Choosing a Paper

At the start of our project, we considered two potential papers:

1. ***A Reputational Theory of Firm Dynamics*** (Board & Meyer-ter-Vehn, 2022)
2. ***Artificial intelligence, algorithmic pricing, and collusion*** (Brown & MacKay, 2023)

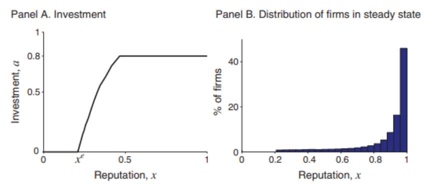


Figure 1: Consumers Observe Firms' Investment

The first paper appeared relatively straightforward, with simple graphs that seemed easy to replicate (see Figure 1 from Board & Meyer-ter-Vehn (2022)). This made it an attractive choice initially.

However, after reviewing the replication package for the second paper (Brown & MacKay (2023))—described in the ***README*** file (Open), we noticed that, despite its complexity, it offered some compelling advantages.

The package also includes multiple well-documented Stata scripts, such as ***07_summary_stats.do*** (Open), which are used to generate the tables and figures in the paper. Initially, we found the concepts in this paper quite complex and the graphs and tables (see Figure 2 and Figure 3) more challenging to replicate.
Additionally, we discovered that the first paper used **simulated data**, which reduced its appeal for our goals.

Ultimately, we chose to proceed with our "Plan B": the Calvano et al. paper. Despite its complexity, it uses real, proprietary data and provides a thoroughly organized and transparent replication package—factors that aligned better with our objectives and offered a more enriching learning experience.

	(1)	(2)	(3)	(4)
Retailer *B*	0.064	0.047	0.146	0.117
	(0.000)	(0.001)	(0.000)	(0.001)
Retailer *C*	0.092	0.107	0.171	0.187
	(0.000)	(0.001)	(0.000)	(0.001)
Retailer *D*	0.249	0.289	0.307	0.337
	(0.000)	(0.001)	(0.000)	(0.001)
Retailer *E*	0.284	0.366	0.340	0.419
	(0.000)	(0.001)	(0.000)	(0.001)
Product fixed effects	Yes	Yes	Yes	Yes
Period fixed effects	Yes	Yes	Yes	Yes
Sold at all retailers			Yes	Yes
On or after July 1, 2019		Yes		Yes
Observations	3,606,956	677,650	1,186,571	234,696

Figure 2: Price Differences for Identical Products Relative to Retailer A

Issues when coding

Throughout our journey, we encountered several challenges, which we can summarize as follows:

1. Interpreting the Stata scripts and mapping each section to its corresponding chart or table in the paper.

2. Managing GitHub publishing—dealing with data size limitations and the dilemma between using a public or private repository.

3. Troubleshooting project structure issues, such as realizing the index file was mistakenly placed outside the main working directory.

4. Handling R crashes—there were moments when R would freeze or become unresponsive; using Ctrl+C proved useful to “unstuck” it.

5. Fine-tuning visualizations—adjusting scaling, colors, and layout to match the presentation style in Brown & MacKay (2023), while ensuring the underlying data aligned correctly.

