The Consumer Value of Ad Personalization: Evidence from Meta

Daniel Deisenroth Utsav Manjeer Nils Wernerfelt*

November 4, 2026

Abstract

An apparent tension that has emerged in regulating digital advertising centers on the use of detailed consumer data. If advertisers can use more data in their ad delivery, there is evidence ad effectiveness will increase; in contrast, some believe the use of such data may pose potential privacy risks or other harms to consumers. These privacy concerns have motivated a large share of regulation around the world, yet there is little evidence on how consumers actually value being served ads based on more versus less data. In this paper, we conduct a large-scale field experiment with 12 million Facebook users where some were randomly assigned to receive less personalized ads. Our treatment condition leverages the 'less personalized ads' experience that Meta created in response to demands from the European Commission. Combining both log data and custom surveys, we estimate effects on more than 40 outcome variables to paint a holistic picture of the consumer experience. We have two main findings. First, under decreased ad personalization, not only do the average levels of ad engagement fall substantially, but users report both less satisfaction with their ads experience and a decreased user valuation for the platform overall. Second, leveraging detailed user demographics from Meta, we conduct a heterogeneity analysis to study if any sizable subgroups exist that appear to increase their engagement meaningfully under decreased ad personalization. We find no evidence of such groups. On net, our results are consistent with ad personalization in our context conveying benefits, not costs, to consumers along the margins we study. Any account of potential harms from ads personalization can be weighed against these results.

*Daniel Deisenroth (Meta, dbdeisenroth@meta.com); Utsav Manjeer (Meta, utsavmanjeer@meta.com); and Nils Wernerfelt (Kellogg, Northwestern; nils.wernerfelt@kellogg.northwestern.edu). Deisenroth and Manjeer are employees of Meta; Wernerfelt is contractor at Meta for data access, and his salary is donated to charity.

1 Introduction

From being nonexistent a few decades ago, digital advertising is now estimated to be more than 75% of all ad spend in the US (Feger 2024). A major differentiator of digital over other channels is the ability to use detailed consumer data to match consumers and ads. However, as the industry has matured, a key regulatory question has been what data should ad platforms be allowed to use in this matching process?

This question has largely been viewed as a tradeoff between what is good for firms and what is good for consumers. There is growing evidence that advertising effectiveness increases meaningfully if advertisers can use more detailed data to deliver ads (e.g., Goldfarb and Tucker (2011), Korganbekova and Zuber (2023), Wernerfelt et al. (2025)). In contrast, many privacy regulators view increased ad personalization as a cost – potentially a substantial one – to consumers. However, to date, there has been little evidence to explore this claim directly.

Ad personalization on modern digital platforms consists of many complex parts. For example, as inputs into ad personalization, data on consumers can be collected, shared across companies, merged with other data, retained for different periods of time, used to make inferences about other characteristics and so on. Consumers may value each of these components differently depending on the specific context. We do not attempt a holistic evaluation of all these components but rather focus on one, policy relevant dimension: holding all else equal, how does the consumer experience change as more data are used to personalize their ads? As we detail later, this question has been both a central focus of regulators and a necessary component to understand for any broader change in the personalization ecosystem.

Understanding how consumers value the degree of personalization in their ads is hard for several reasons. First, many observable proxies one may consider using for consumer utility have ambiguous welfare predictions in theory. For example, if more ad personalization leads to more purchases, it could be good for consumers if they are discovering new products they like; alternatively, it could be bad for consumers if the ads are somehow inducing unwanted purchases. Second, even directly asking consumers about their preferences in this setting is challenging due to the well known privacy paradox (Athey et al. 2017a). The stated preferences of consumers may differ substantially from their revealed preferences, requiring a careful research design to disentangle. Finally, methodologically, while some consumers naturally receive more personalized ads than others, quasi-experimental methods may fail to properly account for the substantial endogeneity issues that arise in online ad delivery (Gordon et al. 2023).

