The impact of sustainability programs on consumer purchase behavior: Evidence from Amazon

Davide Proserpio* Amazon & USC

Ali Goli† Amazon

Tyler Mangini‡ Amazon

Ken Lau§ Amazon

Daniela Yu¶ Amazon

April 2025

*Email: dproserp@amazon.com. Davide Proserpio is an associate professor of marketing at the University of Southern California. He currently works for Amazon as an Amazon Scholar.

†Email: agoli@amazon.com

‡Email: tmangini@amazon.com

§Email: lauuken@amazon.com

¶Email: yudn@amazon.com

Abstract

In 2020, Amazon launched the Climate Pledge Friendly (CPF) program to make it easy for customers to discover and shop for products with sustainability certifications. In this paper, we measure the causal impact of products qualifying for CPF on consumer purchase behavior. Using a dataset of about 45,000 products spanning three categories, and a Difference-in-Differences identification strategy, we show that joining CPF leads to a 13.3% increase in Gross Merchandise Sales, 12.5% increase in sales, and 4.4% increase in shipped units in the next 12 weeks after adoption. These results are robust to different estimators, two identification strategies, longer time windows (24 and 48 weeks), and controlling for price, promotions, ad spend, ratings and reviews, and search rank. Importantly, we show that not accounting for these controls inflates estimates by 19-27%. These differences highlight the importance of controlling for concurrent marketing efforts when evaluating sustainability programs, as sellers often increase support for certified products. We conclude the paper by showing that products with low visibility as measured by product page views, and consumable products benefit the most from the program.

1 Introduction

Environmental and social responsibility have become increasingly important factors for consumers deciding where and how they shop. Recent large-scale studies document this trend: McKinsey and NielsenIQ (2023) found that 78% of US consumers consider sustainable lifestyle important, while Simon-Kucher & Partners (2022) showed 71% of consumers worldwide factor sustainability into purchasing decisions. These preferences vary meaningfully across demographics and markets, with GlobeScan (2022) documenting substantial variations in sustainability priorities across 31 markets. In 2024, the share of customers who felt it was important to buy from socially responsible companies reached an all-time high of 66%.[1] The growing emphasis on corporate responsibility not only impacts where customers shop, but also what they choose to buy, and 80% indicate they would be willing to pay more for sustainable products, also an all-time high in 2024.[2] As demand grows, retailers and online stores are launching programs intended to help customers discover products and services that match their values.

Launched in September 2020, Amazon's Climate Pledge Friendly (CPF) program highlights products with sustainability features backed by at least one of the 54 sustainability certifications part of the program.[3] The program has grown to include 1.4M products, which customers discover using badges, filters, recommendations, and a dedicated storefront. Whether programs like this benefit consumers depends on comprehension and trust (Delmas and Grant, 2014), cultural and political beliefs (Aneja et al., 2023; Kim and Liu, 2023), and willingness to pay. While consumers claim to prioritize sustainability in surveys and research studies, there is ongoing debate about the extent to which these stated preferences translate into real-world purchase decisions, a phenomenon called the "value-action gap" (Essiz et al., 2023).

[1] https://goodmustgrow.com/cms/resources/ccsi/gmg2023ccsifinal.pdf

[2] https://www.firstinsight.com/white-papers-posts/consumer-expectations-for-sustainable-retail-compendium

[3] https://www.amazon.com/b?node=21221608011

To measure how customer demand for sustainability translates to purchase behavior, this paper examines the impact of qualifying for Amazon's CPF program on product sales and volume using two datasets. The first (and main dataset for our analyses) includes about 45,000 U.S. products for which we observe weekly economic outcomes for a period of about three years, from July 2021 to January 2024. The second dataset includes another 45,000 products from five European countries (the United Kingdom, Germany, France, Italy, and Spain) for which we observe economics outcomes over the same time period.

To measure the causal impact of qualifying for CPF on purchase behavior, we use the staggered adoption of the program across the products we observe. We implement a Difference-in-Differences strategy that compares outcomes before and after products join CPF with a baseline of changes in outcomes for products that never joined or had not-yet-joined CPF over the same time period. Using this strategy, we find that the CPF program positively impacts consumer purchase behavior in both the U.S. and European regions. We quantify consumer purchase behavior using three outcomes: (1) Gross Merchandise Sales (GMS) which are computed using the following formula: GMS = Product Price $\times$ Quantity + Shipping + Giftwrap − Returns (Customer Accommodations) − Promotional Discounts, (2) Ordered Product Sales (OPS) to which we also refer to as sales, i.e., gross sales revenue including any governmental taxes, and (3) Net Shipped Units or volume of products sold. Using the U.S. dataset and our preferred specification that accounts for ad spend, price, promotions, and review ratings, we find a 13.3% increase in GMS, a 12.5% increase in OPS, and a 4.4% increase in Net Shipped Units, in the 12 weeks after joining the program. These results hold when considering longer time windows (24 or 48 weeks before and after adoption) and are robust to two staggered DD estimators (Two-Way Fixed Effect and Gardner (2022)'s two-stage approach). In addition, we replicate the U.S. findings for products sold in European countries where we find a 16.8% increase in GMS, a 15.5% increase in OPS, and a 4.2% increase in Net Shipped Units. Moreover, we show that our results are not driven by changes in search ranks potentially affected by Amazon artificially increasing the visibility of CPF products. Finally,

to reinforce the causal interpretation of our results, and reduce concerns about differences between CPF and non-CPF products driving our estimates, we implement three additional checks. First, we implement an alternative identification strategy that benefits from the fact that a product can be sold in multiple countries, and can be part of the CPF program in one country but not the other. We, therefore, follow a similar approach to that discussed in Chevalier et al. (2018) and Proserpio and Zervas (2017), and implement a cross-country Difference-in-Differences strategy. This strategy compares changes in performance for the same product across two countries, one where the product is part of the CPF program and one where it is not. In doing so, we allay concerns related to the effect we estimate being driven by differences between treated and control units since, in this case, they are exactly the same. Second, we show that our results hold if we limit our dataset to only eventually treated products, i.e, products that eventually become CPF. Third, we show that our results hold using a more balanced dataset of matched products using propensity score matching.

We conclude the analyses by studying the heterogeneous impact of the CPF program focusing on two product dimensions: visibility and category. We find that the CPF program has a higher impact on sales and volume for less visible products suggesting that the program helps consumers discover products they would not have otherwise purchased. We also find that consumable products—products that are intended to be used and replaced over a relatively short period of time such as food and personal care items—benefit the most.

Overall, these results indicate that consumers' self-reported preferences for more sustainable products can translate to purchase decisions when retailers offer programs—like Amazon's Climate Pledge Friendly—that enable shoppers to easily discover and shop for products that qualify for sustainability certifications.

2 Literature review

Programs like Amazon's Climate Pledge Friendly (CPF) help consumers identify more sustainable products, but their effectiveness can vary based on consumer values and whether these stated sustainability preferences actually translate into purchase decisions. More broadly, product labeling has emerged as a strategy to influence consumer choices across various domains like health and sustainability. By providing simplified cues about desirable or undesirable product attributes, labels have the potential to aid consumers in making better decisions aligned with their values and preferences. However, there is an ongoing debate around the true effectiveness of such voluntary labeling initiatives.

In the nutritional domain, research on mandatory labeling schemes has found such labels can significantly impact choices like reducing energy and fat intake (Shangguan et al., 2019), though the magnitude varies across categories (Dubois et al., 2021) and label designs (Ikonen et al., 2020). Front-of-pack nutritional warning labels that explicitly flag unhealthy products, which are mandatory in nature, have proven particularly effective in shifting demand toward healthier options (Araya et al., 2022).

In the context of voluntary labeling, Rao and Ursu (2024) examined the impact of voluntary nutrition labels on consumer demand. Brecko and Kim (2024) analyzed firm and consumer decisions around products with sustainability features in health and beauty categories. Their findings suggest sustainability occupies a unique position – while consumers exhibit increasing preference for sustainability, they still prioritize other non-sustainability product attributes. As a result, brands face varying incentives, with smaller brands using sustainability for strategic differentiation while large brands have limited motivation to offer sustainable products.

Research has shown that the effectiveness of such programs relies on consumer awareness and trust. Delmas and Grant (2014) show how labeling a product with a third-party certification of environmentally friendly practices (to which the authors refer to as eco-labels) affects products' price premiums in the wine industry. The authors find that consumers are

not willing to pay a premium for wine eco-labels but that certified but unlabeled wine enjoys a significant premium. The authors argue that this is due to the lack of understanding and trust in the eco-labels. Indeed, past research highlighted that it is important that firms, policy-makers and accreditation organizations educate consumers about eco-labels because in doing so they can increase pro-environmental consumer behavior (Taufique et al., 2017).

Similarly, Borin et al. (2011) show that it is important that these labels provide clear explanations of the environmental benefits in order to affect consumer purchase behavior and perception of product quality and value. The effectiveness of sustainability messaging can also depend on customer values, as discussed by Buerke et al. (2017). The results suggest that investing in sustainable practices can influence purchase behavior, especially for consumers who value sustainability.

