiences Amazon brand entry, and control categories within a $\pm 12$ month window around the entry month, $t = 0$. Two features are informative. First, DPI in treated categories rises after entry and remains elevated, which is consistent with stronger data precision and informational capacity at, or immediately following, entry. Second, more importantly, the mean difference is not zero immediately before entry, and treated categories already exhibit systematically higher DPI than control categories. This pre-entry level gap implies nonrandom selection, that is, Amazon appears to prospect and prioritize categories with higher data power prior to entry. In other words, DPI functions both as a leading signal for entry and as a channel through which post-entry differences persist.

*[Insert Figure 4 Here]*

These patterns have direct implications for estimation and interpretation. Because selection is likely, I move beyond raw means and estimate probit models in which entry is the dependent variable, traditional category-level variables, including price, margins, and HHI, are included as controls, and DPI is added as a data-centric predictor. If DPI remains a significant predictor after these adjustments, this would provide evidence that data power plays a structural, leading role in platform entry decisions. From a policy perspective, quantifying the link between data access, algorithmic allocation, and market entry can inform antitrust and platform-governance design. For third-party sellers, the results speak to strategic choices about category focus, product launch timing, and investments in advertising and reviews. For consumers, broader private label variety

may lower prices and improve match quality, yet data-driven entry can also raise concerns about longer-run concentration and innovation.

## 4.1 Predictive Power of DPI versus Conventional Measures

This section examines whether Amazon's first private label entry at the category–month level is explained by traditional product–market variables, and whether a data–centric measure, the Data Power Index (DPI), adds incremental predictive power. I estimate the following probit specification at the category–month level:

$$\Pr\big(\text{Entry}_{c,t}=1\big)=\Phi\Big(\alpha\,\text{DPI}_{c,t-1}+\mathbf{X}_{c,t-1}^{\top}\boldsymbol{\theta}+\delta_c+\gamma_y+\mu_m\Big), \tag{11}$$

where $\text{Entry}_{c,t}$ indicates whether category $c$ experiences the first Amazon brand entry in month $t$, $\text{DPI}_{c,t-1}$ is the one–month lag of DPI, and $\mathbf{X}_{c,t-1}$ contains one–month lags of conventional product market predictors and Amazon product quality measures, such as average product price, sales per product, margin rate, Herfindahl–Hirschman Index (HHI), and consumer metrics, such as the average number of reviews, the average rating, and the number of products. I include category fixed effects $\delta_c$, year fixed effects $\gamma_y$, and month fixed effects $\mu_m$. Standard errors are clustered at the category–month level. All regressors are lagged by one month in order to mitigate simultaneity.

Table 7 reports probit estimates of the probability of first Amazon private label entry. Column (1) includes only the lagged DPI with fixed effects. The DPI coefficient is 1.89, statistically significant at the 1% level. This indicates that even without additional controls, categories with higher data power are more likely to see first–time entry in the subsequent month. Then, column (2) adds conventional product market factors, namely average price, sales per product, margin rate, and category HHI. The coefficient of DPI increases to 2.34 and is statistically significant at the 1% level. While sales per product shows a positive association, other traditional controls such as average price, margin rate, and HHI are small and statistically insignificant. This shows that conventional proxies for market power and entry incentives largely fail to explain Amazon's timing and category choice, whereas DPI does.

Column (3) replaces traditional controls with the Amazon product quality variables used to construct DPI, such as the average number of reviews, the average rating, and the number of prod-

ucts at the category level. DPI remains significant with a coefficient of 1.95, statistically significant at the 5% level. The number of products loads negatively with a coefficient of $-0.65$ and is statistically significant at the 1% level, implying that already crowded categories are less attractive near-term entry targets due to cannibalization risk and the difficulty of overtaking entrenched variety. Column (4) pools all controls and constitutes the full specification. DPI remains a positive predictor with a coefficient of 2.28, statistically significant at the 10% level. While a few controls like average rating and sales per product show significance in this pooled model, the negative association with the number of products remains robust. These results reinforce that Amazon prioritizes categories where data power is high and congestion is not excessive.

On the other hand, column (5) includes only conventional product-market predictors without DPI. The model has limited explanatory power, and most standard controls are insignificant, indicating that prices, margins, and conventional market-concentration proxies do not capture the platform's selection mechanism well. Similarly, column (6) includes only the Amazon product-quality variables used to construct DPI. Average reviews and ratings are weak predictors on their own, while the number of products remains negatively significant with a coefficient of $-0.72$, statistically significant at the 1% level. Comparing columns (3) and (6) with the DPI specifications shows that single Amazon product-quality measures are not the determinative factors, whereas the structured integration of such signals into DPI, which captures informational precision and match quality, has substantial predictive content.

Across all specifications, conventional variables that usually proxy market power and entry incentives do not forecast Amazon's brand entry, while DPI does. Economically, Amazon appears to enter where demand can be read more precisely from data and where platform tools, search, recommendation, advertising, and logistics, can amplify that informational advantage, while avoiding immediately crowded spaces where cannibalization risks are high. The estimates are consistent with the graphical evidence. Figure 3 shows that Amazon–affiliated brands expand product variety notably after 2015, while non–Amazon brands are comparatively flat over the same period. In Figure 4, DPI in treated categories rises after entry and remains elevated, and importantly, the pre–entry difference is not zero. Treated categories already exhibit higher DPI immediately before entry, which implies nonrandom selection based on data power. In other words, DPI functions as a leading signal for entry, and it is also a channel through which post–entry differences persist.

*[Insert Table 7 Here]*

Accordingly, the economic interpretation is straightforward. Traditional variables that usually proxy market power and entry incentives do not forecast Amazon's brand entry, while DPI does. Amazon appears to enter first where demand can be read more precisely from data, and where search, recommendation, advertising, and logistics can amplify that informational advantage. These findings align with the theoretical insight in Kirpalani and Philippon (2020), in which greater information precision strengthens the platform's outside option, that is, a copycat threat, and weakens the seller's outside option through a consumer access externality, thereby shifting bargaining power.

# 5 Impact on Sellers, Consumers, and Firms after Amazon Brand Entry

In the preceding sections, I have shown that the Data Power Index (DPI) strongly predicts the likelihood of Amazon's private brand entry across product categories. This finding implies that Amazon strategically leverages its data advantage to enter markets where its informational precision provides a competitive edge. Such data-driven entry is not merely a platform expansion decision but a structural intervention that reshapes the equilibrium among multiple marketplace participants.

In this section, I evaluate the multifaceted consequences of this vertical integration across three main dimensions. First, I examine the product market dynamics, focusing on how incumbent third-party sellers respond to intensified competition, potential demand diversion, and margin compression. Second, I assess the impact on consumer welfare, specifically looking at changes in product variety, perceived quality, and the resolution of informational frictions. Third, I analyze the broader corporate implications for firms operating on Amazon, investigating how they adjust their innovation portfolios and how their underlying firm risk is affected by the platform's entry.

To systematically quantify these effects across different participants, I employ a staggered difference-in-differences (DID) framework. The baseline estimation model is specified as follows:

$$Y_{i,c,t} = \alpha + \beta \text{Treat}_{c,t} + X'_{i,c,t}\theta + \gamma_i + \delta_t + \varepsilon_{i,c,t}, \tag{12}$$

where $i$ indexes the unit of observation (stores or firms), $c$ denotes Amazon product categories, and $t$ refers to the month (or quarter, depending on the data frequency). The treatment variable, $\text{Treat}_{c,t}$, is an indicator that equals one for all periods following Amazon's initial private-label entry into category $c$. The coefficient of interest, $\beta$, captures the average treatment effect on the treated (ATT). The vector $X_{i,c,t}$ includes time-varying controls, while $\gamma_i$ and $\delta_t$ represent individual and time fixed effects, respectively, to absorb time-invariant unobserved heterogeneity and common macroeconomic shocks.

By integrating market-level and firm-level evidence through this empirical strategy, this section aims to reveal how Amazon's data-powered entry strategy alters prices, welfare, innovation, and risk, providing a comprehensive picture of the economic implications of data-driven vertical integration.

## 5.1 Product Market Responses to Amazon Brand Entry

As discussed, platform-owned brand entry fundamentally alters competitive conditions within the product market. Table 8 documents how Amazon's private–brand entry reshapes these product–market outcomes for incumbent third-party sellers at the store–category–month level.

First, I examine the extensive margin of product scope in columns (1) and (2). A key expectation is that product variety may increase temporarily as sellers defensively list new products to maintain visibility against the platform's proprietary label. In column (1), the estimated coefficient on $\text{Treat}_{c,t}$ is 0.04, statistically significant at the 5% level. Given the log-link formulation of the Poisson model, this corresponds to an increase in the expected number of products of approximately 4.1%.[9] Column (2) further shows that rolling three-month net new products increase by 0.10 and significant at the 10% level, which is economically large relative to its baseline mean of 0.0965. This statistically precise expansion is consistent with a short–run defensive listing response, in which sellers introduce product variants to protect their shelf space immediately fol-

[9] In a Poisson regression, the percentage change in the expected count for a dummy variable is calculated as $100 \times (\exp(\beta) - 1)$. Thus, $100 \times (\exp(0.04) - 1) \approx 4.08\%$. Evaluated at the baseline mean of 4.186 products, this translates to roughly 0.17 additional products per store–category.

lowing Amazon's entry.

Second, I evaluate the pricing response. When a platform vertically integrates, it gains the ability to utilize preferential placement, sponsored slots, and algorithmic self-preferencing to intensify price competition (Hagiu et al., 2022; Zennyo, 2022; Long and Amaldoss, 2024). In column (3), the estimated coefficient is $-0.06$ and statistically significant at the 1% level. Interpreted in a log–linear way, this corresponds to a price decline of about -5.8%.[10]. This implies that after Amazon's private brand enters a category, third-party sellers are forced to cut their average product prices by approximately 6% to remain competitive.

Third, columns (4) and (5) examine sales and margins per product, which capture the ultimate impact on seller profitability. The literature documents that algorithmic self-preferencing and consumer trust in the platform's brand inevitably lead to demand diversion from independent sellers to the platform (Zhu and Liu, 2018; Crawford et al., 2022). In column (4), the coefficient on the treated indicator is $-9.57$, significant at the 1% level. Relative to its mean of 80.12 units, this represents a substantial 11.9% decline in sales per product. In column (5), the coefficient is $-1.54$, also significant at the 1% level, implying that the average margin per product falls by about 17.4% of its mean value of 8.867. Along with the price decline in column (3), these results indicate that Amazon's entry compresses product–level profitability and reallocates demand away from incumbents in a data–intensive, information–advantaged competitive environment.

Overall, Table 8 shows that Amazon's private–brand entry is followed by a modest short-run increase in the number of listed products, but ultimately leads to lower prices, weaker sales, and compressed margins per product. Prices fall by about 6%, while sales per product and margins decline by roughly 12% and 17%, respectively. This pattern is consistent with Amazon using data–driven placement to redirect demand toward its own label, compressing third–party unit economics and triggering defensive assortments as sellers try to survive. Strategically, these results suggest that third–party sellers cannot rely solely on cutting prices. In a setting where the platform controls search and recommendation, investing in information capital—such as verified reviews, stable ratings, and search–friendly quality signals—becomes central for maintaining visibility and volume.

*[Insert Table 8 Here]*

[10] $100 \times (\exp(-0.06) - 1) \approx -5.8\%$

Building on these baseline static results, I further assess the validity of the identifying assumption and explore the dynamic treatment effects using the Callaway and Sant'Anna (2021) estimator for staggered adoption settings. Figure 5 reports the event–study estimates of the treatment effect on the logarithm of average product price. The pre–treatment coefficients fluctuate tightly around zero, providing reassuring evidence of parallel pre–trends. In contrast, the post–treatment periods display a clear and persistent decline in prices, consistent with the static DID estimates reported in Table 8.

*[Insert Figure 5 Here]*

Similarly, Figure 6 plots the dynamic effects of Amazon's private-brand entry on sales per product. Consistent with the demand diversion hypothesis, the pre-treatment estimates exhibit no significant anticipatory trends, whereas the post-entry periods reveal a sharp and sustained drop in incumbent sales volume. Together, these dynamic patterns strengthen the interpretation that Amazon's data-driven entry causes a significant and lasting reduction in both product prices and third-party sales, rather than merely reflecting pre–existing downward trajectories in the targeted categories.

*[Insert Figure 6 Here]*

## 5.2 Consumer Surplus: Reallocation from Sellers to Consumers

This paper next studies how Amazon's private–brand entry affects consumer welfare on the platform. Pre-purchase information, such as the number and credibility of consumer reviews, plays a central role in shaping welfare gains from digital markets. Reimers and Waldfogel (2021) show that ratings and reviews in the digital environment bridge the gap between consumers' ex ante expectations and ex post utility, thereby increasing consumer surplus by reducing information frictions. Following their approach, I use product level information, specifically star ratings to construct a log–sum measure of consumer surplus, $\Delta CS$, and express it in dollar terms.