Mindful of these concerns, our approach is to leverage a field experiment on Facebook

where a large sample of users were randomly assigned to receive either normally personalized ads or less personalized ads. For users in each experience, we analyze behavior across a wide range of log data and custom surveys to depict a holistic image of the consumer experience. In addition, following Brynjolfsson et al. (2025), we ran incentivized deactivation studies of users in each experience, allowing us to compare user valuations of Facebook with more vs. less personalized ads. Our between-user design allows us clean, causal estimates of differences in user valuations and experiences without risk of being contaminated by the privacy paradox or having to rely on strong, quasi-experimental assumptions.

For our intervention, there are many ways of operationalizing 'less personalized' ads (LPA). To make our analysis as policy-relevant as possible, the counterfactual we use is the version of Facebook that was created in response to regulatory demands from the European Commission. Similar policies are being considered in other jurisdictions now, making this counterfactual relevant not only for the ongoing one in the EU but for outside of that as well. We describe the product in more detail in the following section, but at a high level, ad delivery in our treatment condition can only use three categories of data: (i) content the user views during the same session, (ii) ads the user engages with, and (iii) the user's age, gender, and location. This cuts out a substantial amount of data – for example, external browsing activity, past purchasing behavior, and more detailed demographics. In effect, it makes advertising on Meta closer to contextual advertising.

Given that intervention, we look at four broad categories of outcome variables. First, leveraging detailed log data, we analyze standard ad engagement metrics such as clicks, conversions, comments, shares, likes, reports, page likes, and x-outs. Second, we study broader on platform engagement metrics, such as time spent, sessions, posting behavior, and activity levels. Third, we analyze data from two custom surveys: one that asks users about their ad experience on platform and another that asks users about how satisfied they are with purchases they made through Facebook ads. These surveys allow us to tell a more detailed story than what would be allowed by log data alone. Finally, as noted above, we ran an incentivized deactivation experiment to compare user valuations for Facebook overall for users receiving personalized versus less personalized ads. This last experiment allows us to more directly address the consumer welfare question. Collectively, these metrics paint a broad picture of the consumer reaction to the ads themselves.

With this set of outcome variables, we conduct two main analyses. The first analysis aims to understand the typical user experience: if we look at the average user behavior under more vs. less personalized ads, how do our outcome variables differ?

Our main effects are consistent with a worse consumer experience under decreased personalization. First, nearly all engagement with ads falls under decreased personalization –

for example, ad clicks, ad comments, ad likes, onsite conversions, self-reported ad relevance and ads satisfaction all fall while ad x-outs rise. The magnitudes are substantial and suggest an altered consumer experience. Overall engagement metrics with the platform move substantially less, suggesting the effects of less personalized ads are much more concentrated on the ads experience. Second, not only do consumers convert more under increased personalization, but they seem to enjoy what they buy – self-reported purchase satisfaction increases as does the probability of purchasing from the advertiser eight months after the experiment. Finally, and importantly, users in the less personalized condition report a median monthly valuation of Facebook that is 7.7% less than that of users in the business as usual arm ($p = .0180$). In other words, consumers value the overall product significantly (and substantially) less when the ads in it are less personalized.

The above results pertain to estimates of the average user. A common concern, however, is that there are subpopulations that not only prefer decreased personalization, but prefer it so much that overall it is optimal to decrease personalization. Our second main analysis aims to shed light on this concern. To do this, we leverage detailed user data from Meta with non-parametric methods for identifying heterogeneous treatment effects across subpopulations (Wager and Athey 2018). These methods allow the data to inform what subgroups may exist that demonstrate positive treatment effects as well as how large their treatment effects are.

In this second analysis, we find the large majority of users are predicted to decrease their ad engagement under less personalization. Further, those who are predicted to increase their engagement (or decrease their negative engagement) with ads have relatively small estimated treatment effects. Grouping users by quantiles of predicted conditional average treatment effects (CATEs) to increase power, we find similar patterns. Finally, looking at user characteristics, our results suggest those users who engage more with the platform experience more adverse effects of less personalized ads. Together, though our analysis cannot rule out the existence of such groups, in our data we detect no evidence of sizable subgroups who engage meaningfully more with either the ads or the platform under less personalized ads.

Our main contribution is large-scale, experimental evidence on consumer valuations of ad personalization as experienced on social media platforms. As aforementioned, much of the literature to date has focused on the firm side, and understanding the benefits of ad personalization to them (e.g., Korganbekova and Zuber (2023), Wernerfelt et al. (2025)). These papers have some suggestive results on the consumer side – via lower returns and more long term purchases, respectively – but our analysis is able to both directly study consumer utility and analyze a much richer set of outcomes.