Counter to the findings on customers' stated preferences, Essiz et al. (2023) study the drivers of the so called "green gap", i.e., consumers generally hold favorable values toward green consumption, but they often struggle to translate these values into actual purchase decisions. The authors show that risk-aversion can affect conversion, and that female consumers have greater consistency between their values and purchase behavior. The gap between stated preferences and real purchase decisions highlights that programs like Amazon's Climate Pledge Friendly have the potential to impact consumer purchase behavior, but their success is contingent upon consumer awareness, clarity of the program, perceived authenticity of the sustainability claims, and consumers' intrinsic risk aversion. We add to this literature—which mostly relies on surveys–by empirically studying the effectiveness of CPF on shifting customer shopping behavior.

The introduction of new labeling policies, whether voluntary or mandatory, is often concomitant with firms adjusting other marketing mix variables like pricing or promotions. As such, it is important to account for these potential supply-side responses when assessing the full impact of labeling regulations. For instance, some studies suggest that even with mandatory nutritional labeling policies, manufacturers may adjust pricing in equilibrium, poten-

tially raising prices of labeled unhealthy products or unlabeled healthier alternatives (Pachali et al., 2023).

Closely related to our research, Feng et al. (2024) and Wang and He (2024) also study the impact of Amazon's CPF program, but rely on sales rank data from Keepa rather than actual sales figures as we do. Both papers find that CPF leads to higher sales ranks. In addition, Feng et al. (2024) show that the effect is larger for products targeting older individuals and males, while Wang and He (2024) show that the effect is larger for hedonic goods compared to utilitarian goods, and for lower-reviewed products compared to higher-reviewed products. However, an important limitation of sales rank is that changes in rank cannot be easily translated into economic impact. For instance, a large improvement in sales rank among low-selling items (e.g., moving from rank 100,000 to 50,000) might represent just a few additional units sold, while a small rank improvement among top sellers (e.g., moving from rank 20 to 10) could represent thousands of units sold. Moreover, sales ranks are category-specific and depend on factors such as category concentration, catalog size, and sales distribution within each category. For example, a rank of 1,000 in Electronics might represent 500 monthly units sold because it is a category with high concentration and high unit sales, while the same rank in Home & Kitchen could represent 50 units sold due to its larger catalog size and more dispersed sales distribution. This makes it difficult to compare cross-category treatment effects.

An additional limitation of these papers is their inability to observe and account for concurrent marketing activities like advertising spend and price promotions by sellers on the Amazon platform. By accessing proprietary data, we can observe and account for all marketing activities including ad spend and promotions performed by sellers on Amazon, thereby increasing our ability to isolate and precisely measure the impact of CPF. This is particularly important in our setting, since sellers decide when to join the program and the timing may coincide with changes in marketing activities. Comparing our estimates with and without marketing controls reveals that not controlling for such variables can lead to inflated

estimates of the treatment effect on sales metrics, with differences ranging from 19-20% for GMS and sales to 27% for shipped units. Therefore, an important contribution of our paper is to provide estimates that are likely less biased that those reported in this papers simply because we observe actual sales and sellers actions that can affect them.

Finally, Chu et al. (2025) also study CPF, but focus on the cost-benefits analysis of participating in the program for one large Amazon sellers part of the consumer electronics industry. The authors find that while CPF increases product visibility, the extra sales that the program generates are not enough to offset the certification costs for the specific seller they study.

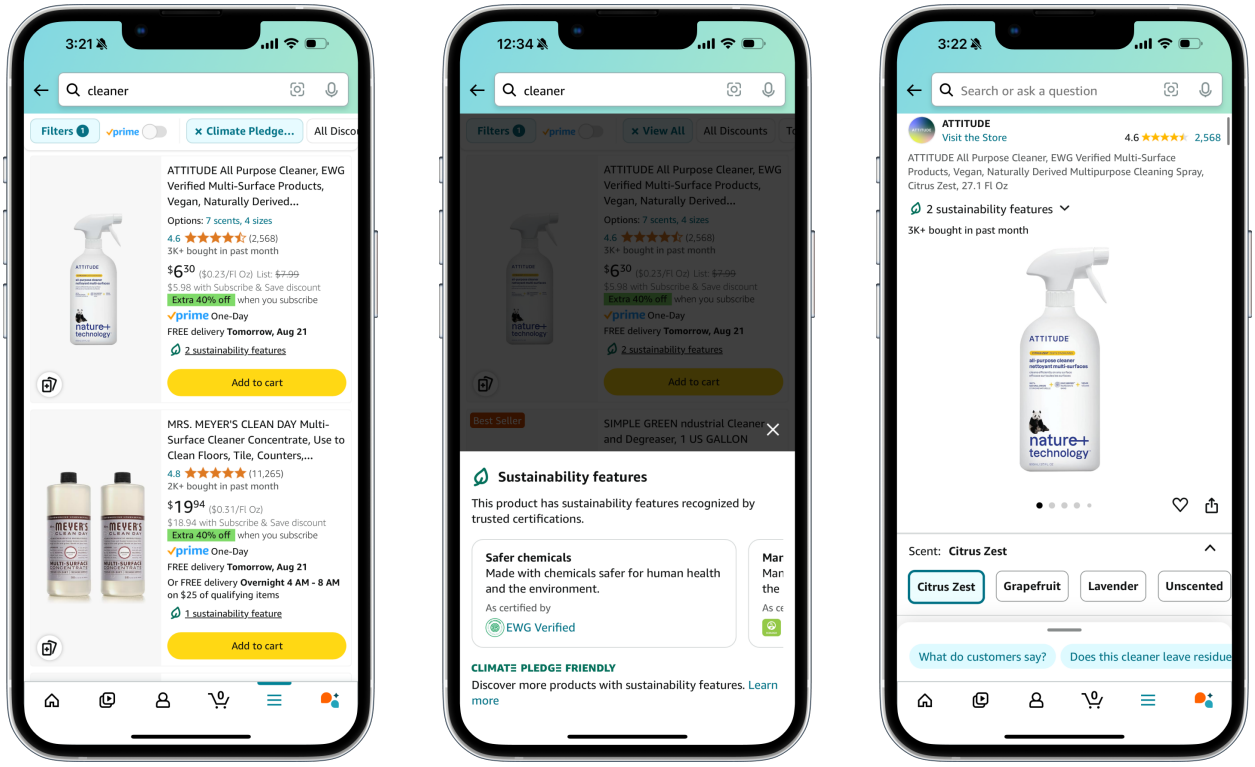
3 Empirical context and data

Launched in September 2020, Climate Pledge Friendly (CPF) is an Amazon program that helps customers discover products with sustainability features. To qualify for CPF, a product must be accredited by at least one of 54 third-party certifications or Amazon's own certifications (Compact by Design or Pre-owned Certified)[4]. Qualified products receive a badge that appears throughout a customer's shopping journey with a green leaf icon and the number of sustainability features a product qualifies for (see Figure 1). Consumers can click on the badge to get additional information about the sustainability feature(s), and the certifications that substantiate the claim. Beyond badges, customers can discover CPF products using a search filter, product recommendations, or a dedicated storefront (see Figure 2). CPF's onsite experience helps customers quickly identify products with sustainability features and verify the claims are backed by robust certifications. CPF's features are intended to make it easy for Amazon customers to make purchase decisions that align with their values.[5]

[4]See: `https://www.amazon.com/b?node=21221609011&ref_=a20m_us_spcs_cpf` and `https://www.amazon.com/s/browse/?node=23911980011&ref_=a20m_us_spcs_cpf`

[5]These images are representative of the current US customer experience and are subject to change.

Figure 1: From left to right, example of the CPF badge in search results, information available on click, and the CPF badge on the product page.