Graph Replication

For our first replication, we used the data source ***analysis_data.dta*** and followed the script ***07_summary_stats.do*** (Open) to reproduce ***Figure 1: Example Time Series of Prices for Identical Products across Retailers***, found on page 118 of Brown & MacKay (2023).

```
library(haven)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(janitor)
```

```
Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```r
library(patchwork)
# Load and clean
data <- read_dta("C:/Users/Christian Casas/OneDrive - studhsf/Documents/Masters - HS Freseni
  janitor::clean_names()

# Filter for Xyzal 80ct Tablet (non-multipack)
data_xyzal <- data %>%
  filter(
    brand == "Xyzal",
    form == "Tablet",
    size == 80,
    multipack == 1,
    flag_imputed_price != 1
  ) %>%
  distinct(website, period_id, .keep_all = TRUE)

# Get upper Y-axis limit for clean breaks
x_max <- max(ceiling(max(data_xyzal$price, na.rm = TRUE) / 200) * 200, 12000)


# Filter for Claritin
data_claritin <- data %>%
  filter(
    brand == "Claritin",
    form == "Tablet",
    size == 70,
    multipack == 1,
    flag_imputed_price != 1
```

```
  ) %>%
  distinct(website, period_id, .keep_all = TRUE)

x_max <- max(ceiling(max(data_claritin$price, na.rm = TRUE) / 200) * 200, 12000)

#1
p1 <- ggplot(data_xyzal, aes(x = period_id, y = price, color = website)) +
  geom_line(size = 0.9) +
  scale_color_manual(values = c("A" = "black", "B" = "#98bf64", "C" = "darkblue", "D" = "bro
  scale_x_continuous(limits = c(0, x_max), breaks = seq(0, x_max, by = 2400)) +
  labs(title = "Panel A. Xyzal, tablets, 80 count", x = "Hours Elapsed in Sample", y = "Pric
  theme_minimal(base_size = 14) +
  theme(legend.position = "bottom", legend.title = element_blank())
```

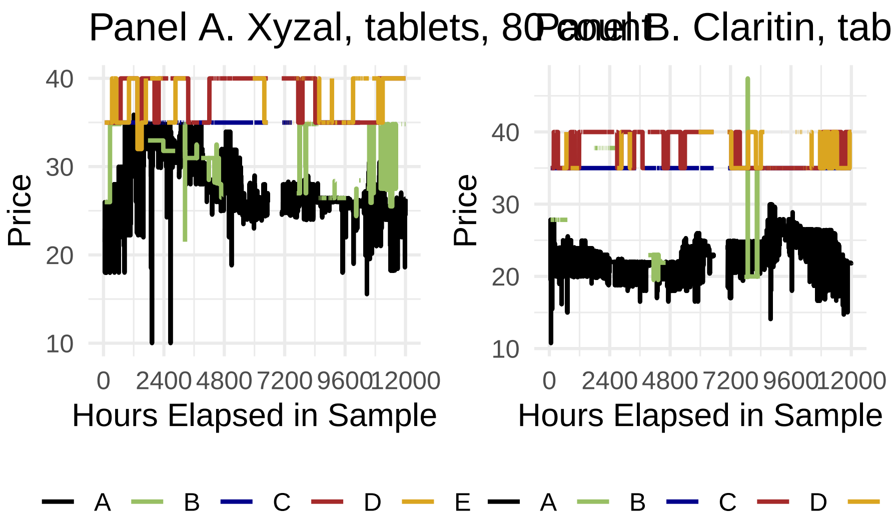
```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
# First plot
p2 <- ggplot(data_claritin, aes(x = period_id, y = price, color = website)) +
  geom_line(size = 0.9) +
  scale_color_manual(values = c("A" = "black", "B" = "#98bf64", "C" = "darkblue", "D" = "bro
  scale_x_continuous(limits = c(0, x_max), breaks = seq(0, x_max, by = 2400)) +
  labs(title = "Panel B. Claritin, tablets, 70 count", x = "Hours Elapsed in Sample", y = "P
  theme_minimal(base_size = 14) +
  theme(legend.position = "bottom", legend.title = element_blank())

# Display plots side by side
p1 + p2 + plot_layout(ncol = 2)
```

```
Warning: Removed 5714 rows containing missing values or values outside the scale range
(`geom_line()`).
```

```
Warning: Removed 8410 rows containing missing values or values outside the scale range
(`geom_line()`).
```



Original from ***The Effect of Gender on Credit Access*** (Brown & MacKay (2023)):

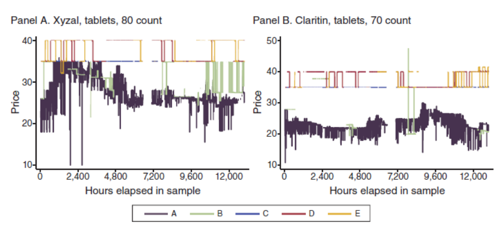


Figure 3: Example Time Series of Prices for Identical Products across Retailers

Replication:

```
Warning: Removed 5714 rows containing missing values or values outside the scale range
(`geom_line()`).
```

```
Warning: Removed 8410 rows containing missing values or values outside the scale range
(`geom_line()`).
```