Our work also builds on the literature that has studied consumer benefits of recommender systems. Two recent examples include Donnelly et al. (2024) who study how personalized product rankings affect consumer surplus on Wayfair and Zielnicki et al. (2025) who analyze how user engagement on Netflix changes under more personalized recommender systems. Our work complements these by being tailored specifically to digital advertising and the policy questions in that space.

Finally, Mustri et al. (2022) and Beknazar-Yuzbashev et al. (2025) both run clever experiments where users are exposed to varying levels of ad targeting. The former finds suggestive evidence of randomized ads conveying lower consumer surplus than more targeted ads; the latter finds no significant difference in user behavior across more vs. less personalized conditions. Our work builds on these by running at a larger scale and focusing on our specific, policy-relevant counterfactual.

Last, our results come with caveats. First, our analysis by construction is partial equilibrium. Were all users to be exposed to less personalized ads (LPA), then the set of advertisers and campaigns (amongst other actors) would shift. It is hard to extrapolate to that general equilibrium counterfactual from our results, but we note (i) current regulatory discussions focus on offering LPA to users, meaning the long term equilibrium may not be an all or nothing one, and (ii) we believe the partial equilibrium analysis is still informative about how large the general equilibrium effects would need to be in order to reverse our findings. Second, the quality of our heterogeneity analysis is limited by the quality of our demographics and sample size. Despite leveraging detailed demographic data and responses from large-scale surveys, our statistical power was still limited. We limit our heterogeneity analysis to log data accordingly.

The rest of the paper is organized as follows. Section 2 provides more background on digital advertising today and how data enters into it; Section 3 overviews our experiment, both the intervention itself and the experimental design; Section 4 describes the data; Section 5 covers our results; and Section 6 concludes.

2 Background

To understand how the data restrictions could affect users, it is useful to understand how data typically enters the ad delivery funnel. This section provides a high-level overview; see Wernerfelt et al. (2025) or Gordon et al. (2023) for additional descriptions.

There are two main ways data enters the ad delivery funnel at Meta: through 'targeting' and 'optimization.' On Meta, the former term is used to describe the target audience that the advertiser specifies; the latter term is used to describe how Meta's algorithms match

ads to users within that target audience.[1] Similar processes exist at nearly every major ad platform.

Under targeting, an advertiser may specify a certain demographic or custom audience (e.g., in-store customers who have provided their email addresses, or demographic profiles such as 'recent movers') to show the ads to. When ad auctions start on Meta, as part of the process the platform restrict to bids from campaigns where the focal user is in the target audience. This restriction is the first place where data enter the ad delivery process.

The second place occurs under delivery 'optimization'. A brief intuition for this process is as follows. Suppose an advertiser wants to drive likes on their Facebook page. Meta could show the advertiser's ads uniformly amongst the target audience. However, it is very likely that some users will engage with the ad more than others. Suppose that women are not liking the focal advertiser's Facebook page after seeing the ad. Meta will observe this quite quickly after the campaign has started. Upon realizing this, Meta could keep distributing the ads uniformly to men and women, but that would likely not be good for either the advertiser – as their ad budget would be spent on people who are not liking their page – or the users, as women would be seeing ads they do not engage with and men could have preferred to see them instead. A way to improve outcomes here (under some assumptions) would be for Meta to train a model to predict the probability a user would like the focal advertiser's Facebook page and then dynamically allocate ad budget to ensure impressions go to users who would like the relevant Facebook page. In this case, dynamically Meta would 'optimize' delivery to only show the ads to men.

Note how useful this process could be for advertisers. They may not know who their target audience truly is or have the data themselves to identify them; rather, Meta can use their optimization algorithms and data to find the users most likely to convert.[2] Nearly every major ad platform today uses a variant of delivery optimization to match users to ads (e.g., Amazon, Google, Meta, TikTok, and X.)