I implement this welfare measure at two levels of aggregation: (i) $\Delta CS(\text{Store, stars, \$})$, which computes the contribution to surplus at the individual third-party store level, and (ii) $\Delta CS(\text{Market, stars, M\$})$, which computes the broader market–level (category) surplus that accrues to consumers in millions

of dollars. To isolate the informational value of ratings, both specifications include the logarithm of average product price and other informational variables as time-varying controls.[11]

Figure 7 illustrates the cross–category heterogeneity in $\Delta CS$ during 2018 to benchmark the implementation. Categories are ordered by average monthly surplus. Large, review-sensitive categories such as *Home & Kitchen*, *Clothing, Shoes & Jewelry*, and *Home Improvement* sit at the top of the distribution, generating surplus gains on the order of $1.0 to $1.5 million per month. In contrast, long–tail categories contribute much smaller amounts. The dashed horizontal line marks a reference level of $265,000, underscoring the heavy-tailed nature of consumer welfare generation across the platform's diverse ecosystems.

*[Insert Figure 7 Here]*

Having established that the $\Delta CS$ measures deliver plausible magnitudes, I turn to how Amazon's private brand entry affects consumer welfare. Table 9 presents the staggered DID specifications linking Amazon's entry to monthly changes in consumer surplus. Column (1) focuses on the store–level contribution to consumer surplus generated by third–party sellers. The coefficient on the treated–category indicator is $-103.38$, statistically significant at the 1% level. Relative to the baseline mean of $266.20, this represents a substantial decline of about 38.8% in the store–level surplus contribution. Thus, after Amazon's private brand enters a category, individual third–party stores generate significantly less incremental consumer surplus per month. This pattern is consistent with the concern that algorithmic favoring of the platform's private brands redirects review visibility toward its own products, reducing consumer exposure to high-quality third-party alternatives and eroding the matching gains that independent sellers previously captured (Hagiu et al., 2022; Zennyo, 2022; Shi, 2025).

In contrast, column (2) turns to the broader market–level (category–level) surplus, with the dependent variable expressed in millions of dollars. The coefficient of the main interest is 0.22, statistically significant at the 5% level. Given a mean of $0.186 million, this implies that the total market level consumer surplus rises by about $220,000 per category–month, an increase of roughly 118% relative to the mean. This positive market-wide effect suggests that the combination of intensified price competition (as documented in Section 5.1) and the platform's efficient matching

[11] This paper provides detail about how I calculate the consumer welfare measure following Reimers and Waldfogel (2021) in Appendix B.

mechanisms ultimately expands aggregate consumer welfare, consistent with standard results that improved choice sets raise consumer surplus (Petrin, 2002; Brynjolfsson et al., 2003).

Accordingly, Table 9 highlights a sharp incidence trade–off. Amazon's data-driven private–brand entry substantially reduces the consumer–surplus contribution of individual third–party sellers at the store level by roughly $103 per store–month, yet it simultaneously increases aggregate category level consumer surplus by about $220,000 per month. In other words, total welfare rises, but the underlying dynamics reveal a structural reallocation: welfare shifts away from incumbent sellers and toward consumers as the vertically integrated platform uses its informational advantage and placement power to redirect traffic and sharpen competition (Reimers and Waldfogel, 2021; Zhu and Liu, 2018).

*[Insert Table 9 Here]*

## 5.3 How does Firm in Amazon react to Amazon Brand Entry?

### 5.3.1 Does Market React Negatively to Competition with Amazon Brand?

To examine how Amazon's private brand entry affects the market valuation of incumbent firms, I conduct an event study centered on each Amazon brand's initial entry date. Following the standard event study framework (Fama et al., 1969; MacKinlay, 1997), I estimate firm-specific abnormal returns using the Fama–French four-factor model, which adjusts for market, size, value, and momentum factors. Specifically, I estimate expected returns over a 100-day estimation window ending 30 trading days prior to the event date $T^*$ as follows:

$$R_{i,t} = \alpha_i + \beta_{m,i} R_{m,t} + \beta_{s,i} SMB_t + \beta_{h,i} HML_t + \beta_{u,i} UMD_t + \varepsilon_{i,t}, \tag{13}$$

where $R_{i,t}$ and $R_{m,t}$ denote the daily excess returns of firm $i$ and the market, respectively. Abnormal returns are then computed as:

$$AR_{i,t} = R_{i,t} - (\hat{\alpha}_i + \hat{\beta}_{m,i} R_{m,t} + \hat{\beta}_{s,i} SMB_t + \hat{\beta}_{h,i} HML_t + \hat{\beta}_{u,i} UMD_t). \tag{14}$$

I aggregate abnormal returns into cumulative abnormal returns (CAR) over various event win-

dows surrounding Amazon's entry date:

$$CAR_i(a,b) = \sum_{t=a}^{b} AR_{i,t}, \tag{15}$$

where $[a,b]$ represents event windows such as [0,0], [–1,1], [–3,3], and [–5,5]. To capture both the immediate and persistent market reactions, I further compute CARs over an extended window from –30 to +180 trading days relative to the event date.

Figure 8 presents the mean CAR across 289 Amazon brand entry events. The solid purple line represents the average CAR, while the dotted lines indicate the 95 percent confidence interval. The figure shows a pronounced and sustained decline in cumulative abnormal returns following Amazon's entry, reaching approximately –15% to –20% within six months. This persistent downward trend suggests that investors interpret Amazon's private brand launches as a negative competitive signal, anticipating intensified competition, data-driven advantages, and margin pressure for incumbent firms.

These findings are consistent with prior research documenting that platform owners' entry into complementary or competing markets tends to reduce the market value of partner or rival firms. Cennamo et al. (2018) show that platform entry into complementor domains negatively affects partners' valuations, while Wen and Zhu (2019) provide evidence that platform self-entry crowds out complementors by leveraging superior data access. My results extend this literature by providing direct stock market evidence that Amazon's private brand entry triggers negative abnormal returns among same-category firms, implying that investors anticipate redistribution of market value toward the platform.

*[Insert Figure 8 Here]*

### 5.3.2 Does Firm Face Higher Risks When Amazon Brand Enter?

Next, Table 10 examines how Amazon's private–brand entry relates to firm–level risk. The dependent variables are: (i) the rolling volatility of operating income, (ii) idiosyncratic return volatility, (iii) total return volatility, and (iv) market beta $\beta^{MKT}$. Following Irvine and Pontiff (2009), I interpret these measures as linking product–market competition to firms' cash–flow and return

risks. The operating–income volatility measure follows the spirit of Makaew and Maksimovic (2020), who use operating variability to capture real–side instability under intensified competition.

In column (1), the coefficient on operating–income volatility is 0.062, statistically significant at the 5% level. Given a mean of 0.317, this represents an increase of about 20% in the volatility of operating income, indicating that the stock–market effects are mirrored by instability in underlying operations. Columns (2) and (3) focus on monthly return risks. In column (2), the coefficient on the treatment indicator for idiosyncratic volatility is 0.008, statistically significant at the 5% level. Relative to the mean of 0.0788, this corresponds to an increase of about 10% in idiosyncratic volatility. In column (3), the coefficient for total volatility is 0.012, significant at the 1% level, which is roughly 12% of its mean of 0.0986. If these volatilities are interpreted as monthly standard deviations, the absolute changes map into annualized volatility increases of about 2.8 and 4.2 percentage points, respectively. Column (4) shows that market beta $\beta^{MKT}$ also rises: the treatment coefficient is 0.087, significant at the 5% level, which is about 9% of the its mean. It implies that firms in the treated categories become more sensitive to aggregate shocks after Amazon's entry, compared to other firms in the control categories.

Therefore, these estimates are consistent with a competition–to–volatility transmission: stronger product market competition increases cash–flow instability, which in turn raises firm–level return volatility, as in Irvine and Pontiff (2009). The operating–income results echo the mechanism emphasized by Makaew and Maksimovic (2020) that intensified competition makes real operations more volatile. The rise in $\beta^{MKT}$ dovetails with Baker et al. (2023), who show that erosion of customer bases and intangible capital heightens firms' exposure to systematic risk. It implies that Amazon's private–brand entry plausibly weakens incumbents' brand capital and customer stickiness, amplifying both idiosyncratic and systematic components of risk. Overall, Amazon's entry not only elevates stock–return volatility but also increases operating volatility and market sensitivity, pointing to a broad reshaping of the risk environment for incumbent firms in affected categories.

*[Insert Table 10 Here]*

### 5.3.3 Does Firm Adjust Innovation Strategy?

Lastly, platform entry can alter incumbent firms' innovation strategies, which is central for understanding the link between product–market competition and innovation. Prior work typically measures innovation using patent counts, citations, and quality or generality metrics (e.g., Kogan et al., 2017) and shows that competitive shocks shift the balance between exploration and exploitation (e.g., Aghion et al., 2005; Bloom et al., 2013). In this context, I use the average age of a firm's patent portfolio as an auxiliary outcome that summarizes whether the knowledge base tilts toward newer technological cohorts versus mature, exploitation–oriented domains (e.g., Ma, 2025), and I study how this measure responds when Amazon launches a private brand within a category.

In column (1) of Table 11, the dependent variable is the average age of active patents. The coefficient on the treatment indicator is $-0.646$, statistically significant at the 1% level. Relative to the mean age of 12.63 years, this corresponds to a decline of about 5.1%. In column (2), where portfolio age is measured by application date, the treatment coefficient is $-0.355$, significant at the 10% level, which is roughly 2.6% of its mean of 13.79 years. Economically, these magnitudes indicate that, following Amazon's entry, incumbents' patent portfolios become systematically younger.

Combined with the volatility evidence, the results are consistent with a competition–to–innovation mechanism in which heightened product market pressure increases cashflow uncertainty and pushes firms toward more recent technological cohorts (Irvine and Pontiff, 2009; Makaew and Maksimovic, 2020). The pattern also aligns with the idea that platform entry can erode brand capital and customer attachment, raising systematic risk exposure (Baker et al., 2023) and encouraging quicker portfolio renewal rather than reliance on older assets. In short, Amazon's private brand entry is associated with a measurable rotation of incumbents' patent portfolios toward younger technologies, consistent with theories that stronger competition increases the returns to exploration and experimentation (Aghion et al., 2005; Bloom et al., 2013; Ma, 2025).

*[Insert Table 11 Here]*

# 6 Conclusion

This paper studies how a platform's accumulated data functions as a fundamental mechanism of market power. To bridge the gap between economic theory and empirical observation, I introduce the Data Power Index (DPI), a novel measure that quantifies the precision with which a platform can read demand from its own marketplace signals. Using two decades of Amazon consumer interactions, I demonstrate that the DPI acts as a critical state variable in digital competition. Most notably, the DPI robustly predicts Amazon's strategic transitions from a marketplace operator to a direct competitor via private-label entry. This predictive power remains highly significant even after controlling for conventional product-market characteristics and traditional concentration metrics like the Herfindahl-Hirschman Index (HHI), which fail to consistently forecast the platform's behavior. This establishes that Amazon prioritizes markets not based on traditional structural dominance, but where its informational precision allows for tighter targeting and quicker learning.

Exploiting this data-driven entry, I document an important economic trade-off at the center of current regulatory and antitrust debates. On the demand side, Amazon's entry lowers search frictions, reduces prices, and significantly expands aggregate category-level consumer surplus. However, this allocative efficiency comes at a severe distributional cost. The platform systematically reallocates value away from the incumbent third-party sellers, who experience compressed per-product margins and a sharp contraction in their individual surplus contributions. Furthermore, this competitive shock extends well beyond the immediate marketplace. U.S. public firms exposed to entered categories face heightened financial uncertainty. To survive this intensified information-based competition, these firms are forced to adapt by implementing leaner operating cost structures and aggressively rotating their patent portfolios toward younger technologies.

By integrating product-market outcomes, welfare redistribution, and corporate adaptations into a unified framework, this paper makes three broad contributions. First, it provides a transparent and measurable state variable (DPI) that successfully maps unobservable data power to strategic firm behavior, replacing traditional structural metrics. Second, it delivers an empirical account of the complex trade-offs inherent in platform vertical integration, speaking directly to modern antitrust concerns, such as the FTC's recent lawsuit. Third, it bridges digital platform economics with corporate finance by showing how data-driven shocks travel from the platform ecosystem to

the corporate sector. Ultimately, this paper demonstrates that in the modern digital economy, competitive advantage and market power are no longer dictated merely by physical scale or structural concentration, but by the accumulation, precision, and weaponization of information.

# References


Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt, 2005, Competition and innovation: An inverted-u relationship, *The quarterly journal of economics* 120, 701–728.

Anderson, Simon, and Özlem Bedre Defolie, 2024, Hybrid platform model: Monopolistic competition and a dominant firm, *The RAND journal of economics* 55, 684–718.