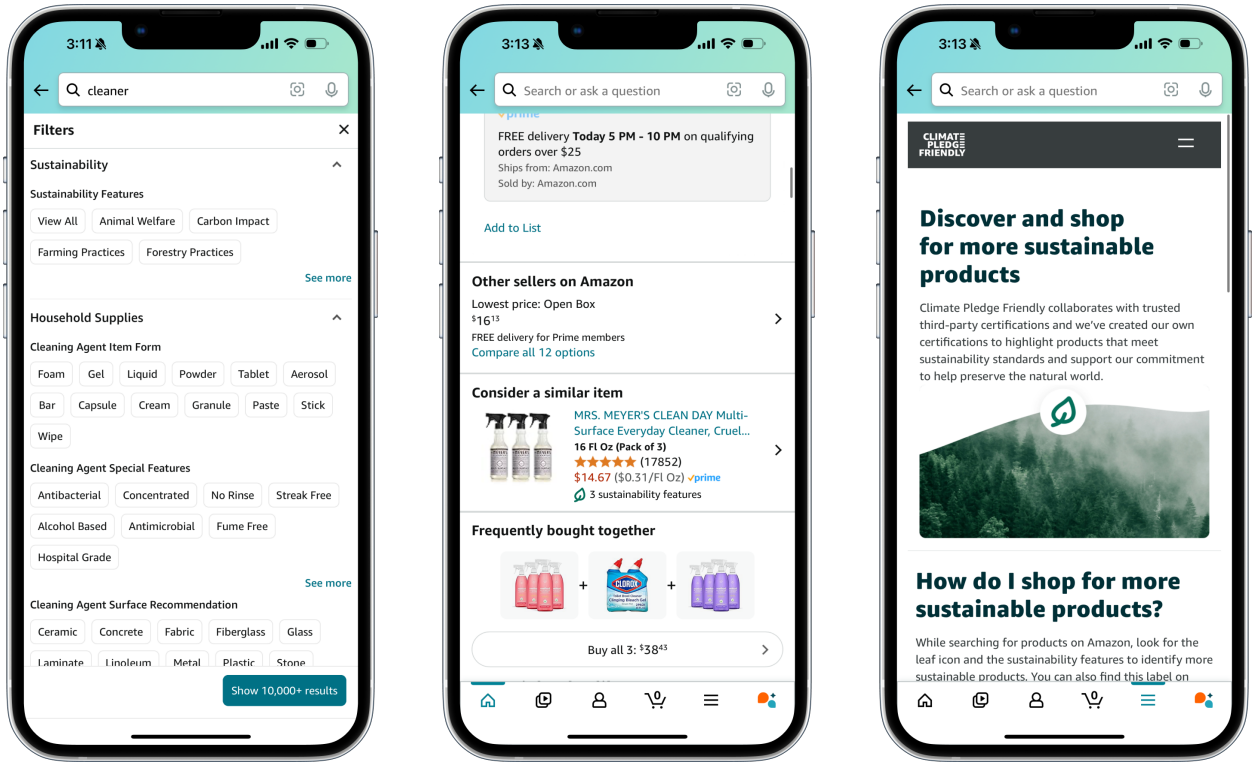


3.1 Data

U.S. products All our analyses are performed on Amazon proprietary data. Our main dataset consists of a sample of about 55,000 U.S. products. 50,000 of them were randomly sampled from the population of products that joined the CPF program during our observation period, and 5,000 of them were randomly sampled from the population of products which were never part of the CPF program. We observe these products for a period covering about three years, from July 2021 to January 2024. For each product, we collected information about its product category, ratings and reviews, price, price discounts (if any), and advertising expenditures for several channels.[6] In addition, we collect information about three measures of consumer purchase behavior: (1) Gross Merchandise Sales (GMS) (computed using the following formula: GMS = Product Price × Quantity + Shipping + Giftwrap −

[6]The ad channels are: Display, Video and Audio (DVA), Over the Top TV (OTT), Sponsored Product (SP), Sponsored Video Brands (SVB), Sponsored Brand (SB), and Sponsored Display (SD).

Figure 2: From left to right, an example of a CPF filter, recommendation, and dedicated storefront.



Returns (Customer Accommodations) − Promotional Discounts), (2) Order Product Sales (OPS) which we also refer to as sales, i.e., Gross Sales Revenue including any governmental taxes, and (3) Net Shipped Units (volume of units sold).

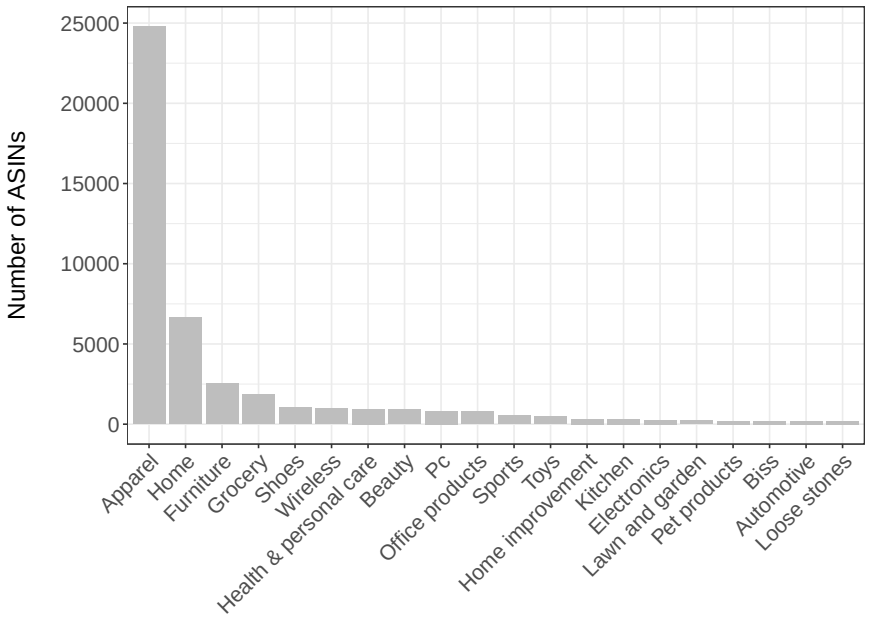
We aggregate the data at the product-week level. In addition, for all of the analyses reported in this paper, we focus on products that qualified for CPF and were never removed from the program, and for which we could match them with sales outcomes and the variables described above. Moreover, we removed products with an unknown category leaving us with 45,361 products, which are part of three categories, consumables, hardlines, and softlines, with softlines and hardlines being the largest categories with 26,302 and 15,073 products, respectively.[7] Softlines products make up the majority due to 24,800 Apparel products,

[7]Consumables are products that are regularly consumed and replenished. Examples include beauty, health and personal care, cleaning supplies, and grocery. Hardlines are durable goods typically made of rigid materials like plastic, metal, and wood. Examples include small appliances, hardware, automotive parts, sporting goods, and toys. Softlines are products related to fashion including apparel and accessories. Examples include clothing, footwear, handbags, and luggage.

which often have multiple variations and colors. 36,046 of these products eventually joined the CPF program while the rest never did.[8]

In Figure 3, we plot the number of products by product group for the top-20 groups. Most of the products in our dataset are part of Apparel, Home, and Furniture, with a long tail of products in categories such as Grocery, Shoes, or Electronics.

Figure 3: Number of products by category for the U.S. dataset



European products In addition to U.S. products, we replicate some of the analyses performed on the U.S. dataset using a dataset of products from five European countries: the United Kingdom, Germany, France, Italy, and Spain. The procedure to create the dataset is similar to that used to generate the U.S. dataset. The final EU dataset contains 44,835 products,of which 40,821 joined the CPF program at some point and 4,020 never joined it. We observe outcomes for these products over the same time period as the U.S. We plot the

[8]Note that there are more than 5,000 products that we consider as never joined the program because for some products the period for which we observed sales outcomes did not overlap with the period during which the product was part of the CPF program. This often occurred for products that we observed being part of the program for less than three days. We obtain consistent results if we drop these products rather than consider them controls.

number of products by product group for the top-20 groups in Figure 8 in Appendix A.

4 Empirical strategy

The main issue in estimating the impact of the CPF program on product performance is that brands self-select products to certify, which can be problematic if brands decide to qualify products for CPF at a specific point in time (e.g., when sales are decreasing) or if the adoption is part of a larger marketing strategy that can affect other marketing tactics such as advertising spend. In both cases, we may measure a positive effect associated with CPF, when in reality lower pre-CPF sales and advertising were the drivers of the effect.

We try to address this self-selection issue in several ways. Our first approach uses the staggered adoption of CPF among our products. This variation allows us to implement a Difference-in-Differences (DD) identification strategy with staggered treatment adoption (De Chaisemartin and d'Haultfoeuille, 2023; Goodman-Bacon, 2021). Using this identification strategy, we measure changes in outcomes before and after joining CPF with respect to changes in outcomes for products that did not (nevertreated) or did not yet adopt (not-yet-treated) CPF over the same period. Allowing not-yet-treated (in addition to nevertreated) to be part of the control products reduces concerns about differences between treated and nevertreated products that can drive the results. Moreover, the staggered adoption of the CPF program allows us to rule out common shocks to CPF products that affected their performance right after adoption (through time fixed effects). To further isolate the impact of CPF, we focus on a period of 12 weeks before and after qualifying for CPF to limit the possibility of confounders affecting our estimates (but show that results hold when considering longer time windows). Finally, given the recent research that demonstrated that staggered DD estimates may be biased in the presence of heterogeneity (Goodman-Bacon, 2021), we show that estimates are consistent when using the classic Two-Way Fixed Effect estimator and the recent two steps approach proposed by Gardner (2022).

In addition to relying on a staggered DD identification strategy, we account for several factors affecting product performance. A benefit of our proprietary dataset is that we observe concomitant marketing activities implemented on Amazon, allowing us to control for usually unobserved variables such as advertising spend and promotions (in addition to the usual controls such as price, average ratings, and number of reviews). To further validate our results, we show that they hold when we use a dataset of products from five European countries. While it is possible that the same confounder or set of confounders affects in a similar way products in the same country, it is less likely (but not impossible) that the same confounder is tied to six different countries. Moreover, we show that our results are not driven by changes in search ranks potentially caused by Amazon preferencing CPF products. Finally, we perform three additional checks to reduce concerns about our results being driven by differences between CPF and non-CPF products First, we benefit from the fact that the same products are listed on Amazon across different countries and a product may be part of the CPF program in one country but not the other. In doing so, and similarly to several papers including Proserpio and Zervas (2017) and Chevalier et al. (2018), we implement a cross-country DD. This strategy compares outcomes for a product part of the CPF program, before and after the joining it, with outcomes of the same product in a different country and not part of the CPF program. This strategy effectively eliminates concerns about differences between treated and control units driving the results.[9] Second, we show that our results hold if we limit our dataset to only eventually treated products, i.e, products that eventually become CPF. Third, we perform matching using propensity score via a logistic regression to create a more balanced panel of products.