Mechanistically, how optimization is conducted is via the ad auction. Specifically, as described in Tadelis et al. (2023), the final bid is a function of the monetary value from the advertiser, times a weight for expected click through rate and expected conversion rate, where the conversion is the advertiser-specified objective. Hence, in an ad auction for a given user, the ads where Meta predicts the user is likely to convert have their bids scaled

[1] In more recent times, advertisers have been able to use Meta to optimize delivery even including users outside of the advertiser-specified target audience.

[2] This discussion has avoided any mention of incrementality – it could be that Meta is simply showing ads to users who were already going to convert. Wernerfelt et al. (2025) provide a more in-depth discussion of this point; note also in their experiment the number of conversions in their holdouts is very small, suggesting many of the observed conversions are in fact incremental.

up, increasing their chances of winning and making each user more likely to see ads they are likely to convert on.

Calculation of these bidding weights thus require various types of data to effectively match users with ads. For example, information on past purchases the user has made, their on platform engagement, and high dimensional embeddings of their behavioral and demographic characteristics. The data restrictions we analyze limit the ability of the ad platform to use certain kinds of data in the estimation of those weights. We provide more detail on this below.

3 Experiment

3.1 Intervention

There are many ways that ads could be made 'less personalized.' We focus on one, policy relevant example: in 2024, Meta launched a 'less personalized ads' (LPA) version in the European Union to satisfy regulatory demands from European regulators.[3]

Users in the EU could opt into LPA, in which case their ads would only be based on direct interactions with ads such as views and clicks, and limited demographic and contextual information. The demographics include age, gender, imprecise location such as city, and device or browser type; the contextual information includes only data from the concurrent user session. For example, barring the direct ad interactions, a user could not be shown content based on activities taken during any other session, nor based on broader demographics and user behaviors.[4]

Operationally, this meant only limited data could be used to match users to target audiences and compute auction weights in Meta's optimization algorithms. For example, if an advertiser specified a mandatory target audience of people who had visited their website, then no one in LPA could match that criterion and would thus not see ads from that campaign. Similarly, in predicting expected click and conversion rates, Meta could only use the greatly restricted set of data mentioned earlier. It is an empirical question – and part of what we assess in this paper – how much these changes affect the user experience.

We chose this specific intervention for a few reasons. First, it is directly policy relevant. The European Commission has interpreted the Digital Markets Act (DMA) to require Meta

[3]Public-facing documentation on this product from Meta is available at https://www.facebook.com/help/825123042877166.

[4]As an example of broader user behaviors, Wernerfelt et al. (2025) find evidence data on user purchasing activities can substantially increase ad effectiveness on Meta. (This is consistent with a large past literature in marketing, for example, Rossi et al. (1996).) In LPA, such data cannot be used in ad delivery.

to provide its EU users access "to a less personalised but equivalent alternative [to Facebook] without a subscription fee."[5] Our experiment was designed as a test of exactly the 'less personalized' product Meta created in response and launched in November 2024. Hence, it is important to know how much utility (or disutility) consumers get from this less personalized version of Facebook. If consumers dislike it, the optimal regulation may look different than if they greatly benefit from it.

Second, beyond the DMA, regulation that would restrict similar data from being used in ad delivery is being considered in several other countries (e.g., Brazil, India). Hence the results of this analysis speak not only to the ongoing debate in Europe but also the nascent discussions in other countries.

Finally, and more broadly, several of the data that are restricted under LPA – for example, offsite tracking data – are frequently considered under different potential privacy-related regulations and product changes (e.g., Wernerfelt et al. (2025)). While past research has focused on advertiser-side effects of such data loss, there is relatively little known about the consumer-side impacts.

3.2 Design

We analyze two different datasets. First, we leverage a randomized experiment of 1.5 million European Facebook users where 90% were assigned to receive LPA during the fall of 2024. Treatment assignment was done at the account level, meaning across all devices and logins during our experimental window, the experimental arm was fixed. Further, as new users in the EU would sign up for Facebook during our experiment, a random sample of them would be hashed into either control or treatment for our experiment. This setup is identical to that in (Brynjolfsson et al. 2025), as their study used the same underlying experimentation infrastructure. For all of our log data-based analyses, we focus on this subsample.