Arcidiacono, Peter, Paul B Ellickson, Carl F Mela, and John D Singleton, 2020, The competitive effects of entry: Evidence from supercenter expansion, *American Economic Journal: Applied Economics* 12, 175–206.

Baker, Scott R, Brian Baugh, and Marco Sammon, 2023, Customer churn and intangible capital, *Journal of Political Economy Macroeconomics* 1, 447–505.

Basker, Emek, 2007, The causes and consequences of wal-mart's growth, *Journal of Economic Perspectives* 21, 177–198.

Baye, Michael R, John Morgan, Patrick Scholten, et al., 2006, Information, search, and price dispersion, *Handbook on economics and information systems* 1, 323–375.

Bennati, Luca, 2024, Vertical integration and e-commerce competition: Evidence from amazon marketplace .

Bergemann, Dirk, and Alessandro Bonatti, 2024, Data, competition, and digital platforms, *American Economic Review* 114, 2553–2595.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan, 2022, The economics of social data, *The RAND Journal of Economics* 53, 263–296.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen, 2013, Identifying technology spillovers and product market rivalry, *Econometrica* 81, 1347–1393.

Brynjolfsson, Erik, Yu Hu, and Duncan Simester, 2011, Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales, *Management science* 57, 1373–1386.

Brynjolfsson, Erik, Yu Hu, and Michael D Smith, 2003, Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers, *Management science* 49, 1580–1596.

Callaway, Brantly, and Pedro HC Sant'Anna, 2021, Difference-in-differences with multiple time periods, *Journal of econometrics* 225, 200–230.

Cennamo, Carmelo, Hakan Ozalp, and Tobias Kretschmer, 2018, Platform architecture and quality trade-offs of multihoming complements, *Information Systems Research* 29, 461–478.

Chevalier, Judith A, and Dina Mayzlin, 2006, The effect of word of mouth on sales: Online book reviews, *Journal of Marketing Research* 43, 345–354.

Clemons, Eric K, Guodong Gordon Gao, and Lorin M Hitt, 2006, When online reviews meet hyperdifferentiation: A study of the craft beer industry, *Journal of management information systems* 23, 149–171.

Crawford, Gregory S, Matteo Courthoud, Regina Seibel, and Simon Zuzek, 2022, Amazon entry on amazon marketplace, *CEPR Discussion Papers* .

Demary, Vera, and Christian Rusche, 2018, *The economics of platforms*, number 123 (IW-Analysen).

Dinerstein, Michael, Liran Einav, Jonathan Levin, and Neel Sundaresan, 2018, Consumer price search and platform design in internet commerce, *American Economic Review* 108, 1820–1859.

Eisenmann, Thomas, Geoffrey Parker, and Marshall Van Alstyne, 2011, Platform envelopment, *Strategic management journal* 32, 1270–1285.

Etro, Federico, 2021, Product selection in online marketplaces, *Journal of Economics & Management Strategy* 30, 614–637.

Fama, Eugene F, Lawrence Fisher, Michael C Jensen, and Richard Roll, 1969, The adjustment of stock prices to new information, *International economic review* 10, 1–21.

Gutierrez Gallardo, German, 2021, The welfare consequences of regulating amazon, *Available at SSRN 3965566* .

Hagiu, Andrei, Tat-How Teh, and Julian Wright, 2022, Should platforms be allowed to sell on their own marketplaces?, *The RAND Journal of Economics* 53, 297–327.

Hall, BH, A Jaffe, and M Trajtenberg, 2005, Market value and patent citations journal issue: Rand journal of economics, *The RAND Journal of Economics* 36.

Hausman, Jerry, and Ephraim Leibtag, 2007, Consumer benefits from increased competition in shopping outlets: Measuring the effect of wal-mart, *Journal of Applied Econometrics* 22, 1157–1177.

Hou, Yupeng, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley, 2024, Bridging language and items for retrieval and recommendation, *arXiv preprint arXiv:2403.03952* .

Irvine, Paul J, and Jeffrey Pontiff, 2009, Idiosyncratic return volatility, cash flows, and product market competition, *The Review of Financial Studies* 22, 1149–1177.

Kirpalani, Rishabh, and Thomas Philippon, 2020, Data sharing and market power with two-sided platforms, Technical report, National Bureau of Economic Research.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, 2017, Technological innovation, resource allocation, and growth, *The quarterly journal of economics* 132, 665–712.

Lee, Robin S, 2013, Vertical integration and exclusivity in platform and two-sided markets, *American Economic Review* 103, 2960–3000.

Long, Fei, and Wilfred Amaldoss, 2024, Self-preferencing in e-commerce marketplaces: The role of sponsored advertising and private labels, *Marketing Science* 43, 925–952.

Ma, Song, 2025, Technological obsolescence, *The Review of Financial Studies* hhaf059.

MacKinlay, A Craig, 1997, Event studies in economics and finance, *Journal of economic literature* 35, 13–39.

Makaew, Tanakorn, and Vojislav Maksimovic, 2020, Competition and operating volatilities around the world, *Journal of Financial and Quantitative Analysis* 55, 517–547.

Martens, Bertin, 2021, An economic perspective on data and platform market power, *Available at SSRN 3783297* .

OECD, 2022, The evolving concept of market power in the digital economy, Working Paper 278, OECD Competition Law and Policy Working Papers.

Parker, Geoffrey G, and Marshall W Van Alstyne, 2005, Two-sided network effects: A theory of information product design, *Management science* 51, 1494–1504.

Paulsen, Derek, Yash Govind, and AnHai Doan, 2023, Sparkly: A simple yet surprisingly strong tf/idf blocker for entity matching, *Proceedings of the VLDB Endowment* 16, 1507–1519.

Petrin, Amil, 2002, Quantifying the benefits of new products: The case of the minivan, *Journal of political Economy* 110, 705–729.

Reimers, Imke, and Joel Waldfogel, 2021, Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings, *American Economic Review* 111, 1944–1971.

Shi, Peng, 2025, Welfare-optimal policies for sponsored advertising in a two-sided marketplace, *Available at SSRN* .

Sun, Monic, 2012, How does the variance of product ratings matter?, *Management science* 58, 696–707.

Syverson, Chad, 2011, What determines productivity?, *Journal of Economic literature* 49, 326–365.

Wen, Wen, and Feng Zhu, 2019, Threat of platform-owner entry and complementor responses: Evidence from the mobile app market, *Strategic Management Journal* 40, 1336–1367.

Zennyo, Yusuke, 2022, Platform encroachment and own-content bias, *The Journal of Industrial Economics* 70, 684–710.

Zhu, Feng, and Qihong Liu, 2018, Competing with complementors: An empirical look at amazon. com, *Strategic management journal* 39, 2618–2642.

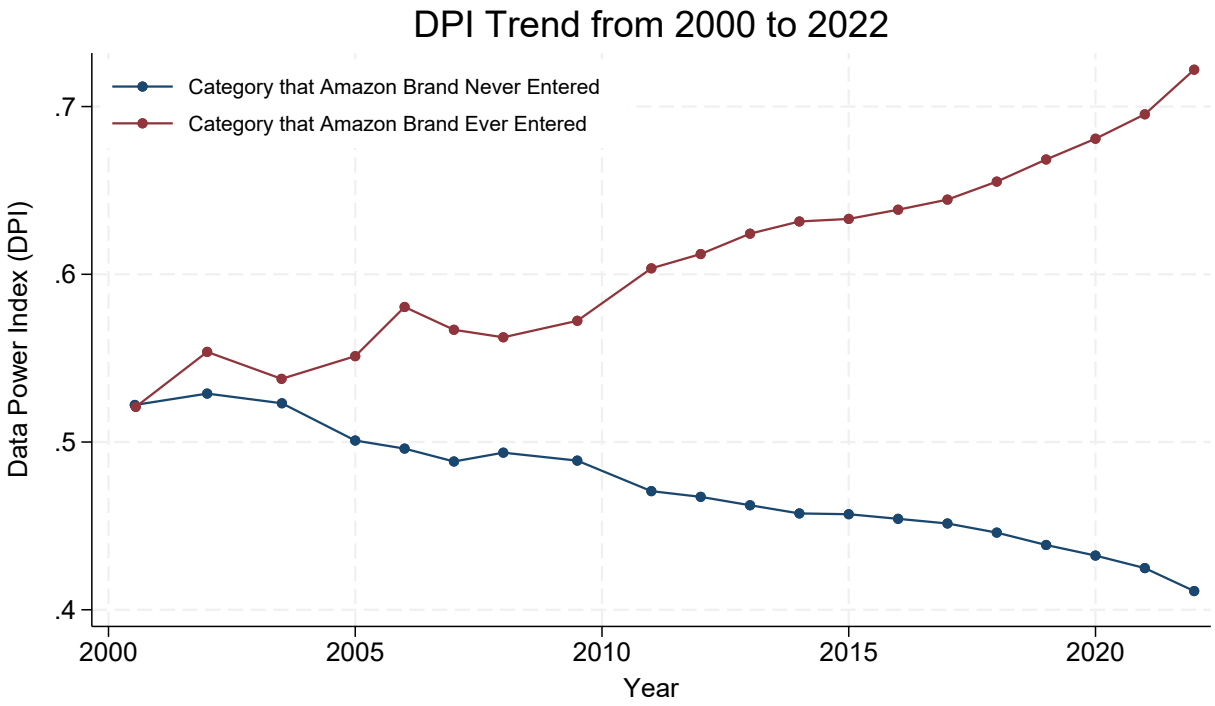



**Figure 1. DPI Trend from 2000 to 2022 by Amazon Entry Status**
This figure plots the time-series evolution of the Data Power Index (DPI) from 2000 to 2022, split by Amazon's private label entry. The red line represents the average DPI for categories that Amazon brands eventually entered ("Category that Amazon Brand Ever Entered"), while the blue line represents the average DPI for categories that remained untouched by Amazon brands ("Category that Amazon Brand Never Entered"). The graph demonstrates a clear divergence over time: categories targeted by Amazon exhibit a steady and significant increase in their informational richness, whereas non-targeted categories show a gradual decline in their DPI.

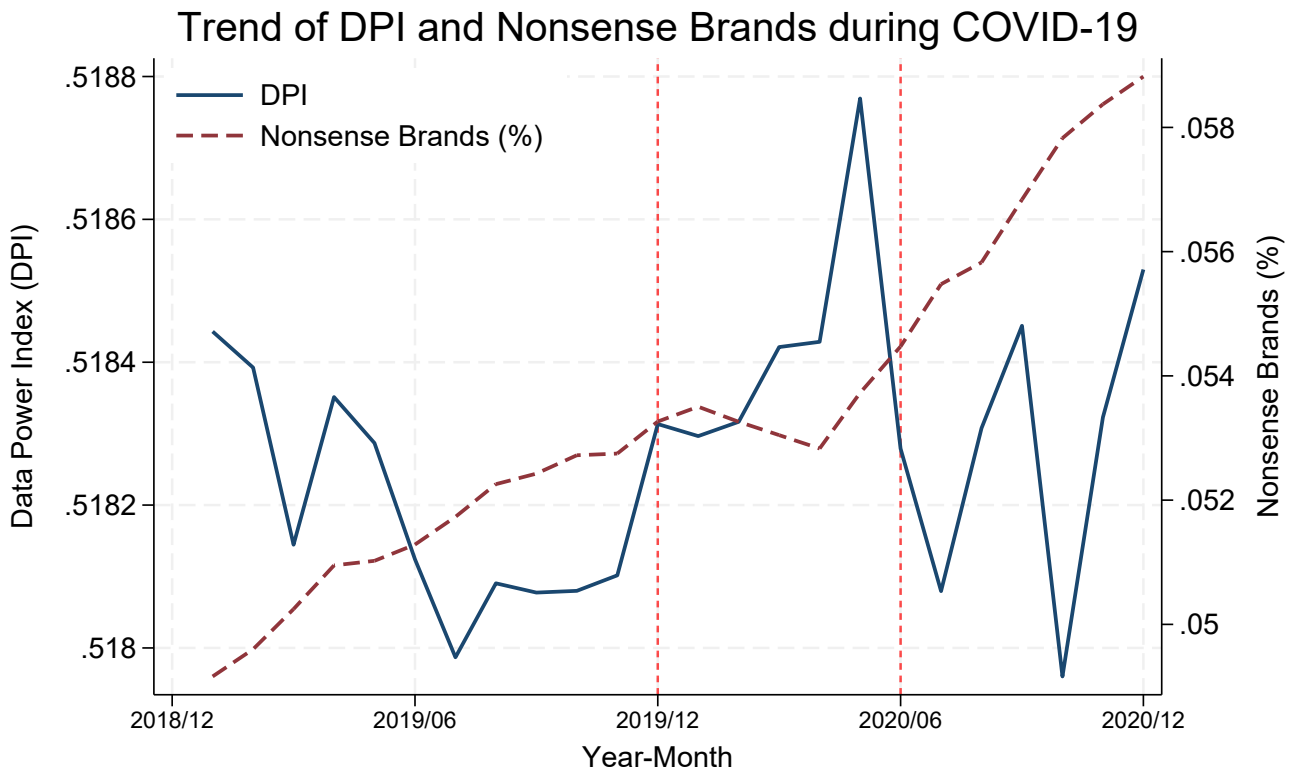



**Figure 2. Trend of DPI and Nonsense Brands during COVID-19**
This figure plots the monthly evolution of the overall Data Power Index (DPI) and the share of nonsense brands from December 2018 to December 2020. The solid blue line, plotted against the left axis, tracks the average DPI across all categories over this time period. The dashed red line, plotted against the right axis, represents the average percentage of nonsense brands within categories. The two vertical red dashed lines (December 2019 and June 2020) denote key time points associated with the onset and progression of the COVID-19 pandemic.