We describe these analyses and results in detail next.

[9]Note that while the Cross-country DD is helpful and provides additional evidence supporting our main findings, the estimated effects are not ATT but LATE (Local ATE). This means that what we measure is the effects of CPF products that comply with the selection criteria (i.e., they are sold in both countries and are CPF in the UK but not Germany); these products may be different from the average CPF products and therefore results may differ. This is the reason why this is not our main identification strategy.

5 Staggered DD

The main identifying assumption behind any DD is the parallel trends assumption, i.e., that treated and control units' outcomes would have evolved in the same way in the absence of the treatment. Since this assumption is untestable (we do not observe the counterfactual outcome for treated units had they not been treated) researchers rely on checking pre-treatment trends to partially verify this assumption. In addition the treatment should be exogenous or quasi-exogenous, i.e., uncorrelated with the error term. We provide evidence in support of the parallel trends assumption in Section 5.1. To reduce concerns about the treatment being endogenous, in Section 6, we show that our estimates are robust to the inclusions of a large number of controls. Moreover, to try to isolate the effect of the CPF program, we limit our main analysis to a period of 12 weeks before and after the treatment to reduce the likelihood of advertisers implementing other non-observed actions that can affect the outcomes we study.[10]

This strategy is usually implemented using a Two-Way Fixed Effect (TWFE)—unit (in our case product) and time (in our case year-week)—OLS regression. However, recent research has shown that if the treatment is heterogeneous between groups (in our case, groups of products joining the CPF program at the same time) and over time, TWFE may generate biased estimates (De Chaisemartin and d'Haultfoeuille, 2023). To avoid this issue, econometricians have developed several new estimators for staggered DD that are robust to treatment heterogeneity (Borusyak et al., 2021; Callaway and Sant'Anna, 2021; Gardner, 2022). While these approaches differ slightly in their assumptions (formulation of the parallel trends assumptions, types of treatment effects possible to identify, and data or treatment conditions necessary for identification and estimation), all of them produce consistent and unbiased results in the presence of treatment heterogeneity.

We use TWFE as our main model because it scales well to millions of observations and

[10] In Section 6, we show that our results hold for longer windows of 24 and 48 weeks, or the complete dataset.

allows us to easily include controls. To test whether treatment heterogeneity across groups is affecting our estimates, we rely on the two-stage estimator proposed by Gardner (2022), which is straightforward to implement and also scales relatively well. The key identifying assumption of this approach is that the counterfactual outcome $Y_i(d, 0)$ can be characterized by $Y_i(d, 0) = g_i + \tau_t$, i.e., an additive function of a group (of products) fixed effects, g_i, which are defined based on the time period in which the products joined the CPF program, and a time trend, τ_t. This additivity assumption seems reasonable in our case, as the group fixed effects capture time-invariant differences across groups joining at different times, while time fixed effects account for seasonality or common demand shocks across products.

The TWFE model takes the following form:

$$Y_{it} = \beta_1 \text{After}_{it} + \text{X}_{it}^{'}\gamma + \alpha_i + \tau_t + \epsilon_{it}, \tag{1}$$

where i and t index products and time (weeks), respectively. Y_{it} is either the logarithm of GMS, sales, or net shipped units (to deal with zeros with add one to these quantities) of product i at year-week t. After_{it}, the coefficient of interest, is a binary indicator which is one after product i joined the CPF program, and zero otherwise. $X_{it}^{'}$ is a vector of controls which includes total ad spend on Amazon, price, discounts, average ratings, and number of reviews. α_i and τ_t are product and year-week fixed effects, respectively; product fixed effects account for time invariant product characteristics that can affect sales (e.g., some products are sold by brand names vs. generic/unknown brands), while year-week fixed effects account for time varying shocks to the outcome that are common to all products (e.g., holidays and seasonality). We estimate this model using OLS and cluster standard errors at the product level (Bertrand et al., 2004; Abadie et al., 2023).

Gardner (2022)'s two steps approach relies on the following two stages:

$$\begin{aligned} &(1) \quad Y_{it} = \alpha_{g_i} + \tau_t + \epsilon_{it} \quad \rightarrow \quad \hat{Y}_{it} = Y_{it} - \hat{\alpha}_{g_i} - \hat{\tau}_t, \\ &(2) \quad \hat{Y}_{it} = \beta \text{After}_{it} + \nu_{it}. \end{aligned} \tag{2}$$

In the first step, we estimate the first equation with time and group fixed effects for the set of non-treated observations. For instance in period t, we would be using all groups of products (g) that are not treated up until period t in this specification. The first stage estimates the cross-sectional group fixed effects α_{g_i} and time fixed effects τ_t. In the second stage, we calculate the adjusted outcome $\hat{Y}_{it}$ by subtracting the corresponding group and time fixed effects estimated in the first stage, and then regress it on the treatment dummy.[11]

5.1 Staggered DD results

Event study As is usual in DD analyses, we start by presenting the event study results, which allows us to compare treatment effects between treated and control units before and after the treatment at a weekly cadence. The goal of this analysis is two-fold. First, it allows us to partially verify the parallel trends assumption, i.e., whether before the treatment, treated and control groups behaved similarly in terms of outcomes. Second, the event study allows us to visualize the evolution of the treatment in the post-treatment period.

To estimate the evolution of the treatment effect using TWFE, we use the following specification:

$$Y_{it} = \sum_{k \in [-12,12]} \beta_k \mathbb{1}_{\{t - t_i^* = k\}} + \alpha_i + \tau_t + \epsilon_{it}, \tag{3}$$

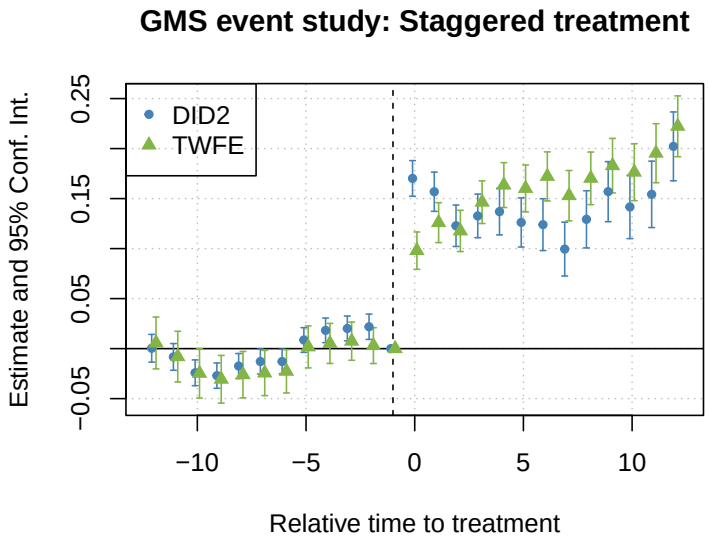
where t_i^* denotes the week where product i joins the CPF program. $\mathbb{1}_{\{t-t_i^*=k\}}$ is a dummy that is one if time period t is k weeks away relative to the time of adoption t_i^*; we refer to

[11] As Gardner (2022) suggests, one way to add controls is to add them to both stages.

this as Interval k. So, for example, Interval 0 is the treatment week, Interval -1 is the week prior to the treatment week, and Interval 1 and is the week right after the treatment week. As before, α_i and τ_t are product and week fixed effects, respectively. As it usually done in event study designs, we estimate Equation 3 setting the Interval -1 as the baseline.

To estimate the event study in the case of Gardner (2022)'s approach, we replace the treatment dummy $After_{it}$ in the second stage with the set of intervals defined in Equation 3.

Figure 4: GMS event study estimates using TWFE and Gardner (2022)'s two-stage approach (DID2). The black vertical dashed line represent the period before joining the CPF program.



We estimate the event study both with TWFE and Gardner (2022), 12 weeks before and after the treatment. We plot the estimated β_k of Equation 3 for the three outcome of interest in Figures 4, 5, 6. The dashed vertical black line identifies the period just before the treatment starts (-1); we plot in green the estimates of the TWFE model and in blue the estimates of Gardner (2022) (DID2 in the figures' legend). There are a few things worth pointing out about these figures. First, in the pre-treatment period ([-12,-1]) we observe estimates that are close to zero in all three plots, suggesting that the parallel trends assumption is satisfied. Second, in the post treatment period ([0,12]), we observe a substantial positive increase in the estimates that coincides with the beginning of the treatment. Third, TWFE and Gardner

Figure 5: Sales event study estimates using TWFE and Gardner (2022)'s two-stage approach (DID2). The black vertical dashed line represent the period before joining the CPF program.