Second, despite that sample size, survey response rates on platform are low, and so in order to be powered for our survey-based analyses, we surveyed a broader sample of users. Specifically, as the LPA product was being developed, there were several experiments that were run that tested different backend features – we surveyed users in those experiments as well as a random sample of the Facebook general population in the countries where our experiment ran. The number of survey responses vary by survey, but the sampling frame for this broader pool was around 12 million users. In the Appendix we provide several analyses on the validity of this combined sample. First, we show the difference in demographics

[5] See https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1085

between pooled treatment and control arms is extremely small. Second, for each survey based metric we test if we can pool both across treatment arms and across control arms. For each combination, we cannot reject homogeneous effects across arms. Third, the R^2 from the regression predicting treatment as a function of pre-treatment demographics is very close to zero. Finally, we rerun all our survey-based analyses with our pre-treatment controls and the effects change negligibly; similarly, re-running our heterogeneity analysis on the larger sample yields very similar results.

4 Data

Our outcome data come from several different sources. Here we break it down by whether it is log or survey data. We also describe our pre-treatment covariates. In the Appendix we provide balance checks on these across experimental conditions.

Log data. For all users in our experiment, we logged all standard interactions they had both with their ads and with the platform. For the former, we have data on: ad clicks, shares, likes, comments, reports, page opens, page likes, x-outs (including the selected reason why), onsite conversions, and long term purchases.[6] For long term purchases, we looked at all ads the user saw during the experiment, and then analyzed purchasing behavior from those advertisers during a week eight months after the experiment concluded. We note that between initial random assignment to LPA vs. PA and data collection for this metric there could have been a number of endogenous intermediate steps (e.g., retargeting, competitor targeting) that could shift this metric up or down. Hence, we take this metric as suggestive.

For the on platform engagement, we have data on: activeness, time spent, sessions, content impressions, posts, reshares, likes, comments, and reactions (e.g., angry, sad). These are common metrics digital platforms use to gauge user engagement.

Survey data. The survey data we use comes from three different surveys: one focused on an incentivized deactivation study to estimate consumer valuations for Facebook, one focused on consumer sentiment toward ads, and a final one focused on consumer satisfaction with purchases they made. In the Appendix we provide the full text of the questions and possible responses.

For the first survey, we logged the accounts who were in treatment and control for our experiment and collected 23,544 responses from users spread across those conditions. We

[6] We winsorize our log outcomes from above for our analyses. We winsorize all our log data at the 99th percentile to cut out the extreme tail; we do this both for our outcome variables and the values of our pre-treatment covariates.

follow Brynjolfsson et al. (2025) and asked consumers if they would be willing to accept a given amount to stop using Facebook for a month. We randomized the amount across users, enabling us to estimate the median willingness-to-accept for users who saw personalized and less personalized ads (see Brynjolfsson et al. (2025) for more details).

For the second, we collected 2,111 responses from users in our experiment, where we asked each several questions about their experiences with ads on Facebook. These questions focus on whether consumers like the ads they see, their relevance, how disruptive ads are, and so forth. This survey aims to portray the subjective user experience with ads on the platform, apart from their specific behavioral log data.

For the final survey, we collected 14,732 responses from users who had been recorded to have made a purchase in the recent past. This survey simply shows users a product Meta had logged the user as purchasing and asks the user whether they were satisfied with their purchase, not satisfied, or did not purchase the given product. This survey very simply gives us an estimate of consumer satisfaction with their purchases.

Pre-treatment covariates. In total, we have 86 pre-treatment covariates. These can be bucketed into a few broad categories:

- **Basic Demographics**: Age, gender, country dummies, days since joined Facebook, friend count, Messenger user, mobile device features, and so forth.
- **Engagement Data**: For 7- and 28- day windows before the experiment started, engagement metrics with both ads and the platform. For example, time spent, sessions, ad clicks, ad x-outs, conversions, etc.
- **Privacy settings**: if the user had opted out of different tracking options (e.g., Apple's ATT, OBA-opt out).

In the Appendix we provide a complete list of all the pre-treatment covariates.

5 Results

We organize the results in two parts. Section 5.1 reports the average effect of moving from personalized to less personalized ads on each of the outcomes introduced in Section **??**: ad engagement, platform engagement, survey-based attitudes, and stated willingness-to-accept (WTA) for the ad experience. These average effects are the natural starting point for any welfare or policy assessment because they describe what happens to the typical user.