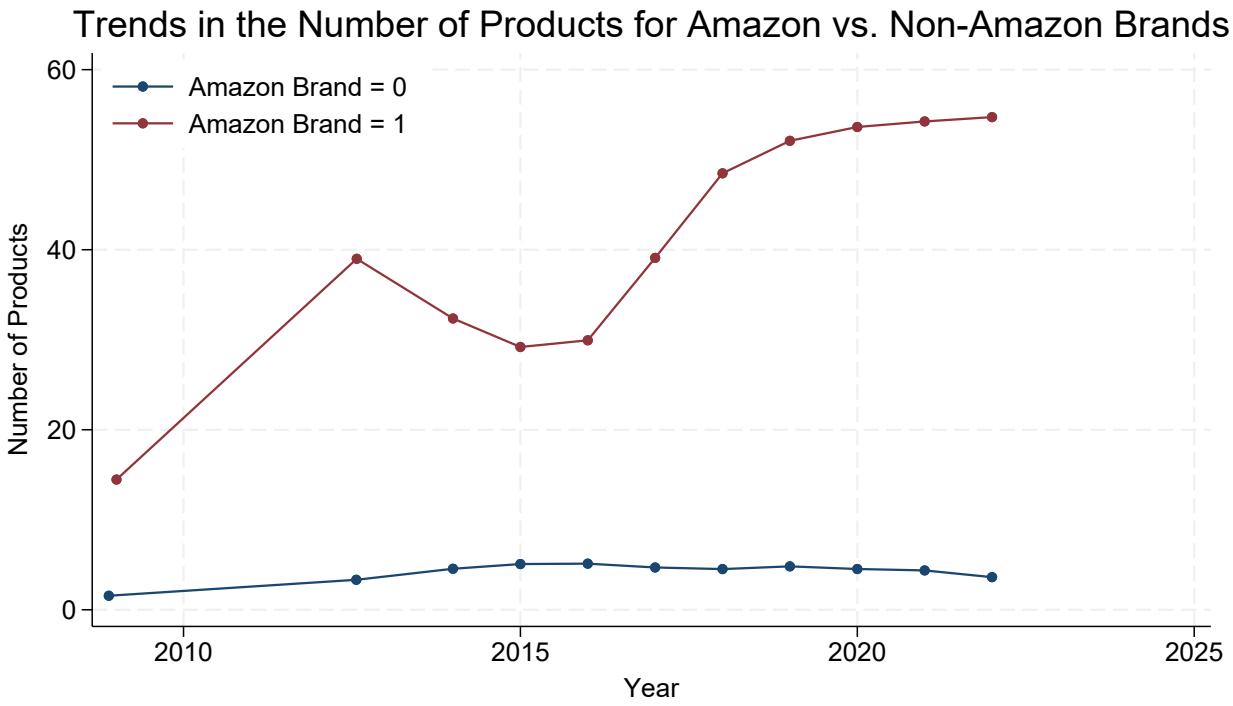



**Figure 3. Trends in the Number of Products for Amazon vs. Non-Amazon Brands, 2000–2022**
This figure plots the average number of products offered by Amazon-affiliated and non-Amazon brands from 2000 to 2022. Each line represents the average number of active products at the category–month level, separately for Amazon brands (in red) and non-Amazon brands (in blue). The figure illustrates that Amazon's private-label brands have expanded their product variety substantially over time, especially after 2015, while the number of products offered by non-Amazon brands has remained relatively stable.

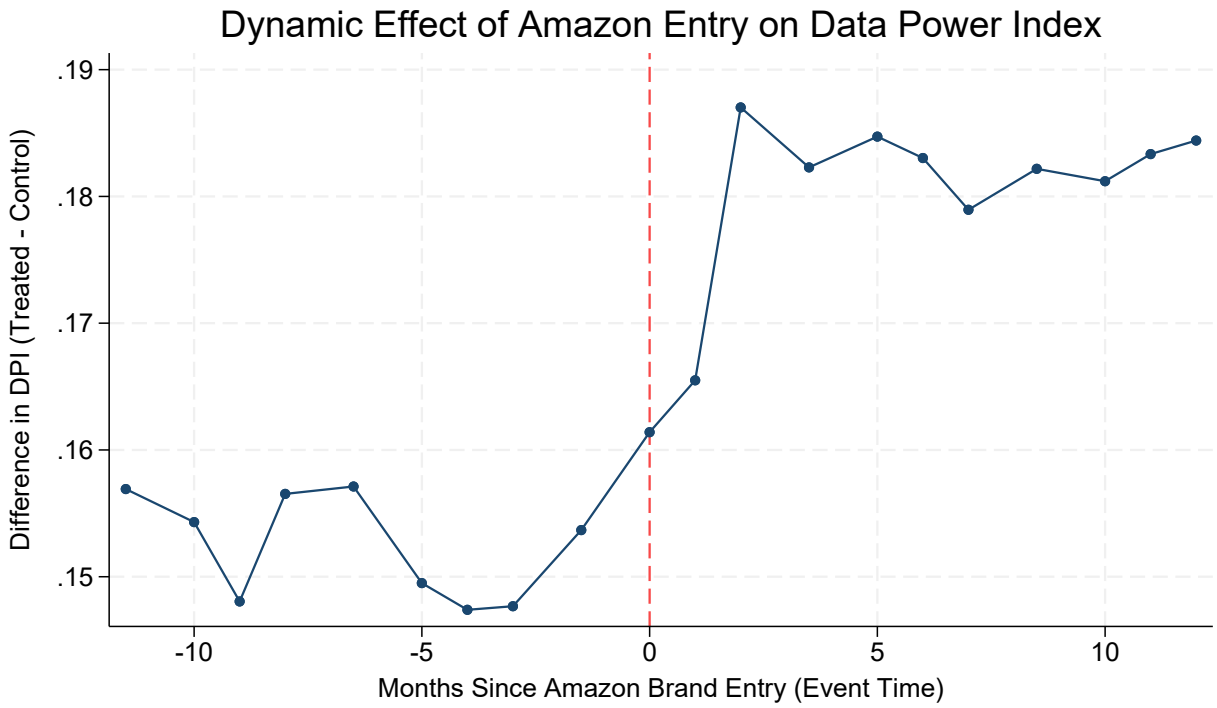



**Figure 4. Data Power Index Around Amazon Brand Entry Between Treated and Control Categories**

This figure plots the mean difference in the Data Power Index (DPI) between treated and control categories within a $\pm 12$-month window around the timing of Amazon brand entry (event time $t = 0$). Each point represents the average DPI difference at a given month relative to the brand entry, and the solid line connects these values over time. The dashed vertical line marks the month of Amazon brand's entry.

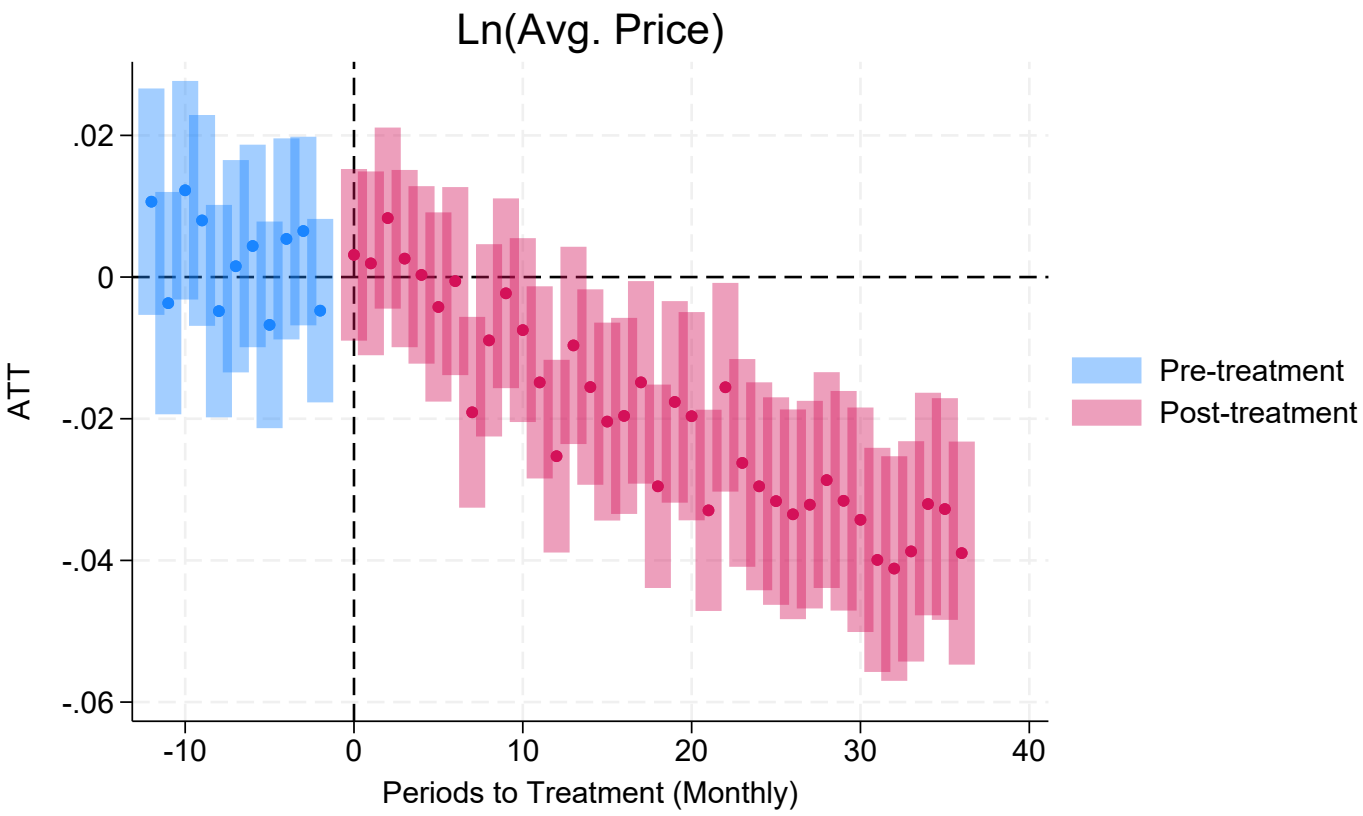



**Figure 5. Dynamic Effects of Treatment on Ln.(Avg. Product Price))**
This event-study figure is produced using the CSDID package in STATA, following the multiple-period difference-in-differences estimator from (Callaway and Sant'Anna, 2021). The x-axis reports event time in months relative to treatment (0 = Months that Amazon brands enter). The y-axis shows the estimated average treatment effect on the treated (ATT) for the logarithm value of average product price. Blue dots and vertical bars represent pre-treatment periods, while pink dots and vertical bars represent post-treatment periods; vertical lines denote 95 percent confidence intervals. The horizontal dashed line indicates a zero effect.