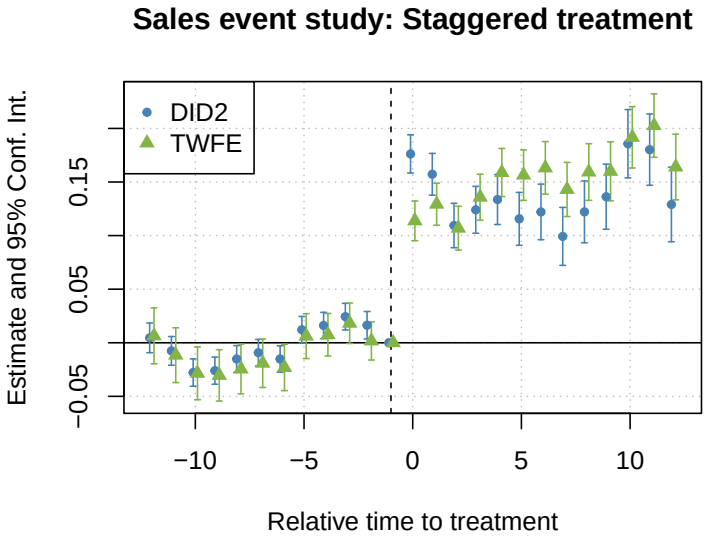
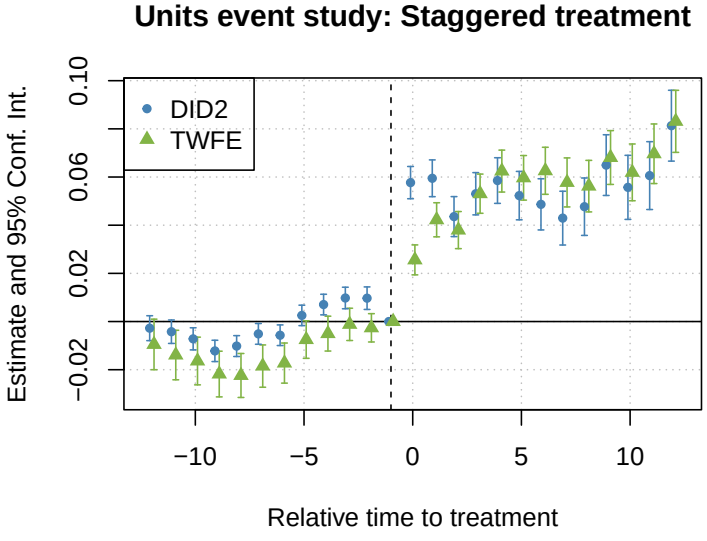


Figure 6: Net Shipped Units event study estimates using TWFE and Gardner (2022)'s two-stage approach (DID2). The black vertical dashed line represent the period before joining the CPF program.



(2022) follows a strikingly similar trend suggesting that treatment heterogeneity is not an issue in our case.

Results We start by presenting estimates without including controls. We report the overall Average Treatment Effects on the Treated (ATT) in Table 1 and Table 2 for TWFE and Gardner (2022), respectively. Consistent with the event studies, we observe that the CPF program has a positive effect on the three outcomes. Considering the estimates for TWFE reported in Table 1, joining the CPF program leads to approximately 16.4% increase in GMS, 15.7% increase in sales, and 6% increase in net shipped units. The estimates obtained with Gardner (2022)'s estimator are very similar to that of the TWFE estimator.

Given the consistent results we obtained with the two estimators, in the rest of the analyses we report estimates using the TWFE estimator and report Gardner (2022)'s results in Appendix B.

Table 1: DD estimates using TWFE

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.152***	0.146***	0.059***
	(0.008)	(0.009)	(0.004)
Observations	1,343,861	1,365,441	1,365,441
R^2	0.657	0.663	0.777

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates obtained using TWFE. All models include product and year-week fixed effects. Standard errors reported in parentheses are clustered at the product level.

6 Sensitivity and robustness checks

In this section, we discuss several sensitivity analyses and robustness checks aimed at reinforcing the causal interpretation of our results.

Table 2: DD estimates using Gardner (2022)

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.142***	0.137***	0.056***
	(0.010)	(0.010)	(0.004)
Observations	1,315,493	1,336,721	1,336,721
R^2	0.002	0.002	0.002

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates obtained using Gardner (2022)'s two-stage approach. All models include product and year-week fixed effects. Standard errors reported in parentheses are clustered at the product level.

6.1 Adding controls

A plausible concern with the results presented above is that brands whose products join the CPF program are also performing additional actions to drive these products' sales. For example, it is possible that brands change their advertising strategy for products part of the CPF program (e.g., increase their spend to better promote them) and this change drive the results we discussed above. It could also be that advertising works better for CPF products and that CPF products get better reviews. Both these effects should increase sales of CPF products. Here we attempt to estimate the *net benefit* of the CPF program by accounting for other factors that drive sales or that may be affected by joining the CPF program. It is worth noting that if the average brand that joins the CPF program also changes their marketing strategy *because* of the CPF program, retailers and brands alike may be more interested in the effect discussed in Section 5.1. This is because this quantity captures the effect on sales for the average brand in out dataset that decides to join the CPF program. However, we believe there is also value in trying to understand how the CPF program alone affects consumer purchase behavior, above and beyond any other marketing activity implemented.

In this section, we show that the estimates discussed above are robust to the inclusion of

Table 3: DD estimates using TWFE and controls

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.125***	0.118***	0.043***
	(0.008)	(0.008)	(0.003)
log Total Ad Spend	0.622***	0.670***	0.279***
	(0.007)	(0.008)	(0.004)
log Price	−0.200***	−0.164***	−0.270***
	(0.033)	(0.032)	(0.017)
log Promo Amount	0.938***	1.151***	0.442***
	(0.011)	(0.011)	(0.006)
Is Reviewed	0.430***	0.409***	0.226***
	(0.072)	(0.071)	(0.030)
Is Reviewed × Cum Avg Ratings	−0.010	−0.011	−0.013**
	(0.014)	(0.014)	(0.006)
Is Reviewed × log Cum Wkly Reviews	−0.278***	−0.305***	−0.079***
	(0.028)	(0.028)	(0.014)
Observations	1,221,424	1,242,984	1,242,984
R^2	0.687	0.704	0.813

Note: *p<0.1; **p<0.05; ***p<0.01

Estimates obtained using TWFE. All models include product and year-week fixed effects. Standard errors reported in parentheses are clustered at the product level.

a wide set of controls that are likely correlated with the outcomes we study. In particular, we account for: (1) the logarithm of total advertising spend computed as the sum of ad spend for the channels discussed in Section 3, (2) the logarithm of the product price and any discounts applied to it; (3) cumulative average ratings which captures products' consumer perceived quality; (4) cumulative number of reviews.[12] We present these estimates using TWFE in Table 3 and using Gardner (2022) in Table 14 in Appendix B. We observe that the estimates decrease compared to those without controls. As discussed above, this is not unexpected, since Amazon's selling partners decide when to join the program, and may decide to change marketing strategy (ad spend, price, and promotions) to support these products. However, the effect of CPF is still large and significant. Using the estimates from Table 3, joining the CPF program leads to approximately 13.3% increase in GMS, 12.5% increase in OPS, and 4.4% increase in net shipped units.

Table 4: DD estimates using TWFE over different time windows

	BW = 24			*BW = 48*		
	log GMS	log Sales	log Units	log GMS	log Sales	log Units
	(1)	(2)	(3)	(4)	(5)	(6)
After	0.131***	0.125***	0.041***	0.120***	0.113***	0.032***
	(0.008)	(0.008)	(0.004)	(0.009)	(0.009)	(0.004)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,733,224	1,764,609	1,764,609	2,445,823	2,491,398	2,491,398
R^2	0.684	0.701	0.811	0.675	0.692	0.807

Note: *p<0.1; **p<0.05; ***p<0.01

Estimates obtained using TWFE. All models include product and year-week fixed effects, and the list of controls described in Section 6. Standard errors reported in parentheses are clustered at the product level.

[12]Since a product can sell and have no reviews and not all products have reviews, similar to Zervas et al. (2017), we add a dummy indicator for whether a product is reviewed and interact it with these two review-related variables.

6.2 Longer time horizons

In the main analyses, we focused on relatively short-term effects (12 weeks). As explained above, the main reason behind this choice is that the longer the time horizon we consider, the more likely that other unobserved actions that affect outcomes can occur. Keeping this limitation in mind, we present estimates looking at 24 and 48 weeks before and after the treatment in Table 4 using TWFE and in Table 15 in Appendix B using Gardner (2022). In column 1-3 of both tables, we present the result using 24 weeks, and in column 4-6 using 48 weeks. Overall, we continue to observe positive and significant effects for both time window. Finally, in Table 5, we present the results using TWFE and the full dataset (Gardner (2022)'s estimates are reported in Table 16 in Appendix B). Again, we continue to see estimates consistent with our main estimates.

Table 5: DD estimates using TWFE and the full dataset

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.087***	0.078***	0.023***
	(0.009)	(0.010)	(0.004)
Controls	Yes	Yes	Yes
Observations	3,698,916	3,764,131	3,764,131
R^2	0.661	0.676	0.796

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates obtained using TWFE. All models include product and year-week fixed effects, and the list of controls described in Section 6. Standard errors reported in parentheses are clustered at the product level.

6.3 European countries

Here we replicate the results obtained with the U.S. dataset using the European dataset described in Section 3. We estimate Equation 1 and 2 using the European dataset and including all controls discussed above. We report these results using TWFE in Table 6 and

using Gardner (2022)'s estimator in Table 17 in Appendix B. We obtain results that are very consistent with those obtained with the U.S. dataset, i.e., we observe a 16.8% increase in GMS, 15.5% increase in sales, and 4.2% increase in shipped units.