In Section 5.2 we then consider heterogeneity. In particular, a natural question to ask is whether there are any sizable subgroups that may respond differently than the average effects? In this section we first ask whether *any* heterogeneity is present along the most natural pre-specified dimensions by estimating a linear interaction model. We then ask the deeper question of whether we can detect evidence of a subset of users who are predicted to respond favorably to less-personalized ads. We leverage the causal forest approach of **?** to estimate the full distribution of conditional average treatment effects (CATEs) and conduct the corresponding analyses.

5.1 Main Effects

Figure 1 reports the headline average effects of less personalized ads (LPA) on the four outcome families. Each estimate is the coefficient on the LPA indicator from an OLS regression of the outcome on a treatment dummy and a constant, with heteroskedasticity-robust standard errors. The figure expresses each effect as a percentage of the control mean; significance at the 5% level is encoded by filled (significant) versus hollow (not significant) markers.

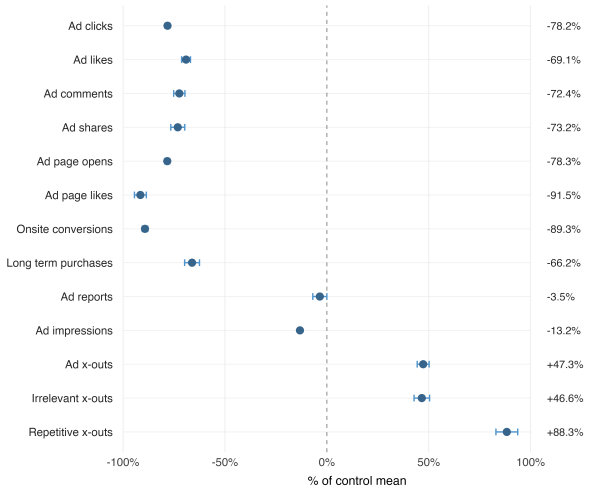
Ad engagement (Panel A). The first order effect of LPA is a large and precisely estimated reduction in several measures of ad engagement: ad clicks fall by 78.2%, ad page opens by 78.3%, ad page likes by 91.5%, onsite conversions by 89.3%, ad shares by 73.2%, ad likes by 69.1%, and ad comments by 72.4%. Long term purchases from advertiser accounts the users saw during the experiment fall by 66.2%, consistent with better match quality under PA.

While increased clicks and other forms of engagement in PA could reflect either better matching or more effective persuasion, the combination with higher self-reported purchase satisfaction (Panel C) and higher WTA (Panel D) suggests the former interpretation.

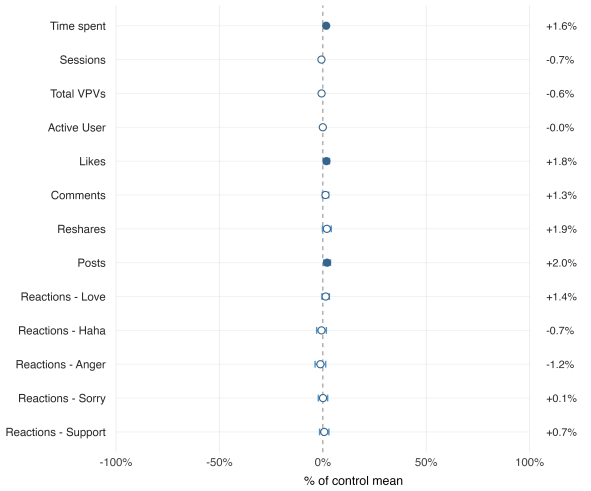
One question that arises is whether the engagement declines are driven by the reduction in ad personalization or simply by the 13.2% reduction in ad load under LPA. We note that the magnitudes here are disproportionate – the ad engagement metrics fall around five to seven times more than the reduction in ad impressions, so engagement per impression drops substantially. Hence, the engagement drop is driven primarily by how users interact with the ads they see, not by how many ads they see.[7]

Working in the opposite direction, users actively dismiss more ads under LPA: ad x-outs