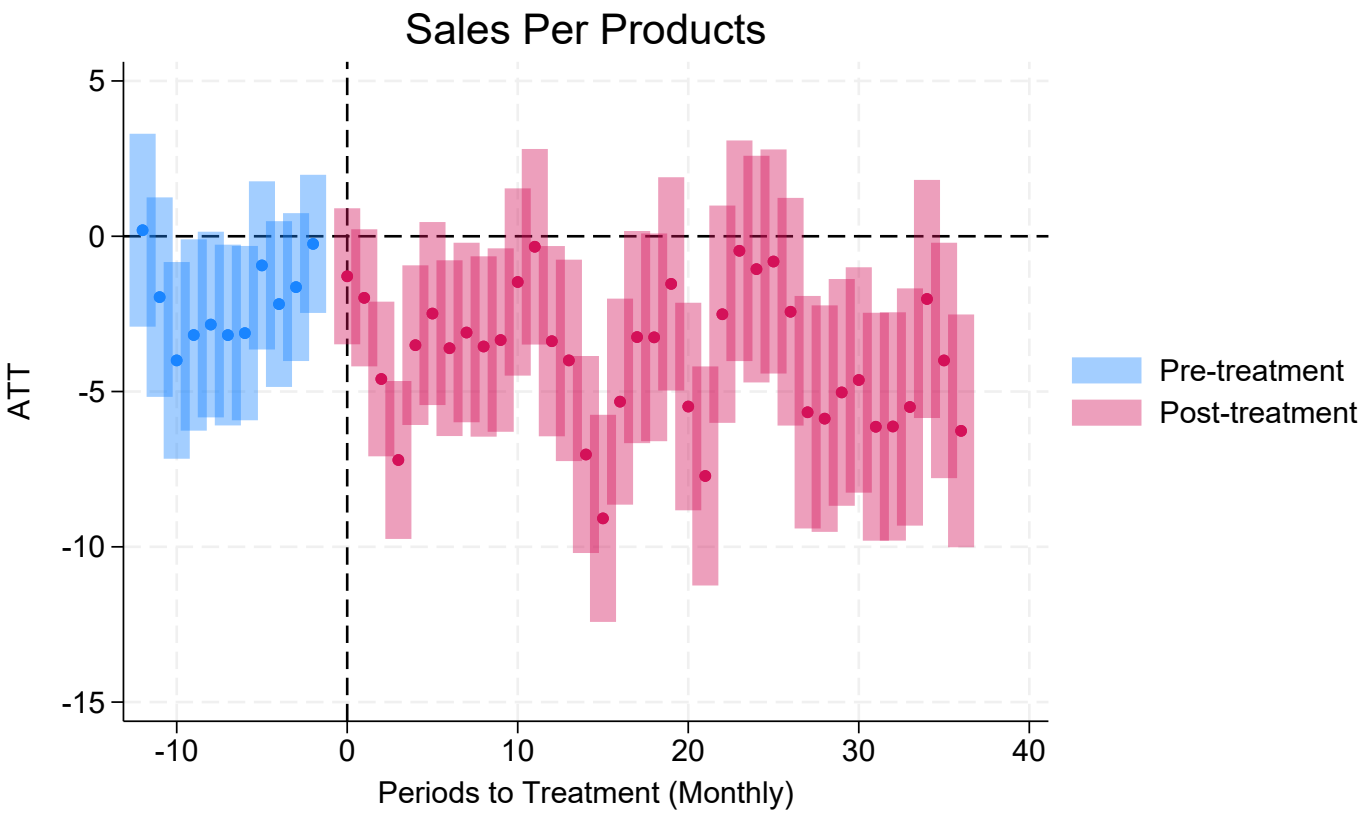



**Figure 6. Dynamic Effects of Treatment on Sales Per Product**
This event-study figure is produced using the CSDID package in Stata, following the multiple-period difference-in-differences estimator from (Callaway and Sant'Anna, 2021). The x-axis reports event time in months relative to treatment (0 = the month that Amazon brands enter). The y-axis shows the estimated average treatment effect on the treated (ATT) for sales per product. Blue dots and shaded bars represent pre-treatment periods, while pink dots and shaded bars represent post-treatment periods. The shaded bars denote the 95 percent confidence intervals, and the horizontal dashed line indicates a zero effect.

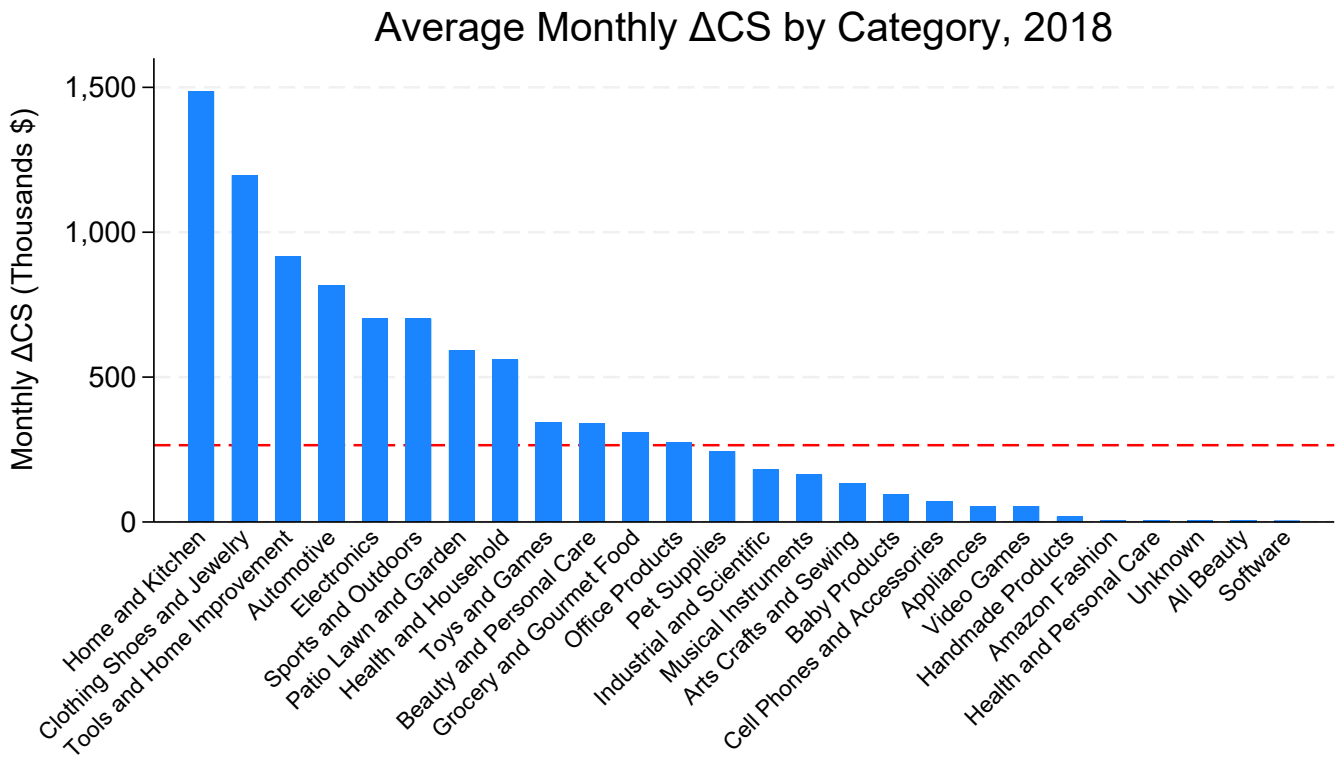



**Figure 7. Average Monthly Consumer Surplus ($\Delta CS$) by Category, 2018**
This bar chart reports mean monthly changes in consumer surplus at the category level for 2018. Categories are ordered from highest to lowest average surplus. The dashed horizontal line marks a $265,000 reference level. Values are in U.S. thousands dollars.

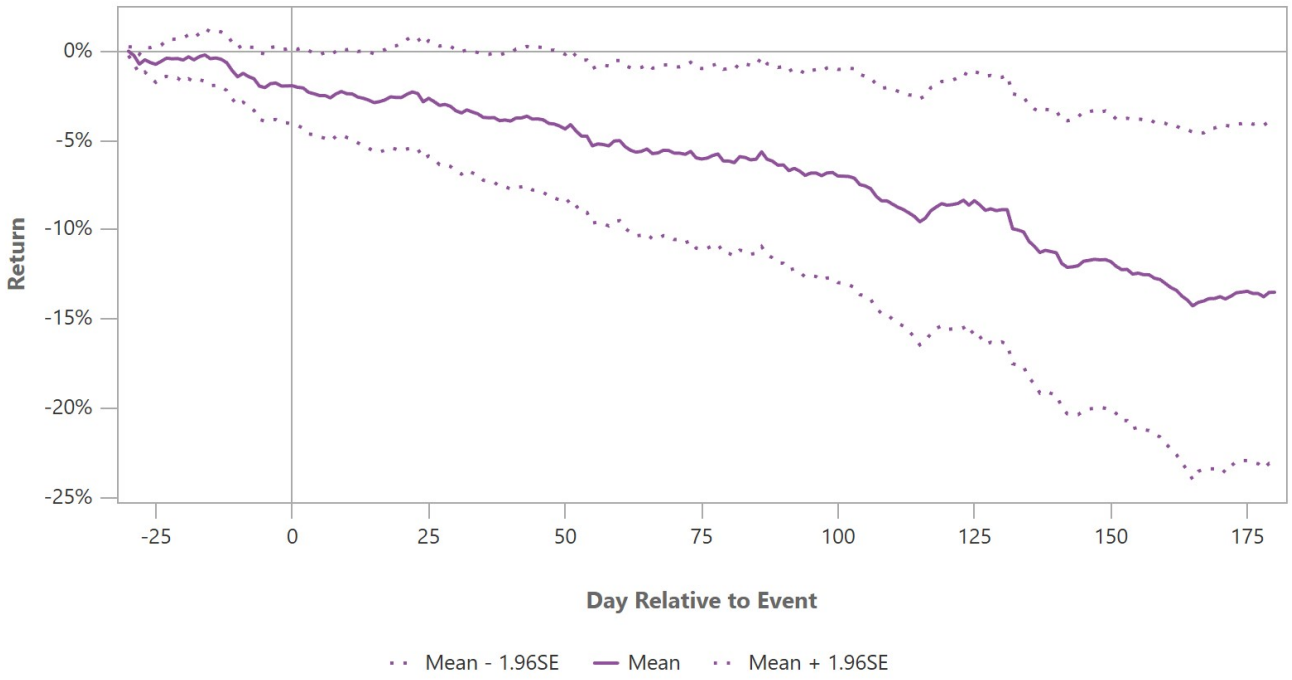



**Figure 8. Cumulative Abnormal Returns (CAR) Around Amazon Brand's Entry**
This figure presents the cumulative abnormal return (CAR) of firms listed in Amazon around Amazon Brand's entry date. The analysis uses the Fama–French Four–Factor Model (including momentum) as the risk adjustment model, with a 100–day estimation window, 30–day gap, and an event window spanning from 30 days before to 180 days after the event date. The y–axis reports cumulative abnormal returns, and the x–axis represents trading days relative to the event. The solid purple line shows the mean CAR across 289 events with non–missing returns, while the dotted lines denote the 95 percent confidence interval.

**Table 1. Summary Statistics of Amazon Product Market**

This table presents summary statistics for Amazon's product market characteristics, constructed at the store–month level. The variables include the number of exiting products, enters, and exits, consumer review and rating, the average product price, total and per-product sales, margin and per-product margin, and total fees (including FBA, storage, and referral fees). All monetary variables are expressed in U.S. dollars. Observations are winsorized at the 1st and 99th percentiles. The sample spans from January 2000 to December 2022.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | N | Mean | Std.Dev. | p10 | Median | p90 |
| # of Existing Products | 10,811,312 | 4.19 | 9.69 | 0.00 | 1.00 | 10.00 |
| # of Product Entry | 10,811,312 | 2.20 | 5.15 | 0.00 | 0.00 | 5.00 |
| # of Product Exit | 10,811,312 | 2.17 | 5.13 | 0.00 | 0.00 | 5.00 |
| # of Consumer Review | 10,811,312 | 15.45 | 110.27 | 0.00 | 1.00 | 27.00 |
| Avg. Consumer Rating | 10,811,312 | 2.60 | 2.11 | 0.00 | 3.58 | 5.00 |
| Avg. Product Price ($) | 4,835,524 | 42.40 | 59.99 | 9.99 | 22.99 | 89.99 |
| Total Sales ($) | 4,835,524 | 667.19 | 1826.37 | 15.22 | 119.50 | 1431.00 |
| Sales per Product ($) | 4,835,524 | 80.10 | 146.55 | 6.66 | 31.60 | 184.43 |
| Margin ($) | 4,835,524 | 530.06 | 1494.55 | 7.82 | 85.20 | 1131.59 |
| Avg. Margin per Product ($) | 4,835,524 | 8.87 | 77.53 | 0.00 | 0.00 | 23.71 |
| Total FBA Fee ($) | 4,835,524 | 56.08 | 1224.45 | 0.00 | 3.07 | 76.32 |
| Total Storage Fee ($) | 4,835,524 | 0.87 | 7.71 | 0.00 | 0.01 | 1.19 |
| Total Referral Fee ($) | 4,835,524 | 139.90 | 1238.55 | 2.22 | 17.24 | 208.09 |
| Sum of Total Fee ($) | 4,835,524 | 196.85 | 1792.18 | 3.40 | 28.18 | 306.57 |

**Table 2. Summary Statistics of Firm-linked in Amazon**

This table presents summary statistics for firm-quarter level variables matched to Amazon's product market data. All financial variables are obtained from Quarterly Compustat and expressed in millions of U.S. dollars. *Firm size (Mktcap) ($M)* is the market capitalization at the end of each quarter. *BM ratio* denotes the book-to-market ratio. *Total assets ($M)*, *sales ($M)*, *net income ($M)*, and *R&D expenditures ($M)* correspond to quarterly values. *Number of Amazon stores* measures the total count of distinct Amazon stores linked to each firm during the corresponding quarter. Observations are winsorized at the 1st and 99th percentiles. The sample period spans 2000–2022.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | N | Mean | Std.Dev. | p10 | Median | p90 |
| Firm size (Mktcap) ($M) | 33,477 | 64,878.34 | 114,346.48 | 170.65 | 5,579.02 | 241,626.19 |
| BM ratio | 33,371 | 0.46 | 0.76 | 0.15 | 0.36 | 0.95 |
| Total Assets ($M) | 33,435 | 52,675.30 | 156,877.94 | 173.78 | 5,790.89 | 155,117.00 |
| Quarterly sales ($M) | 33,538 | 5,436.99 | 9,951.22 | 37.31 | 865.02 | 17,811.00 |
| Net income ($M) | 33,520 | 817.62 | 1,815.05 | -10.02 | 53.64 | 3,511.00 |
| R&D expenditures ($M) | 22,783 | 850.84 | 1,202.57 | 0.00 | 45.43 | 2,599.00 |
| Number of Amazon stores | 34,323 | 15.06 | 24.48 | 1.00 | 3.00 | 63.00 |

**Table 3. Summary Statistics of Data Power Index (DPI)**

This table presents summary statistics for the constructed Data Power Index (DPI), which measures the informational richness of Amazon's product categories. The three versions of the index, the baseline *DPI*, $DPI^{VQ}$, and $DPI^{DQ}$, capture different dimensions of data precision and coverage at the category–month level. All DPI measures are standardized within month to range between 0 and 1. Panel A presents summary statistics for the three DPI measures, including their mean, standard deviation, and percentile distribution across category–month observations. Panel B reports the correlation matrix among the DPI measures. The sample period spans from 2000 to 2022.