Table 6: DD estimates using TWFE and controls for the European countries

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.155***	0.145***	0.041***
	(0.005)	(0.005)	(0.002)
Controls	Yes	Yes	Yes
Observations	1,165,842	1,179,214	1,179,214
R^2	0.658	0.674	0.796

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates obtained using TWFE. All models include product and year-week fixed effects, and the list of controls described in Section 6. Standard errors reported in parentheses are clustered at the product level.

6.4 Accounting for search rank

Another potential concern with the main results is that they might be artificially driven by Amazon preferencing CPF products over non-CPF products. For example, Amazon could boost the visibility of CPF products by lowering their search rank, which would, in turn, increase their sales. To address this concern, we collected weekly search rank data for all the products in our dataset.[13] We then estimate Equations 1 and 2, including the log of search rank as an additional control. We present the results using TWFE in Table 7 below and using Gardner (2022)'s estimator in Table 18 in Appendix B. Because the sample including search rank is slightly smaller than the original sample, in the first three columns of Table 7, we replicate the main results, including the controls discussed above and reported in Table 3. We find estimates that are extremely similar. In Columns 4–6 of Table 7, we report the

[13]Due to privacy regulations, these data are available starting in August 2022.

estimates that include search rank as a control. As expected, lower search rank increases GMS, sales, and shipped units. Moreover, our main estimates are minimally affected by the inclusion of search rank, suggesting that Amazon does not provide any advantage to CPF products.

Table 7: DD estimates using TWFE accounting for search rank

	log GMS	log Sales	log Units	log GMS	log Sales	log Units
	(1)	(2)	(3)	(4)	(5)	(6)
After	0.127***	0.121***	0.045***	0.122***	0.116***	0.043***
	(0.008)	(0.008)	(0.003)	(0.008)	(0.008)	(0.003)
log Search Rank				−0.100***	−0.095***	−0.042***
				(0.004)	(0.004)	(0.002)
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	837,022	852,712	852,712	837,022	852,712	852,712
R^2	0.697	0.718	0.822	0.697	0.718	0.822

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates obtained using TWFE. All models include product and year-week fixed effects. Standard errors reported in parentheses are clustered at the product level.

6.5 Cross-country DD

In this section, we perform an analysis that tries to address the concern that our results may be driven by unobserved differences in attributes (e.g., quality, brand awareness) between CPF and non-CPF products. Specifically, we follow a strategy similar to that adopted by several papers studying the effect of review responses on ratings (Proserpio and Zervas, 2017; Chevalier et al., 2018). In these papers, the authors implement a cross-service DD leveraging the fact that the same offer may be listed on multiple services. Under the assumption that there are no service-specific effects, this strategy effectively reduces concerns about treated units being different than control units, since treated and controls are exactly the same.

In our specific case, we benefit from the fact the same product is listed on Amazon in

multiple countries and the same product may have joined the CPF program in one country but not the other. We randomly sample about 4,000 products that eventually join CPF in the UK but not in Germany. We then estimate the following specification:

$$Y_{ijt} = \beta_1 \text{Treated}_{ij} + \beta_2 \text{After}_{it} + \beta_3 \text{Treated}_{ij} \times \text{After}_{it} + \text{X}_{ijt}^{'}\gamma + \alpha_i + \tau_t + \epsilon_{ijt}, \tag{4}$$

where Y_{ijt} is the outcome of product i in country j at year-week t. Treated_{ij} is an indicator of whether product i is sold in the treated country j (i.e., UK), After_{it} is an indicator of whether product i has joined the CPF program at time t. X_{ijt} is a vector of the same controls described above, i.e, total ad spend, price and promotions, average ratings and number of reviews. Finally, we include in the model year-week fixed effects τ_t and product-pair fixed effects α_i, i.e., a fixed effect for each product. We estimate Equation 4 using OLS and clustering standard errors at the product-pair level.

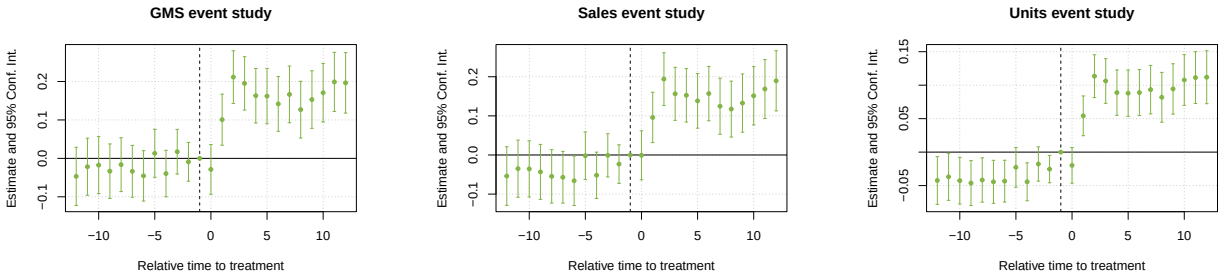
Event study As we have done for the main analysis, we start by presenting event studies for the three outcomes. To do so, in a similar way as we did for the staggered DD model, we compute relative-to-the-treatment-week weekly intervals and then we estimate the following model:

$$Y_{ijt} = \beta_1 \text{Treated}_{ij} + \beta_2 \text{After}_{it} + \sum_{k \in [-12,12]} \theta_k \mathbb{1}_{\{t - t_i^* = k\}} \times \text{Treated}_{ij} + \alpha_i + \tau_t + \epsilon_{ijt}, \tag{5}$$

where everything is like in Equation 4 but replace After_{it} in the interaction term with a set of weekly dummies around the treatment as defined in Equation 3. We estimate Equation 5 using OLS and setting the baseline interval to be -1, i.e., the interval right before the treatment starts. We continue to cluster standard errors at the product-pair level. We plot the θ_k estimates for GMS, sales, and shipped units in Figure 7. For the three outcomes, we observe that in the pre-treatment period estimates are close to zero, providing support for the

parallel trends assumption.[14] Further, in the post-treatment period, we see a positive jump in all three outcomes, suggesting a positive effect associated with joining the CPF program.

Figure 7: Cross-country DD event studies. The black vertical dashed line represent the period before joining the CPF program.



Results Having validated the assumption behind the cross-country DD strategy, we proceed to estimate Equation 4 and obtain the overall ATT of the CPF program. As we have done for the main analysis, we focus on a period of 12 weeks before and 12 weeks after joining the program. We present these results in Table 8. In columns 1-3, we present the estimates without controls, and in columns 4-6 with controls. We continue to observe a positive and significant effect of the CPF program on all three outcomes. In addition, and similar to what we observed in the main analysis, adding controls reduces the estimates suggesting that selling partners support CPF products with other marketing activities. Despite this, even after accounting for all the controls, we find that CPF leads to a 9.6% increase in GMS, a 8.9% increase in sales, and a 8% increase in shipped units for identical products across countries.

6.6 Additional robustness checks

We conclude this section, by discussing two additional robustness checks aimed at reducing concerns about unobservable differences between CPF and non-CFP products that can drive

[14]For Net Shipped Units, estimates are slightly negative in the pre-treatment period. This is due to the baseline period being slightly more positive than the rest of the pre-treatment periods. For example, if we had chosen Interval -12 as a baseline, we would have observed statistically insignificant pre-treatment estimates.

Table 8: Cross-country DD

	log GMS	log Sales	log Units	log GMS	log Sales	log Units
	(1)	(2)	(3)	(4)	(5)	(6)
After × Treated	0.168***	0.170***	0.119***	0.092***	0.086***	0.077***
	(0.023)	(0.023)	(0.012)	(0.017)	(0.016)	(0.008)
After	0.003	0.001	−0.022**	0.013	0.012	−0.017**
	(0.020)	(0.019)	(0.010)	(0.016)	(0.015)	(0.007)
Treated	0.885***	0.882***	0.486***	0.128***	0.134***	0.045***
	(0.038)	(0.037)	(0.020)	(0.022)	(0.021)	(0.011)
Controls	No	No	No	Yes	Yes	Yes
Observations	165,360	167,040	167,040	165,189	166,869	166,869
R^2	0.584	0.585	0.649	0.747	0.765	0.825

Note: *p<0.1; **p<0.05; ***p<0.01

Estimates obtained using the cross-country DD. All models include product-pair and year-week fixed effects. Standard errors reported in parentheses are clustered at the product level.

our results.

First, we show that our results hold if we only consider eventually treated products, i.e, products that eventually join the CPF program. We report these results using TWFE in Table 9.