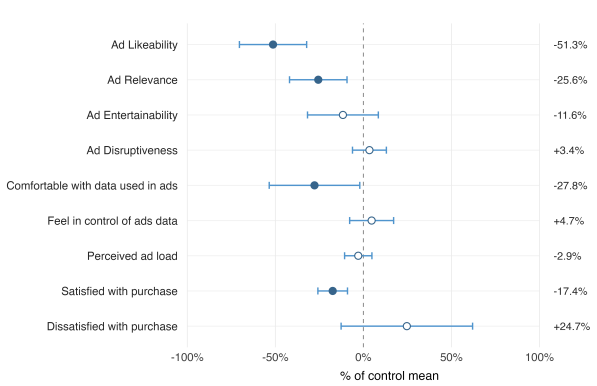
[7]Note we are estimating the effect of assignment to the LPA product, which includes the reduction in ad load and ad personalization. Our counterfactual speaks to the policy-relevant regulation, but given the point estimates, we can see that the reduction in ad personalization played a substantive role itself.



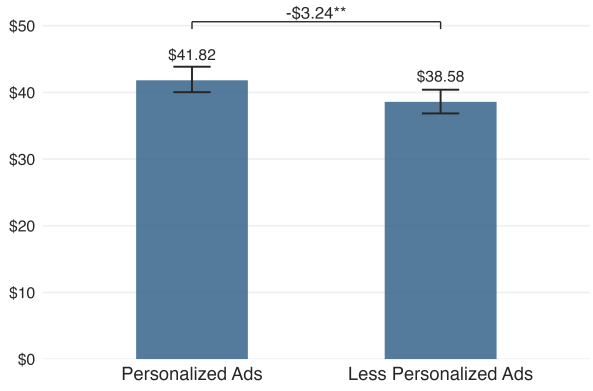
(a) Ad-engagement outcomes.



(b) Platform-engagement outcomes.



(c) Survey outcomes.



(d) Stated WTA for personalized ads ($).

Figure 1: Average effects of less personalized ads. Panels A-C report the OLS coefficient on the LPA indicator as a percentage of the control mean, with 95% confidence intervals; filled markers denote $p < 0.05$. Panel D reports the median elicited WTA in each arm with bootstrap 95% confidence intervals; the p-value for the difference is 0.0180 in two-sided t-test.

rise by 47.3%, with irrelevant-content dismissals up 46.6% and repetitive-content dismissals up 88.3%. Ad reports decrease under LPA, but given the drop in ad impressions, ad reports per ad impression actually increase.

In short, across these thirteen outcomes, LPA delivers ads users engage with markedly less and dismiss markedly more.

Platform engagement (Panel B). Effects on platform-level behavior are an order of magnitude smaller and almost uniformly insignificant, even despite the sample size. Time spent on the platform rises by 1.6%,[8] posts by 2.0%, reshares by 1.9%, and likes by 1.8%; sessions, total video views, and the active-user margin are within tenths of a percent of zero. None of the five Facebook reaction categories shows a movement distinguishable from zero. The platform engagement results indicate that the reduction in ad engagement quality documented in Panel A does not translate into a meaningful contraction of overall platform use over the horizon we observe.

Survey attitudes (Panel C). Users in the LPA arm rate the ads they see as substantially less likeable (−51.3% relative to control), less relevant (−25.6%), and they report lower satisfaction with their recent purchases from Facebook ads (−17.4%). Reported ad-entertainability is lower (−11.6%) but not statistically distinguishable from zero.

The LPA arm reports being substantially less comfortable with how data are used in ads (−27.8%, though this does not survive a Bonferroni correction). We suspect this may be because users find the less personalized ads more noticeable, perhaps raising the salience of ad targeting data. We also note that LPA users do not feel significantly more in control of the data used in their ad targeting, nor do they perceive their ad load as different. These effects are relatively noisily estimated, however, so we do not lean on them too heavily.

Stated WTA for personalized ads (Panel D). The median user under personalized ads asks for $41.82 in compensation to deactivate the platform for one month versus $38.58 for the median user under less personalized ads. The difference of −$3.24 is statistically significant and of a relatively sizable magnitude: about 7.7% of the control (PA) median valuation. This is our best estimate of how ad personalization affects consumer welfare.