**Panel A: Summary Statistics**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | N | Mean | Std.Dev. | p10 | Median | p90 |
| *DPI* | 6,723 | 0.52 | 0.28 | 0.13 | 0.52 | 0.92 |
| $DPI^{(VQ)}$ | 6,723 | 0.51 | 0.27 | 0.14 | 0.51 | 0.87 |
| $DPI^{(DQ)}$ | 6,723 | 0.49 | 0.23 | 0.17 | 0.49 | 0.80 |

**Panel B: Correlation Matrix**

| | *DPI* | $DPI^{VQ}$ | $DPI^{DQ}$ |
|---|---|---|---|
| *DPI* | 1 | | |
| $DPI^{VQ}$ | .91 | 1 | |
| $DPI^{DQ}$ | .77 | .75 | 1 |

**Table 4. The Relationship between Data Power Index (DPI) And Product Market Dynamics**
This table reports panel regressions linking the lagged category–month *Data Power Index* (DPI) to store outcomes in Amazon's marketplace. We estimate the following equation:

$$Y_{i,c,t} = \alpha + \beta\,\mathrm{DPI}_{c,t-1} + X'_{i,c,t-1}\theta + \delta_i + \gamma_t + \varepsilon_{i,c,t},$$

where $Y_{i,c,t}$ is the product market outcome variable as follows: (1) the number of existing products, and (2) rolling three–month of net new products, (3) logarithm value of average product price, (4) sales per product, and (5) average margin per product. The independent variable, $\mathrm{DPI}_{c,t-1}$, is the lagged category level DPI constructed in equation (7) to (9). The vector $X_{i,c,t-1}$ collects time varying controls, and we include store fixed effects and year–month fixed effects. Standard errors are two-way clustered at store and year-month level. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively. The sample covers from January 2000 to December 2022.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Dependent | # of Existing | Rolling 3-month | Ln(Avg. | Sales | Avg. Margin |
| Variable= | Products | (Δ New Products) | Product Price) | per Product | per Product |
| $DPI_{c,t-1}$ | 0.29*** | 0.47*** | -0.06*** | 27.25*** | -3.35*** |
| | (5.42) | (5.45) | (-6.14) | (8.67) | (-5.09) |
| $HHI_{c,t-1}$ | -0.00 | 0.01 | -0.00 | 0.39 | 0.04 |
| | (-0.80) | (0.88) | (-0.48) | (1.47) | (0.71) |
| $Avg.Rating_{i,c,t-1}$ | 0.12*** | 0.05*** | 0.00*** | 0.28** | -0.03 |
| | (22.43) | (11.92) | (10.18) | (2.57) | (-0.43) |
| $VerifiedReviews_{i,c,t-1}$ | 0.66*** | -0.19*** | -0.01*** | 10.76*** | 0.91*** |
| | (29.14) | (-7.29) | (-5.93) | (20.85) | (3.10) |
| Adjusted $R^2$ | 0.701 | 0.010 | 0.854 | 0.646 | 0.359 |
| Mean of Dep. Var. | 4.231 | 0.0910 | 3.274 | 80.54 | 8.919 |
| Stores FEs | Yes | Yes | Yes | Yes | Yes |
| Year-Month FEs | Yes | Yes | Yes | Yes | Yes |
| Observations | 10,653,937 | 10,653,937 | 4,764,767 | 4,764,767 | 4,764,767 |

**Table 5.**

**The Relationship between Data Power Index (DPI) and Firm Sales & Operating Outcomes**

This table reports firm–quarter panel regressions linking the one–quarter lagged average of the category–level *Data Power Index* (DPI) to firms' sales and operating outcomes. We estimate:

$$Y_{i,c,t} = \alpha + \beta\,\mathrm{DPI}_{c,t-1} + X'_{i,c,t-1}\theta + \delta_i + \gamma_t + \varepsilon_{i,c,t},$$

where $Y_{i,t}$ is, respectively: (1) revenue, (2) sales, (3) cost of goods sold, (4) operating expense, and (5) SG&A expense. All dependent variables are normalized by total assets. $X_{i,t-1}$ collects time–varying controls. All specifications include store fixed effects ($\delta_i$) and year–quarter fixed effects ($\gamma_t$). Standard errors are two–way clustered by firm and year–quarter. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively. The sample spans January 2000 to December 2022.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Dependent Variable= | Revenue | Sales | Cost of Goods Sold | Operating Expense | SG&A Expense |
| $DPI_{c,t-1}$ | -0.0343*** | -0.0343*** | -0.0188** | -0.0309*** | -0.0113*** |
| | (-3.27) | (-3.31) | (-2.24) | (-3.22) | (-3.35) |
| $HHI_{c,t-1}$ | 0.0004 | 0.0004 | 0.0006 | 0.0007 | 0.0001 |
| | (0.73) | (0.76) | (1.30) | (1.31) | (0.29) |
| $FirmSize_{i,c,t-1}$ | 0.0198*** | 0.0197*** | 0.0064*** | 0.0073** | 0.0012 |
| | (6.15) | (6.16) | (2.68) | (2.54) | (1.21) |
| $BM_{i,c,t-1}$ | -0.0018 | -0.0017 | -0.0007 | -0.0019 | -0.0010 |
| | (-0.73) | (-0.71) | (-0.42) | (-0.89) | (-1.47) |
| $Ln(Assets)_{i,c,t-1}$ | -0.0741*** | -0.0737*** | -0.0394*** | -0.0615*** | -0.0232*** |
| | (-8.92) | (-8.93) | (-5.71) | (-7.86) | (-11.58) |
| Constant | 0.7023*** | 0.6980*** | 0.4225*** | 0.6729*** | 0.2602*** |
| | (10.09) | (10.07) | (7.20) | (10.08) | (14.90) |
| Adjusted $R^2$ | 0.910 | 0.911 | 0.928 | 0.920 | 0.893 |
| Mean of Dep. Var. | 0.219 | 0.217 | 0.129 | 0.191 | 0.0653 |
| Store FEs | Yes | Yes | Yes | Yes | Yes |
| Year-Quarter FEs | Yes | Yes | Yes | Yes | Yes |
| Observations | 32,424 | 32,732 | 32,681 | 32,680 | 30,837 |

**Table 6. The Relationship between Data Power Index (DPI) and Innovation Outcomes**

This table reports firm–quarter panel regressions linking the lagged, three–quarter moving average of the category–level *Data Power Index* (DPI(1)) to firms' financial and innovation outcomes. We estimate the following specification:

$$Y_{i,c,t} = \alpha + \beta\,\mathrm{DPI}_{c,t-1} + X'_{i,c,t-1}\theta + \gamma_i + \delta_t + \varepsilon_{i,c,t},$$

where $Y_{i,c,t}$ represents firm–quarter outcomes, including: (1) cumulative number of active patents, (2) patent portfolio age (active), and (3) patent portfolio age (filing). All dependent variables are normalized by total assets. $X_{i,c,t-1}$ denotes time–varying firm controls. All regressions include firm fixed effects ($\gamma_i$) and year–quarter fixed effects ($\delta_t$). Standard errors are two–way clustered by firm and year–quarter. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively. The sample spans January 2000 to December 2022.

| | (1) | (2) | (3) |
|---|---|---|---|
| Dependent Variable= | # of Patents (Active) | Patent Pf. Age (Active) | Patent Pf. Age (Filling) |
| $DPI_{c,t-1}$ | 0.0859* | -2.1455*** | -1.4123** |
| | (1.79) | (-3.05) | (-2.22) |
| $HHI_{c,t-1}$ | 0.0024 | 0.0420 | 0.0281 |
| | (1.12) | (1.44) | (0.99) |
| $FirmSize_{i,c,t-1}$ | 0.0015 | -0.0430 | -0.1382 |
| | (0.06) | (-0.32) | (-1.13) |
| $BM_{i,c,t-1}$ | 0.0005 | -0.1711 | -0.2152* |
| | (0.02) | (-1.52) | (-1.89) |
| $Ln(Assets)_{i,c,t-1}$ | 0.2557*** | -0.9522*** | -0.8999*** |
| | (5.90) | (-4.41) | (-3.92) |
| Constant | 7.0877*** | 22.6939*** | 23.8879*** |
| | (12.45) | (13.30) | (14.06) |
| Adjusted $R^2$ | 0.994 | 0.975 | 0.977 |
| Mean of Dep. Var. | 8589.0 | 12.69 | 13.86 |
| Store FEs | Yes | Yes | Yes |
| Year-Quarter FEs | Yes | Yes | Yes |
| Observations | 27,730 | 27,867 | 27,867 |

**Table 7. Predicting Amazon Brand Entry with DPI and Market Characteristics**

This table reports probit estimates for the probability that category $c$ experiences its first Amazon private label entry in month $t$. The estimation model is

$$\Pr(\text{Entry}_{c,t} = 1) = \Phi\big(\alpha\,\text{DPI}_{c,t-1} + \mathbf{X}'_{c,t-1}\theta + \delta_c + \gamma_y + \mu_m\big),$$

Dependent variable is an indicator equal to one if category $c$ experiences its first Amazon private-label brand entry in month $t$, and zero otherwise. Main independent variable is the lagged category-level DPI $\text{DPI}_{c,t-1}$. All other independent variables are category-level averages measured at $t-1$: average product price, sales per product, margin rate, category sales concentration (HHI), average number of reviews, average rating, and the number of products. Standard errors are clustered at the category-month level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. The sample spans January 2000 to December 2022.