Table 9: TWFE using only eventually treated products

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.101***	0.102***	0.032***
	(0.008)	(0.007)	(0.003)
Controls	Yes	Yes	Yes
Observations	729,036	742,261	742,261
R^2	0.723	0.739	0.847

Note: *p<0.1; **p<0.05; ***p<0.01

Second, we perform matching using propensity score via a logistic regression to create

a more balanced panel and show that we obtain similar results to those presented in the paper. We match on the log of price, log of price promotions, log of total ad spend, log of cumulative weekly reviews, and the cumulative average ratings (we take pre-treatment averages for treated products and use the full period for nevertreated products).[15] We are able to match 31,111 control products to 7,800 treated products. We report the summary statistics before and after matching (the average distance between treated and controls and the standardized mean difference—a measure of how far apart the treated and control group means are, in units of the pooled standard deviation—for all variables used for matching) in Tables 10. We observe that matching does a reasonable job at balancing the variables across treated and control products. Moreover, the standardized mean difference is equal or under 0.1 which is generally considered a sign of good balance. We report the TWFE estimates using the matched data in Table 11. These estimates are very similar to the main estimates reported in Table 1.

Table 10: Matching balance summary

	Before Matching	**After Matching**
Variable	**Std. Mean Diff.**	**Std. Mean Diff.**
Average Distance	0.5498	0.0000
log(Price)	0.1406	-0.1031
log(Promo Amount + 1)	0.2121	0.0248
log(Total Ad Spend+1)	0.2549	0.0180
log(Cum. Weekly Reviews+1)	0.3911	0.0416
Cum. Avg. Ratings	0.4495	0.0078

7 Who benefits more from the CPF program?

We conclude our analyses on the effects of the CPF program by studying treatment heterogeneity. We focus on two product characteristics: visibility and category.

[15] We exclude search rank because due to privacy regulations, this data is available starting in August 2022.

Table 11: TWFE after matching

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.123***	0.116***	0.044***
	(0.010)	(0.010)	(0.005)
Controls	Yes	Yes	Yes
Observations	1,190,939	1,211,993	1,211,993
R^2	0.671	0.689	0.794
Note:			*p<0.1; **p<0.05; ***p<0.01

Visibility We explore heterogeneity in the effect of the CPF program for different levels of product visibility. We rely on a measure of visibility that Amazon computes for all products using page views. Amazon classifies products into five categories using product page view quintiles as cutoffs. To assign a quintile to a product, we look at the 12 weeks long pre-treatment period and take average quintile. We then estimate Equation 1 for each quintile. We report the results using TWFE in Table 12. We observe that most of the gains are for products in lower visibility quintiles, with the effect decreasing as becoming not significant as we move toward high visibility products. However, we caution the reader that the sample size decreases substantially as we move from lower to higher product page view quintiles, which may lead to imprecise estimates for high visibility products. These results suggest that consumer preference and search may make previously low-visibility products organically more visible and therefore increase their sales.

Product category We also estimate the differential impact of the CPF program by product category (consumables, hardlines, and softlines). We report these results in Table 13. We find significant estimates for all three categories, with consumable products experiencing the largest gains across all outcomes, followed by hardlines and then softlines. These results are in line with Borin et al. (2011) who find that the impact of adding environmental information is greater for consumable products. Recall that consumables are non-durable products

Table 12: DD estimates using TWFE by product visibility

	Results by Product Page View Quintile:				
Quintile	5th (top 20%)	4th	3rd	2nd	1st (bottom 20%)
	(1)	(2)	(3)	(4)	(5)
	GMS				
After	0.074	0.089**	0.088**	0.185***	0.131***
	(0.048)	(0.037)	(0.035)	(0.028)	(0.009)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	39,719	56,655	63,160	99,664	903,620
R^2	0.749	0.721	0.688	0.704	0.544
	Sales				
After	0.041	0.078**	0.086**	0.157***	0.127***
	(0.048)	(0.036)	(0.035)	(0.028)	(0.009)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	41,136	58,798	65,329	102,894	915,771
R^2	0.763	0.741	0.708	0.720	0.565
	Units				
After	0.014	0.035**	0.031*	0.077***	0.046***
	(0.020)	(0.016)	(0.016)	(0.012)	(0.004)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	41,136	58,798	65,329	102,894	915,771
R^2	0.876	0.854	0.837	0.827	0.674

Note: *p<0.1; **p<0.05; ***p<0.01

Estimates obtained using TWFE. All models include product and year-week fixed effects, and the list of controls described in Section 6. Standard errors reported in parentheses are clustered at the product level.

Table 13: DD estimates using TWFE by product category

	Product category:		
	Consumables	Hardlines	Softlines
	(1)	(2)	(3)
	GMS		
After	0.260***	0.131***	0.090***
	(0.026)	(0.013)	(0.010)
Controls	Yes	Yes	Yes
Observations	81,658	403,085	736,681
R^2	0.834	0.750	0.543
	Sales		
After	0.247***	0.122***	0.088***
	(0.026)	(0.013)	(0.010)
Controls	Yes	Yes	Yes
Observations	82,151	411,923	748,910
R^2	0.842	0.767	0.568
	Units		
After	0.097***	0.046***	0.021***
	(0.013)	(0.006)	(0.004)
Controls	Yes	Yes	Yes
Observations	82,151	411,923	748,910
R^2	0.915	0.846	0.665

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates obtained using TWFE. All models include product and year-week fixed effects, and the list of controls described in Section 6. Standard errors reported in parentheses are clustered at the product level.

such as food and personal care items. Generally, both these types of products emphasize consumer safety and require stringent quality control measures during manufacturing. This ensures that products are safe for consumption or use and meet the standards set by regulatory bodies. Given that it is for these products that the CPF program performs better, a potential explanation for these results is that the CPF program and certifications may be seen by consumers as an additional check that validates the quality of these products, leading to changes in consumer purchase behavior.

8 Conclusions

This paper provides empirical evidence that sustainability programs which help consumers identify products with environmental certifications can positively impact consumer purchase behavior. Using data on tens of thousands of products sold on Amazon in the U.S. and Europe, we find that joining Amazon's Climate Pledge Friendly program leads to significant increases in sales revenue, product sales volume, and overall consumer spending on those products.

Across multiple analyses using different estimation approaches and robustness checks, the results consistently show positive impacts of Climate Pledge Friendly certification on product performance. Importantly, by leveraging internal data on sellers' marketing activities, we find that controlling for advertising spend, pricing, and promotions substantially affects the magnitude of these estimates. While unadjusted effects suggest increases of approximately 16% in both gross merchandise sales and sales and 6% in shipped units volume, accounting for these marketing activities reveals more modest increases of 13.3% in gross merchandise sales, 12.5% in sales, and 4.4% in shipped units volume. These differences highlight the importance of controlling for concurrent marketing efforts when evaluating sustainability programs, as sellers often increase support for certified products. Notably, we find larger impacts for consumable product categories like grocery and personal care items compared

to durable products like furniture and apparel.

These positive effects on consumer demand persist over longer time horizons up to 48 weeks after certification. The impacts are even larger for products with lower initial visibility to consumers. This suggests the program helps surface sustainable options consumers might otherwise overlook. Importantly, we replicate these findings across the U.S. as well as five major European countries—the UK, Germany, France, Italy and Spain.

Overall, these findings indicate that despite often-discussed gaps between stated sustainability preferences and real purchase decisions, providing clear sustainability certifications and making sustainable choices more discoverable can effectively influence consumer purchasing in favor of environmentally-friendly products, especially in frequently purchased consumable categories. As demand for sustainable goods continues growing, programs like Climate Pledge Friendly offer a way for retailers and brands to meet this demand while incentivizing more sustainable practices. It is important to note that these findings pertain specifically to Amazon's Climate Pledge Friendly program and may not be generalizable to all sustainability programs or online retailers. Future research should examine the applicability of these results to other contexts and platforms.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge, "When should you adjust standard errors for clustering?," *The Quarterly Journal of Economics*, 2023, *138* (1), 1–35.

Aneja, Abhay, Michael Luca, and Oren Reshef, "The Benefits of Revealing Race: Evidence from Minority-owned Local Businesses," Technical Report, National Bureau of Economic Research 2023.

Araya, Sebastian, Andres Elberg, Carlos Noton, and Daniel Schwartz, "Identifying food labeling effects on consumer behavior," *Marketing Science*, 2022, *41* (5), 982–1003.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How much should we trust differences-in-differences estimates?," *The Quarterly journal of economics*, 2004, *119* (1), 249–275.

Borin, Norm, Douglas C Cerf, and Ragi Krishnan, "Consumer effects of environmental impact in product labeling," *Journal of Consumer Marketing*, 2011, *28* (1), 76–86.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting event study designs: Robust and efficient estimation," *arXiv preprint arXiv:2108.12419*, 2021.

Brecko, Kristina and Yewon Kim, "Sustainability and Strategic Differentiation: Rising Preferences and Divergent Brand Strategies in Unregulated Consumer Markets," *Kilts Center at Chicago Booth Marketing Data Center Paper (forthcoming)*, 2024.

Buerke, Anja, Tammo Straatmann, Nick Lin-Hi, and Karsten Müller, "Consumer awareness and sustainability-focused value orientation as motivating factors of responsible consumer behavior," *Review of Managerial Science*, 2017, *11*, 959–991.

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of econometrics*, 2021, *225* (2), 200–230.

Chaisemartin, Clément De and Xavier d'Haultfoeuille, "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey," *The Econometrics Journal*, 2023, *26* (3), C1–C30.

Chevalier, Judith A, Yaniv Dover, and Dina Mayzlin, "Channels of impact: User reviews when quality is dynamic and managers respond," *Marketing Science*, 2018, *37* (5), 688–709.