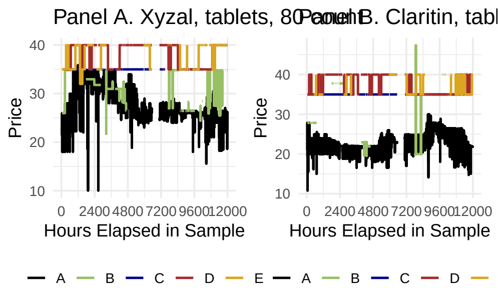


Table Replication

For our second replication, we followed the same logic as with the graph and used the data source ***analysis_data.dta*** and followed the script ***07_summary_stats.do*** (Open) to reproduce ***Table 1—Daily Statistics for Hourly Price Data***, found on page 117 of Brown & MacKay (2023).

We are currently facing the following key issues:

```
library(haven)
library(dplyr)
library(tidyr)
library(janitor)
library(gt)
library(tibble)

# Load and clean
df <- read_dta("C:/Users/Christian Casas/OneDrive - studhsf/Documents/Masters - HS Fresenius
  clean_names() %>%
  filter(!is.na(price), flag_imputed_price != 1)


# Ensure proper types
df <- df %>%
  mutate(
    date = as.Date(date),
    price_change = as.integer(price_change),
    is_observed = as.integer(is_observed)
  )

# Construct abs_price_change only where price_change occurred
df <- df %>%
  group_by(website, product_website_id) %>%
  arrange(date, .by_group = TRUE) %>%
  mutate(abs_price_change = if_else(price_change == 1, abs(price - lag(price)), NA_real_)) %
```

```
  ungroup()

# Collapse to daily product-level data
daily_df <- df %>%
  group_by(website, product_website_id, date) %>%
  summarise(
    n_price_change = sum(price_change, na.rm = TRUE),
    abs_price_change = sum(abs_price_change, na.rm = TRUE),
    observations = sum(is_observed, na.rm = TRUE),
    has_price_change = as.integer(any(price_change == 1)),
    price = mean(price, na.rm = TRUE),
    .groups = "drop"
  )

# Collapse to website-date level
summary_df <- daily_df %>%
  group_by(website, date) %>%
  summarise(
    n_products = n(),
    n_price_change = sum(n_price_change, na.rm = TRUE),
    abs_price_change = sum(abs_price_change, na.rm = TRUE),
    has_price_change = sum(has_price_change, na.rm = TRUE),
    observations = sum(observations, na.rm = TRUE),
    price_mean = mean(price, na.rm = TRUE),
    price_sd = sd(price, na.rm = TRUE),
    price_10 = quantile(price, 0.10, na.rm = TRUE),
    price_90 = quantile(price, 0.90, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  mutate(
    price_change_per_product = n_price_change / n_products,
    has_price_change_per_product = has_price_change / n_products,
    obs_per_product = observations / n_products,
    avg_abs_price_change = abs_price_change / n_price_change
  )

# Compute final summary stats per website
stats_by_website <- summary_df %>%
  group_by(website) %>%
  summarise(
    `Count of Products` = mean(n_products, na.rm = TRUE),
    `Observations per Product` = mean(obs_per_product, na.rm = TRUE),
    `Price: Mean` = mean(price_mean, na.rm = TRUE),
    `Price: 10th Percentile` = mean(price_10, na.rm = TRUE),
    `Price: 90th Percentile` = mean(price_90, na.rm = TRUE),
```

```r
    `Mean Absolute Price Change` = mean(avg_abs_price_change, na.rm = TRUE),
    `Price Changes per Product` = mean(price_change_per_product, na.rm = TRUE),
    `Share of Products with a Price Change` = mean(has_price_change_per_product, na.rm = TRU
    .groups = "drop"
  )

# Add "Total" row across all websites
total_row <- summary_df %>%
  summarise(
    website = "Total",
    `Count of Products` = mean(n_products, na.rm = TRUE),
    `Observations per Product` = mean(obs_per_product, na.rm = TRUE),
    `Price: Mean` = mean(price_mean, na.rm = TRUE),
    `Price: 10th Percentile` = mean(price_10, na.rm = TRUE),
    `Price: 90th Percentile` = mean(price_90, na.rm = TRUE),
    `Mean Absolute Price Change` = mean(avg_abs_price_change, na.rm = TRUE),
    `Price Changes per Product` = mean(price_change_per_product, na.rm = TRUE),
    `Share of Products with a Price Change` = mean(has_price_change_per_product, na.rm = TRU
  )

# Combine
final_stats <- bind_rows(stats_by_website, total_row)

# Round for display
final_stats_rounded <- final_stats %>%
  mutate(across(where(is.numeric), ~ round(.x, 2))) %>%

    column_to_rownames("website") %>%
  t() %>%
  as.data.frame()%>%
rownames_to_column("Statistic")


# Display
final_stats_rounded %>%
  gt() %>%
  tab_header(title = "Table 1 - Daily Statistics for Hourly Price Data")
```