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Dependent Variable= | Amazon Private Label Entry (=1) | | | | | |
| $DPI_{c,t-1}$ | 1.89*** | 2.34*** | 1.95** | 2.28* | | |
| | (2.93) | (3.28) | (2.10) | (1.86) | | |
| $Avg.ProductPrice_{c,t-1}$ | | -0.01 | | -0.01* | -0.01 | |
| | | (-1.46) | | (-1.72) | (-1.38) | |
| $SalesperProduct_{c,t-1}$ | | 0.02** | | 0.03*** | 0.02** | |
| | | (2.22) | | (3.50) | (2.21) | |
| $MarginRate_{c,t-1}$ | | 0.18 | | -0.14 | 0.28 | |
| | | (0.18) | | (-0.15) | (0.30) | |
| $HHI_{c,t-1}$ | | -0.01 | | -0.03 | -0.01 | |
| | | (-0.17) | | (-0.58) | (-0.20) | |
| $Avg.\#Reviews_{c,t-1}$ | | | 0.03 | -0.17 | | 0.07 |
| | | | (0.32) | (-1.26) | | (0.81) |
| $Avg.Rating_{c,t-1}$ | | | 0.35 | 1.28*** | | 0.49 |
| | | | (0.86) | (3.61) | | (1.14) |
| $Avg.\#Products_{c,t-1}$ | | | -0.65*** | -0.64** | | -0.72*** |
| | | | (-2.94) | (-2.28) | | (-3.13) |
| Constant | -4.05*** | -5.82*** | -1.83 | -4.69** | -3.61*** | -0.53 |
| | (-6.41) | (-5.47) | (-1.38) | (-2.18) | (-4.12) | (-0.45) |
| Pseudo $R^2$ | 0.031 | 0.084 | 0.068 | 0.138 | 0.066 | 0.054 |
| Year FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| Month FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| Category FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| Observations | 574 | 564 | 574 | 564 | 564 | 576 |

**Table 8. The Impact of Amazon Brand Entry on Product Market Outcomes**

This table reports panel regressions estimating the impact of Amazon private–brand entry on product–market outcomes at the store–category–month level. We estimate:

$$Y_{i,c,t} = \alpha + \beta \text{Treat}_{c,t} + X'_{i,c,t}\theta + \gamma_i + \delta_t + \varepsilon_{i,c,t},$$

where $\text{Treat}_{c,t}$ equals one after Amazon's private–brand entry in category $c$. The dependent variables are, respectively: (1) the number of existing products, and (2) rolling three–month of net new products, (3) logarithm value of average product price, (4) sales per product, and (5) average margin per product. The vector $X_{i,c,t-1}$ collects time varying controls, and we include store fixed effects and year–month fixed effects. Standard errors are two-way clustered at store and year-month level. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively. The sample covers from January 2000 to December 2022.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Dependent | # of Existing | Rolling 3-month | Ln(Avg. | Sales | Avg. Margin |
| Variable= | Products | (Δ New Products) | Product Price) | per Product | per Product |
| $Treat_{c,t}$ | 0.04** | 0.10* | -0.06*** | -9.57*** | -1.54*** |
| | (2.21) | (1.77) | (-11.71) | (-6.55) | (-3.59) |
| Avg. $\text{Rating}_{i,c,t}$ | 0.23*** | 0.06*** | 0.00*** | 2.73*** | 0.11*** |
| | (31.34) | (13.41) | (8.47) | (23.82) | (2.59) |
| #$\text{Verified Review}_{i,c,t}$ | 1.31*** | -0.26*** | -0.05*** | 74.03*** | 8.56*** |
| | (27.80) | (-8.57) | (-20.27) | (61.94) | (20.78) |
| Adjusted $R^2$ | 0.727 | 0.008 | 0.854 | 0.646 | 0.358 |
| Mean of Dep. Var. | 4.186 | 0.0965 | 3.273 | 80.12 | 8.867 |
| Stores FEs | Yes | Yes | Yes | Yes | Yes |
| Year-Month FEs | Yes | Yes | Yes | Yes | Yes |
| Observations | 10,810,326 | 10,810,326 | 4,832,421 | 4,832,421 | 4,832,421 |

**Table 9. The Impact of Amazon Brand Entry On Consumer Welfare**

This table reports panel regressions estimating the effect of Amazon private brand entry on consumer surplus at the store–category–month level. We estimate:

$$Y_{i,c,t} = \alpha + \beta \text{Treat}_{c,t} + X'_{i,c,t}\theta + \gamma_i + \delta_t + \varepsilon_{i,c,t},$$

where $\text{Treat}_{c,t}$ equals one after Amazon's private brand entry in category $c$. The dependent variables are: (1) $\Delta CS$ (Store, stars, M$), and (2) $\Delta CS$ (Market, stars, M$). $\Delta CS$ denotes the month-over-month change in consumer surplus computed via the log-sum formula, and "stars" uses ratings only to map pre-purchase information into utility. "Store" computes the contribution to surplus at the store level, "Market" computes category-level surplus, and "M$" indicates the unit of million dollars. The vector $X_{i,c,t}$ includes time-varying controls such as ln(Average Product Price), *Average Rating*, and *Number of Verified Reviews*. All specifications include store fixed effects ($\gamma_i$), category fixed effects, and year–month fixed effects ($\delta_t$). Standard errors are two-way clustered by store and year–month. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. The sample spans January 2000 to December 2022.

| | (1) | (2) |
|---|---|---|
| Dependent Variable = | $\Delta CS$ (Store, stars, $) | $\Delta CS$ (Market, stars, M$) |
| $Treat_{c,t}$ | -103.38*** | 0.22** |
| | (-6.53) | (2.07) |
| Adjusted $R^2$ | 0.662 | 0.581 |
| Mean of Dep. Var. | 266.2 | 0.186 |
| Controls | Yes | Yes |
| Stores FEs | Yes | |
| Category FEs | | Yes |
| Year-Month FEs | Yes | Yes |
| Observations | 4,832,421 | 6,932 |

**Table 10. The Impact of Amazon Brand Entry on Firm Risks**

This table reports panel regressions estimating the effect of Amazon private brand entry on firm-level volatility and systematic risk. We estimate:

$$Y_{i,c,t} = \alpha + \beta \text{Treat}_{c,t} + X'_{i,c,t}\theta + \gamma_i + \delta_t + \varepsilon_{i,c,t},$$

where $\text{Treat}_{i,t}$ equals one after Amazon's private brand entry in the corresponding product category. The dependent variables are: (1) $\sigma_{i,t}^{Operating\ Income}$, (2) Idiosyncratic Volatility, (3) Total Volatility, and (4) $\beta_{i,t}^{MKT}$. Firm-level control variables include *Firm Size*, *BM Ratio*, ln(Assets), and factor loadings on SMB, HML, and UMD. All regressions include store fixed effects ($\gamma_i$) and year–month fixed effects ($\delta_t$). Standard errors are two-way clustered by firm and year–month. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. The sample spans January 2000 to December 2022.

| | (1) $\sigma_{i,t}^{\text{Operating Income}}$ | (2) Idiosyncratic Volatility | (3) Total Volatility | (4) $Beta_{i,t}^{MKT}$ |
|---|---|---|---|---|
| $Treat_{c,t}$ | 0.062** | 0.008** | 0.012*** | 0.087** |
| | (2.62) | (2.20) | (2.68) | (2.12) |
| $FirmSize_{i,c,t}$ | -0.076*** | -0.011*** | -0.013*** | -0.112*** |
| | (-4.14) | (-3.19) | (-3.64) | (-3.54) |
| $BM_{i,c,t}$ | -0.021 | -0.003 | -0.002 | -0.013 |
| | (-0.83) | (-1.10) | (-0.69) | (-0.48) |
| $Ln(Assets)_{i,c,t}$ | 0.044 | 0.002 | 0.007 | 0.232*** |
| | (1.27) | (0.47) | (1.51) | (4.96) |
| $Beta_{i,c,t}^{SMB}$ | | | | -0.098*** |
| | | | | (-4.74) |
| $Beta_{i,c,t}^{HML}$ | | | | -0.040 |
| | | | | (-1.13) |
| $Beta_{i,c,t}^{UMD}$ | | | | 0.143*** |
| | | | | (4.01) |
| Constant | 0.587** | 0.151*** | 0.149*** | -0.053 |
| | (2.49) | (5.21) | (4.53) | (-0.17) |
| Adjusted $R^2$ | 0.682 | 0.843 | 0.822 | 0.619 |
| Mean Value of Dep. Var. | 0.317 | 0.0788 | 0.0986 | 0.947 |
| Store FEs | Yes | Yes | Yes | Yes |
| Year-Month FEs | Yes | Yes | Yes | Yes |
| Observations | 17,870 | 20,190 | 20,190 | 20,190 |

**Table 11. The Impact of Amazon Brand Entry On Innovations**

This table reports panel regressions estimating the effect of Amazon private brand entry on firms' patent portfolio age at the firm–category–month level. We estimate:

$$Y_{i,c,t} = \alpha + \beta \text{Treat}_{c,t} + X'_{i,c,t}\theta + \gamma_i + \delta_t + \varepsilon_{i,c,t},$$

where $\text{Treat}_{c,t}$ equals one after Amazon's private brand entry in category $c$. The dependent variables are: (1) *Patent Portfolio Age (Active)* and (2) *Patent Portfolio Age (Filing)*. Both variables measure the average age (in years) of patents currently active or filed by the firm within the category. The vector $X_{i,c,t}$ includes time-varying controls such as *Firm Size*, *BM Ratio*, and ln(Assets). All specifications include store fixed effects ($\gamma_i$), and year–month fixed effects ($\delta_t$). Standard errors are two-way clustered by firm and year–month. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. The sample spans January 2000 to December 2022.

| | (1) | (2) |
|---|---|---|
| | Patent Pf. Age (Active) | Patent Pf. Age (Filling) |
| $Treat_{c,t}$ | -0.646*** | -0.355* |
| | (-2.76) | (-1.73) |
| $FirmSize_{i,c,t}$ | -0.012 | -0.081 |
| | (-0.09) | (-0.69) |
| $BM_{i,c,t}$ | -0.072 | -0.090 |
| | (-1.10) | (-1.27) |
| $Ln(Assets)_{i,c,t}$ | -0.977*** | -0.964*** |
| | (-4.44) | (-4.23) |
| Constant | 21.729*** | 23.249*** |
| | (13.41) | (14.38) |
| Adjusted $R^2$ | 0.975 | 0.977 |
| Mean Value of Dep. Var. | 12.63 | 13.79 |
| Store FEs | Yes | Yes |
| Year-Month FEs | Yes | Yes |
| Observations | 28,121 | 28,121 |

# Internet Appendix for: "(Digital) Platform Data Power"

## A Amazon Private Brand's Entry

**Table IA.1. Amazon Private Label Brand Entry by Category**

This table summarizes the timing and scope of Amazon's private label brand entries across major product categories. For each treated category, we report the month and year when Amazon first launched a private label brand and list the main brands introduced under that category. Excluded categories are Books, CD and Vinyl, Digital Music, Gift Cards, Kindle Store, Magazine Subscriptions, Movies and TV, and Subscription Boxes.

| Amazon Category | Amazon Brand Entry Date | Amazon Brands Launched |
|---|---|---|
| Clothing, Shoes & Jewelry | 2003.12 | Amazon Collection, Amazon Essentials, Amazon Gear, Amazon Aware |
| Home & Kitchen | 2006.02 | Pinzon, Amazon Basics |
| Electronics | 2009.09 | Amazon Basics |
| Tools & Home Improvement | 2010.03 | Amazon Basics, Amazon Commercial |
| Office Products | 2010.04 | Amazon Basics |
| Health & Household | 2013.01 | Amazon Basic Care, Solimo, Amazon Elements |
| Grocery & Gourmet Food | 2014.10 | 365 by Whole Foods Market, Happy Belly, Amazon Fresh |
| Baby Products | 2016.08 | Mama Bear |
| Beauty & Personal Care | 2018.04 | Solimo |

**Table IA.2. The Relationship between HHI And Product Market Dynamics**

This table reports panel regressions linking the lagged category–month HHI to store outcomes in Amazon's marketplace. We estimate the following equation:

$$Y_{i,c,t} = \alpha + \beta\,\mathrm{HHI}_{c,t-1} + X'_{i,c,t-1}\theta + \delta_i + \gamma_t + \varepsilon_{i,c,t},$$

where $Y_{i,c,t}$ is the product market outcome variable as follows: (1) logarithm value of average product price, (2) sales per product, (3) average margin per product, (4) the number of existing products, and (5) rolling three–month of net new products. The independent variable, $\mathrm{HHI}_{c,t-1}$, is the lagged category level Herfindahl–Hirschman Index (HHI). The vector $X_{i,c,t-1}$ collects time varying controls, and we include store fixed effects and year–month fixed effects. Standard errors are two-way clustered at store and year-month level. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively. The sample covers from January 2000 to December 2022.