Chu, Junhong, Xunkun Long, and Shanfeng Zhang, "Can firms Do Well by Doing Environmental Good? Cost-Benefit Analyses of Amazon's Climate Pledge Friendly program," *Available at SSRN*, 2025.

Delmas, Magali A and Laura E Grant, "Eco-labeling strategies and price-premium: the wine industry puzzle," *Business & Society*, 2014, *53* (1), 6–44.

Dubois, Pierre, Paulo Albuquerque, Olivier Allais, Céline Bonnet, Patrice Bertail, Pierre Combris, Saadi Lahlou, Natalie Rigal, Bernard Ruffieux, and Pierre Chandon, "Effects of front-of-pack labels on the nutritional quality of supermarket food purchases: evidence from a large-scale randomized controlled trial," *Journal of the Academy of Marketing Science*, 2021, *49* (1), 119–138.

Essiz, Oguzhan, Sidar Yurteri, Carter Mandrik, and Aysu Senyuz, "Exploring the Value-Action gap in green consumption: Roles of risk aversion, subjective knowledge, and gender differences," *Journal of Global Marketing*, 2023, *36* (1), 67–92.

Feng, Xiaohang (Flora), Xiao Liu, Shunyuan Zhang, and Kannan Srinivasan, "Sustainability and Competition on Amazon," *Available at SSRN 4958107*, 2024.

Gardner, John, "Two-stage differences in differences," *arXiv preprint arXiv:2207.05943*, 2022.

GlobeScan, "Healthy and Sustainable Living Highlights Report 2022," Technical Report, GlobeScan November 2022.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of econometrics*, 2021, *225* (2), 254–277.

Ikonen, Iina, Francesca Sotgiu, Aylin Aydinli, and Peeter WJ Verlegh, "Consumer effects of front-of-package nutrition labeling: An interdisciplinary meta-analysis," *Journal of the academy of marketing science*, 2020, *48*, 360–383.

Kim, Antino and Che-Wei Liu, "When Good Intentions Backfire: The Asymmetric Effects of Minority-Ownership Markers for Businesses on Online Review Platforms," *Available at SSRN 4149895*, 2023.

McKinsey and NielsenIQ, "Consumers care about sustainability-and back it up with their wallets," Technical Report, McKinsey & Company February 2023.

Pachali, Max J, Marco JW Kotschedoff, Arjen van Lin, Bart J Bronnenberg, and Erica van Herpen, "How do nutritional warning labels affect prices?," *Journal of Marketing Research*, 2023, *60* (1), 92–109.

Proserpio, Davide and Georgios Zervas, "Online reputation management: Estimating the impact of management responses on consumer reviews," *Marketing Science*, 2017, *36* (5), 645–665.

Rao, Anita and Raluca Ursu, "The Impact of Voluntary Labeling," *Marketing Science*, 2024. forthcoming.

Shangguan, Siyi, Ashkan Afshin, Masha Shulkin, Wenjie Ma, Daniel Marsden, Jessica Smith, Michael Saheb-Kashaf, Peilin Shi, Renata Micha, Fumiaki Imamura et al., "A meta-analysis of food labeling effects on consumer diet behaviors and industry practices," *American journal of preventive medicine*, 2019, *56* (2), 300–314.

Simon-Kucher & Partners, "Global Sustainability Study 2022: People are ready for environmental change-if the price is right," Technical Report, Simon-Kucher & Partners September 2022.

Taufique, Khan Md Raziuddin, Andrea Vocino, and Michael Jay Polonsky, "The influence of eco-label knowledge and trust on pro-environmental consumer behaviour in an emerging market," *Journal of Strategic Marketing*, 2017, *25* (7), 511–529.

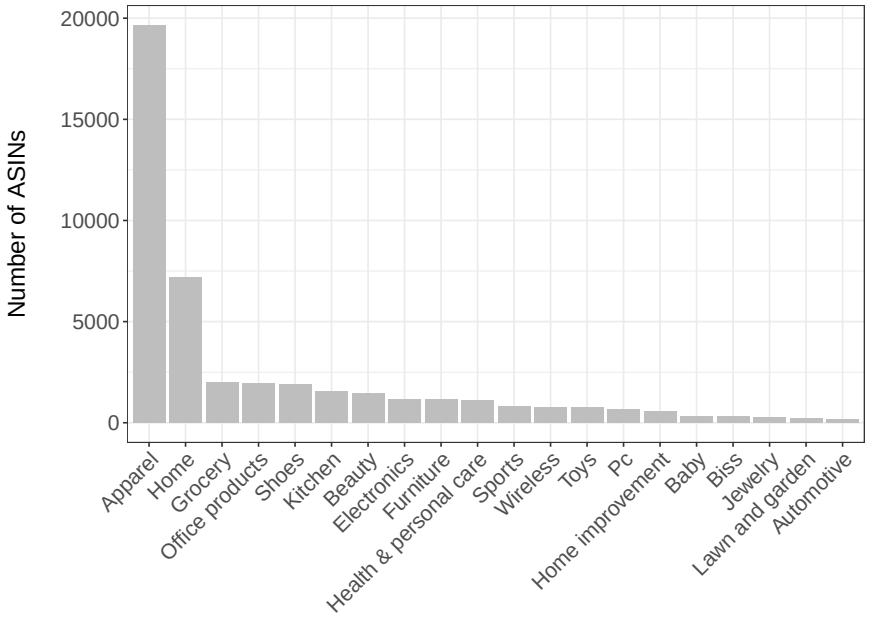
Wang, Caroline and Sherry He, "Do Judge a Product by Its Sustainability Label: Evidence from the Amazon Marketplace," *Available at SSRN*, 2024.

Zervas, Georgios, Davide Proserpio, and John W Byers, "The rise of the sharing economy: Estimating the impact of Airbnb on the hotel industry," *Journal of marketing research*, 2017, *54* (5), 687–705.

Appendix

A Additional figures

Figure 8: Number of products by category for the EU dataset



B Additional tables using Gardner (2022)

In Table 14 and 15, we report Gardner (2022)'s estimates including controls and for different time windows (24 and 48 weeks), respectively. In Table 16, we report the estimates using Gardner (2022)'s estimator and the complete dataset. In Table 17, we report the estimates for the European countries using Gardner (2022)'s estimator. In Table 18, we report the estimates controlling search rank using Gardner (2022)'s estimator.

Table 14: DD estimates using Gardner (2022) with controls

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.116***	0.110***	0.041***
	(0.009)	(0.009)	(0.004)
Controls	Yes	Yes	Yes
Observations	1,202,758	1,223,968	1,223,968
R^2	0.002	0.002	0.003

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates obtained using the Gardner (2022)'s two-stage approach. All models include product and year-week fixed effects, and the list of controls described in Section 6. Standard errors reported in parentheses are clustered at the product level.

Table 15: DD estimates using Gardner (2022) over different time windows

	BW = 24			*BW = 48*		
	log GMS	log Sales	log Units	log GMS	log Sales	log Units
	(1)	(2)	(3)	(4)	(5)	(6)
After	0.146***	0.127***	0.049***	0.170***	0.154***	0.052***
	(0.010)	(0.010)	(0.004)	(0.011)	(0.011)	(0.005)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1,700,315	1,731,018	1,731,018	2,391,314	2,435,748	2,435,748
R^2	0.003	0.003	0.005	0.004	0.003	0.004

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates obtained using the Gardner (2022)'s two-stage approach. All models include product and year-week fixed effects, and the list of controls described in Section 6. Standard errors reported in parentheses are clustered at the product level.

Table 16: DD estimates using Gardner (2022) and the full dataset

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.095***	0.076***	0.014**
	(0.012)	(0.012)	(0.006)
Controls	Yes	Yes	Yes
Observations	3,600,970	3,664,432	3,664,432
R^2	0.002	0.002	0.004

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates obtained using the Gardner (2022)'s two-stage approach. All models include product and year-week fixed effects, and the list of controls described in Section 6. Standard errors reported in parentheses are clustered at the product level.

Table 17: DD estimates using Gardner (2022) with controls for the European countries

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.176***	0.165***	0.051***
	(0.006)	(0.006)	(0.003)
Controls	Yes	Yes	Yes
Observations	1,101,616	1,114,433	1,114,433
R^2	0.006	0.005	0.007

Note: *p<0.1; **p<0.05; ***p<0.01
Estimates obtained using the Gardner (2022)'s two-stage approach. All models include product and year-week fixed effects, and the list of controls described in Section 6. Standard errors reported in parentheses are clustered at the product level.

Table 18: DD estimates using Gardner (2022) accounting for search rank

	log GMS	log Sales	log Units
	(1)	(2)	(3)
After	0.144***	0.134***	0.053***
	(0.010)	(0.010)	(0.004)
log Search Rank	-0.013***	-0.013***	-0.0008
	(0.002)	(0.002)	(0.0009)
Additional controls	Yes	Yes	Yes
Observations	820,735	836,177	836,177
R^2	0.004	0.003	0.006

Note: *p<0.1; **p<0.05; ***p<0.01

Estimates obtained using Gardner (2022)'s two-stage approach. All models include product and year-week fixed effects. Standard errors reported in parentheses are clustered at the product level.