Original from ***The Effect of Gender on Credit Access*** (Brown & MacKay (2023)):

Replication:

Table 1 — Daily Statistics for Hourly Price Data

Statistic	A	B	C	D	E	Total
Count of Products	132.98	43.37	53.19	45.03	38.26	62.61
Observations per Product	20.85	20.41	19.05	21.12	19.11	20.11
Price: Mean	27.18	16.88	17.63	20.92	21.74	20.86
Price: 10th Percentile	9.86	7.25	5.77	6.99	7.95	7.56
Price: 90th Percentile	50.48	27.39	32.82	37.96	38.88	37.47
Mean Absolute Price Change	1.35	2.31	1.12	3.28	3.06	1.91
Price Changes per Product	1.89	0.28	0.01	0.02	0.03	0.45
Share of Products with a Price Change	0.37	0.09	0.01	0.02	0.02	0.10

Table 1—Daily Statistics for Hourly Price Data

Statistic	Retailer *A*	Retailer *B*	Retailer *C*	Retailer *D*	Retailer *E*	All retailers
Count of products	124.9	41.3	49.9	42.5	35.1	58.7
Observations per product	20.9	20.4	19.0	21.1	19.1	20.1
Price: Mean	27.18	16.88	17.63	20.93	21.74	20.86
Price: 10th percentile of products	9.75	6.93	5.53	6.88	7.50	7.32
Price: 90th percentile of products	51.11	28.95	33.30	38.21	39.65	38.21
Mean absolute price change	1.35	2.31	1.12	3.28	3.06	1.91
Price changes per product	1.89	0.28	0.01	0.02	0.03	0.45
Share of products with a price change	0.373	0.089	0.008	0.020	0.024	0.103

Figure 4: Example Time Series of Prices for Identical Products across Retailers

Table 1 — Daily Statistics for Hourly Price Data

Statistic	A	B	C	D	E	Total
Count of Products	132.98	43.37	53.19	45.03	38.26	62.61
Observations per Product	20.85	20.41	19.05	21.12	19.11	20.11
Price: Mean	27.18	16.88	17.63	20.92	21.74	20.86
Price: 10th Percentile	9.86	7.25	5.77	6.99	7.95	7.56
Price: 90th Percentile	50.48	27.39	32.82	37.96	38.88	37.47
Mean Absolute Price Change	1.35	2.31	1.12	3.28	3.06	1.91
Price Changes per Product	1.89	0.28	0.01	0.02	0.03	0.45
Share of Products with a Price Change	0.37	0.09	0.01	0.02	0.02	0.10

Go Back to Index
Next Presentation: Tony

References

Board, S., & Meyer-ter-Vehn, M. (2022). A reputational theory of firm dynamics. *American Economic Journal: Microeconomics*, *14*(2), 44–80. https://doi.org/10.1257/mic.20190376

Brown, Z. Y., & MacKay, A. (2023). Competition in pricing algorithms. *American Economic Journal: Microeconomics*, *15*(2), 109–156. https://doi.org/10.1257/mic.20210158