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Dependent | Ln(Avg. | Sales | Avg. Margin | # of Existing | Rolling 3-month |
| Variable= | Product Price) | per Product | per Product | Products | (Δ New Products) |
| $HHI_{c,t-1}$ | -0.00 | 0.81*** | -0.01 | 0.00 | 0.01 |
| | (-1.36) | (2.83) | (-0.22) | (0.78) | (1.48) |
| $Avg.Rating_{i,c,t-1}$ | 0.00*** | 0.29*** | -0.03 | 0.12*** | 0.05*** |
| | (10.14) | (2.67) | (-0.45) | (22.29) | (11.97) |
| $VerifiedReviews_{i,c,t-1}$ | -0.01*** | 10.89*** | 0.90*** | 0.66*** | -0.19*** |
| | (-6.14) | (21.22) | (3.04) | (28.88) | (-7.20) |
| Adjusted $R^2$ | 0.854 | 0.646 | 0.359 | 0.700 | 0.010 |
| Mean of Dep. Var. | 3.273 | 80.54 | 8.919 | 8.429 | 0.0910 |
| Stores FEs | Yes | Yes | Yes | Yes | Yes |
| Year–Month FEs | Yes | Yes | Yes | Yes | Yes |
| Observations | 4,764,776 | 4,764,776 | 4,764,776 | 10,653,990 | 10,653,990 |

## Table IA.3. The Relationship between DPI And Competitions

This table reports panel regressions linking the lagged category–month *Data Power Index* (DPI) to several competition measures. We estimate the following equation:

$$Y_{c,t} = \alpha + \beta\,\mathrm{DPI}_{c,t-1} + X'_{c,t-1}\theta + \delta_c + \gamma_t + \varepsilon_{c,t},$$

where $Y_{c,t}$ is the market competition measures as follows: (1) Herfindahl-Hirschman Index (HHI), and (2) Entropy measured by seller's sales in category. The independent variable, $\mathrm{DPI}_{c,t-1}$, is the lagged category level Data Power Index. The vector $X_{c,t-1}$ collects time-varying controls, and we include category fixed effects and year–month fixed effects. Standard errors are clustered at the category level. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively. The sample covers from January 2000 to December 2022.

| | (1) HHI | (2) Entropy |
|---|---|---|
| $DPI_{c,t-1}$ | -3.82** | 0.15** |
| | (-2.27) | (2.18) |
| $Avg.\#Reviews_{c,t-1}$ | 1.33*** | -0.07*** |
| | (13.88) | (-12.31) |
| $Avg.Rating_{c,t-1}$ | -1.83*** | 0.64*** |
| | (-3.35) | (16.46) |
| $Avg.MarginPerProduct_{c,t-1}$ | -0.02 | 0.01*** |
| | (-0.41) | (2.68) |
| Adjusted $R^2$ | 0.430 | 0.915 |
| Mean of Dep. Var. | 3.638 | 4.314 |
| Category FEs | Yes | Yes |
| Year-Month FEs | Yes | Yes |
| Observations | 6,096 | 6,096 |

**Table IA.4. Summary Statistics of Consumer Surplus Measures**

This table reports summary statistics for the consumer surplus (CS) measures constructed following the approach of Reimers and Waldfogel (2021), which maps product-level information such as star ratings and prices into mean utilities and converts these into dollar-valued welfare changes using the log-sum formula. The CS measures are computed at both the store–month level and the category–month (market) level. $\Delta CS$(Store, stars,$) reflects welfare changes derived solely from changes in consumer ratings within a store. Similarly, $\Delta CS$(Market, stars,$) summarize aggregated welfare changes at the broader product-market level. All monetary values are expressed in U.S. dollars and the variables are winsorized at the 1st and 99th percentiles. Panel A covers the full sample from January 2000 to December 2022. Panel B restricts the sample to the year 2018 to compare with the sample period of Reimers and Waldfogel (2021).

**Panel A: Full Sample Period**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | N | Mean | Std.Dev. | p10 | Median | p90 |
| $\Delta CS$(Market, stars,$) | 10,811,312 | 989,934 | 977,524 | 50,362 | 686,287 | 2,212,311 |
| $\Delta CS$(Store, stars,$) | 10,805,364 | 119 | 499 | 0 | 0 | 202 |

**Panel B: Sample Perdiod During 2018**

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | N | Mean | Std.Dev. | p10 | Median | p90 |
| $\Delta CS$(Market, stars,$) | 984,497 | 725,760 | 475,096 | 139,182 | 676,869 | 1,390,403 |
| $\Delta CS$(Store, stars,$) | 984,497 | 113 | 474 | 0 | 0 | 192 |

# B Consumer Surplus Measurement

This appendix formalizes my consumer welfare construction and clarifies how it adapts the framework in Reimers and Waldfogel (2021). Their study develops a transparent procedure to map pre-purchase information such as crowd ratings and expert reviews into mean utilities and to evaluate welfare using the log-sum formula from discrete choice demand. They use observed demand responses to information and compute counterfactual surplus by removing the relevant information channel while holding other market features, such as prices, fixed. I follow the same logic. The key differences are that I implement it on Amazon category-by-month markets, recover quantities from sales ranks, and integrate the resulting welfare measures with a difference-in-differences design around the timing of Amazon private label entry.

## 1. Demand System and Definition of Consumer Surplus

Consider a monthly category market with products $j \in \mathscr{J}$. The representative consumer's utility is

$$u_{ij} = \delta_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim \text{Type I Extreme Value.}$$

Under a nested logit structure with nests $g \in \mathscr{G}$ (defined by stores),

$$\delta_j = \ln s_j - \sigma \ln s_{j|g} - \ln s_0, \tag{B-16}$$

where $s_j = q_j/M$, $s_{j|g} = q_j/Q_g$, $Q_g = \sum_{k\in g} q_k$, $s_0 = 1 - Q/M$, $M$ is market size, $Q = \sum_{k\in\mathscr{J}} q_k$, and $\sigma \in [0,1)$ captures within-nest correlation.

To ensure numerical stability when shares approach zero, I add a small constant $\varepsilon = 10^{-6}$ wherever logs of shares are computed. Within category–month $(c,t)$ I define

$$s_{jt} \;=\; \frac{q_{jt}}{M_{ct}}, \qquad s_{0,ct} \;=\; 1 - \sum_{k\in c} \frac{q_{kt}}{M_{ct}}, \tag{B-17}$$

and for nest $g(j)$ the conditional within-nest term is

$$\ln s_{j|g,t} \;=\; \ln\big(s_{jt} + \varepsilon\big) - \ln\big(s_{gt} + \varepsilon\big), \quad s_{gt} \equiv \sum_{k\in g(j)} \frac{q_{kt}}{M_{ct}}. \tag{B-18}$$

The implemented mean utility is therefore

$$\delta_{jt} \;=\; \big[\ln(s_{jt} + \varepsilon) - \ln(s_{0,ct} + \varepsilon)\big] - \sigma \ln s_{j|g,t}. \tag{B-19}$$

The log-sum formula for the nested logit model implies that consumer surplus in state $X$ is

$$CS_X = \frac{M}{\alpha} \ln\left(1 + \sum_{g \in \mathcal{G}} \left[\sum_{j \in g} \exp\left(\frac{\delta_{j,X}}{1-\sigma}\right)\right]^{1-\sigma}\right). \tag{B-20}$$

Following Reimers and Waldfogel (2021), evaluating the welfare impact of pre-purchase information requires distinguishing between ex-ante decision utility and ex-post experienced utility. The change in consumer surplus from the baseline status quo to a counterfactual state without star ratings is given by:

$$\Delta CS = CS_{\text{status quo}} - CS_{\text{counterfactual}} - \frac{M}{\alpha} \sum_{j \in \mathcal{J}} \gamma_j (R_j - \widehat{R}_j) s_j^{\text{counterfactual}}. \tag{B-21}$$

The final term is an ex-post adjustment reflecting the possibility that what is consumed has an ex-post utility that differs from the ex-ante expectation when information is removed.

## 2. Mapping Ranks to Quantities and Recovering Elasticities

I map daily Amazon sales ranks $r_{jt}$ to weekly quantities via the power law

$$q_{jw} = \sum_{t \in w} A\, r_{jt}^{-B}, \tag{B-22}$$

and estimate $(A, B)$ from observed rank dynamics. Because ranks exhibit persistence, I use a partial adjustment specification to isolate contemporaneous responses of ranks to prices and information. In particular,

$$\ln r_{jt} = \theta \ln r_{j,t-1} + a \ln p_{jt} + g \ln R_{jt} + n \ln(\#\text{ratings}_{jt}) \cdot \ln R_{jt} + \text{controls} + \varepsilon_{jt}. \tag{B-23}$$

I then translate the dynamic response into static elasticities of *quantities* using the rank–quantity mapping:

$$\varepsilon_p = \frac{aB}{1-\theta}, \qquad \varepsilon_{R_j} = \frac{\big(g + n \ln(\#\text{ratings}_j)\big)B}{1-\theta}. \tag{B-24}$$

My setting lacks direct quantities at the product-day level, so I recover them from ranks and then back out elasticities. This preserves the spirit of Reimers and Waldfogel (2021) that maps information to demand while adapting to the Amazon environment.

## 3. Mapping Elasticities to Utility Coefficients and Recovering $\delta$

In the status quo I parameterize

$$\delta_j = \delta_{j0} + \alpha p_j + \gamma R_j + \psi_j, \tag{B-25}$$

where $p_j$ is price, $R_j$ is the star rating, and $\psi_j$ captures discrete informational shifts.

To keep the mapping transparent and empirically tractable, I estimate price sensitivity by regressing the market-adjusted log share ratio on price with category–month fixed effects,

$$\ln \frac{s_{jt} + \varepsilon}{s_{0,ct} + \varepsilon} \;=\; \tilde{\alpha}\, p_{jt} \;+\; \eta_{ct} \;+\; u_{jt}. \tag{B-26}$$

To correct for attenuation bias common in direct price regressions and to align the absolute welfare scale with the benchmarks established in the literature, I calibrate the final marginal utility of income, $\alpha$, by scaling $\tilde{\alpha}$. Conditional on $\alpha$ and $\sigma$, I regress the recovered $\delta_{jt}$ on ratings with category–month fixed effects to obtain $\gamma$,

$$\delta_{jt} \;=\; \gamma R_{jt} \;+\; \xi_{ct} \;+\; e_{jt}. \tag{B-27}$$

## 4. Counterfactuals: Removing Stars

Let $\widehat{R}_j$ denote expected quality without stars. I construct counterfactual mean utilities by removing the incremental contribution of ratings,

$$\delta_j^{\text{no stars}} = \delta_j - \gamma(R_j - \widehat{R}_j). \tag{B-28}$$

Consistent with the counterfactual scenarios in Reimers and Waldfogel (2021), I leave product prices unchanged ($\Delta p_j = 0$), focusing exclusively on the informational value of star ratings. In implementation, I construct $\widehat{R}_j$ by predicting expected quality without stars from observable product and market characteristics using fixed effects to minimize bias in the counterfactual expectation.

## 5. Market Definition, Scaling, and Nesting

I define a market as category by month. Nests follow store by default. I construct market size in two steps. First, I interpolate annual population and platform share series to monthly values, denoted $\text{pop}_t$ and $\text{share}_t$, and form the platform-wide market size

$$M_t^{\text{total}} \;=\; \text{pop}_t \times \text{share}_t. \tag{B-29}$$

Second, I scale to category $c$ using the observed category weight $\omega_{ct}$,

$$\omega_{ct} \;=\; \frac{\sum_{j\in c} q_{jt}}{\sum_{c'}\sum_{k\in c'} q_{kt}}, \qquad M_{ct} \;=\; \max\{1,\; \omega_{ct}\, M_t^{\text{total}}\}. \tag{B-30}$$

## 6. Computational Steps

First, I estimate $(A, B)$ in equation (B-22) and construct $q_{jt}$. Second, I estimate (B-23) and compute elasticities via (B-24). Third, I recover $\alpha$ and $\gamma_j$. Fourth, I construct $\widehat{R}_j$ from observables. Fifth, I build $\delta$ under the no-stars scenario using (B-28). Sixth, I compute $CS$ and $\Delta CS$ incorporating the ex-post utility correction using (B-20) and (B-21). Finally, I aggregate to the category level and allocate to the store level using nest-level inclusive value weights.

To show the exact implementation of the nested logit inclusive value, for any utility vector $\{\delta_{jt}\}_{j\in g}$ within nest $g$, I first compute the nest-level sum $I_{gt}$, and then aggregate across all nests to obtain the market-level $G_{ct}$:

$$I_{gt} \;=\; \sum_{j\in g} \exp\Big(\frac{\delta_{jt}}{1-\sigma}\Big), \qquad G_{ct} \;=\; \sum_{g\in c} I_{gt}^{1-\sigma}, \tag{B-31}$$

$$CS_{ct} \;=\; \frac{M_{ct}}{|\alpha|}\ln\Big(1+G_{ct}\Big). \tag{B-32}$$

All $\Delta CS$ objects are defined by replacing $\delta_{jt}$ with the corresponding counterfactual utilities and computing the difference along with the ex-post correction.

## 7. Implementation

For the market unit and quality variable, I implement the procedure at category by month and use star ratings as the main pre-purchase information measure. For counterfactual scenarios, I compute **No Stars** to isolate the contribution of ratings. For the baseline nesting parameter, I set $\sigma = 0.4$, a conventional calibration at this level of aggregation.

To connect the welfare objects to the difference-in-differences design at the store level, I allocate the market-level welfare changes to individual stores (nests) using their relative inclusive value weight in the counterfactual state:

$$w_{g,t} \;=\; \frac{I_{gt}^{1-\sigma}}{G_{ct}}, \qquad \sum_{g\in c} w_{g,t} = 1. \tag{B-33}$$

The store-level treatment variable is defined as:

$$\Delta CS_{g,t}^{\text{Store}} \;=\; w_{g,t} \times \Delta CS_{c,t}^{\text{Market}}. \tag{B-34}$$

For category-level analysis, I simply use the market-level aggregate $\Delta CS_{c,t}^{\text{Market}}$. I then estimate staggered difference-in-differences with store fixed effects and year–month fixed effects using these constructed outcomes. For presentation, I also report welfare in millions of dollars by dividing by $10^6$; this scaling does not affect inference.

Overall, my method applies the Reimers and Waldfogel (2021) welfare logic to Amazon platform data by defining markets at category by month, recovering quantities from ranks, mapping information to mean utility through elasticities, correctly aggregating log-sums under a nested logit structure, and linking the resulting welfare measures in store and category-level empirical designs.