Taken together, the four panels find less personalized ads sharply reduce monetizable engagement with ads, have minimal effects on on platform engagement, worsen users' assessments

[8] We note that ads that use detailed, off site data often lead to users to spend less time on the platform, as users click on the ads and are taken off site to browse. This also makes interpreting the sign of the time spent metric – like other metrics here – unclear from a welfare standpoint.

of the ads they see, and generate a Facebook experience that is overall valued less by users than the Facebook experience with personalized ads.

5.2 Heterogeneity

Apart from average effects, a frequent policy concern is there may be subpopulations who are substantially better off under less personalized ads. For example, personalized ads could help firms exploit or scam subpopulations of vulnerable users. If those subpopulations or their treatment effects are sufficiently large, then decreasing personalization could still be optimal from a social welfare perspective.

Ultimately, our ability to speak to this argument is determined by the user-level demographics we observe and the power afforded to us by our sample size. There may be subpopulations we cannot identify in our data or have the power to precisely estimate. With those caveats, we now explore what heterogeneity we can with our user-level demographics and sample.

Our analysis in this section is structured in two parts. We first repeat our main analysis across basic cuts of our data. In this subsection we study whether we see heterogeneity when users are split by pre-treatment covariates of gender, age, activity levels on the platform, engagement levels with ads, or different privacy settings. In the main text we show results for one, main metric from each of the four panels above (ad clicks, being an active Facebook user, reporting liking their ads experience, and WTA); in the Appendix we provide analogous figures for all our outcome variables of interest across these subgroups. As a preview of our results, in none of these subpopulations do we find evidence that less personalized ads improve ad engagement, liking their ads experience, or the value users place on Facebook. Effects on Facebook activity levels are relatively precise nulls, as we saw for the overall sample.

This first analysis restricts to subpopulations defined by single covariates. For a more in-depth exploration, in our second analysis we pull in 86 individual-level pre-treatment covariates from Facebook and flexibly estimate heterogeneous treatments effects across our outcomes using causal forests Wager and Athey (2018). Despite our sample size, power becomes a limiting factor for some metrics; in particular, while using various measures we detect no evidence of heterogeneity for our survey-based or WTA measures, those estimates are imprecisely estimated. For our log-based metrics, we are able to detect evidence of substantial heterogeneity in our sample. However, we find no evidence of a large subgroup that appears to increase its engagement meaningfully under decreased ad personalization. In the main text we focus on ad clicks, but we provide results from each of our log-based outcomes in the Appendix.

5.2.1 Heterogeneity Across Basic Demographics

We first rerun our main analysis on basic demographic cuts of the data. In Figure 2 we provide evidence on how our treatment effects change across these subpopulations for ad clicks, being an active Facebook user, liking their ads experience, and WTA.

A few observations from Figure 2.

First, while we lose precision – particularly in the survey-based and WTA outcomes – the point estimates are relatively stable across subpopulations. While some subpopulations have insignificant treatment effects, the point estimates for clicks, liking their ads experience, and WTA are all of the same sign as the overall effects.

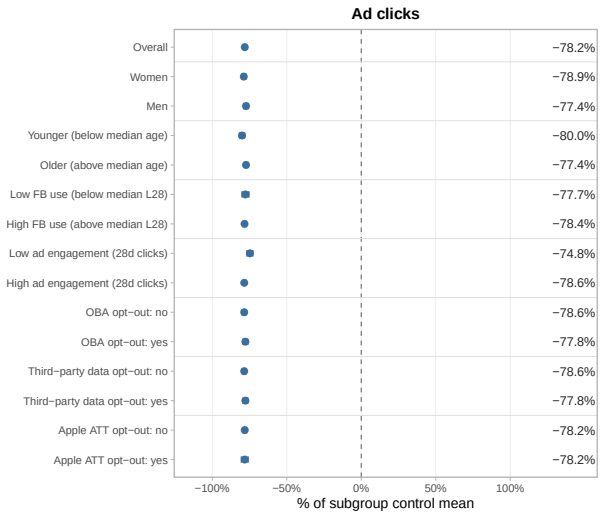
Second, for users with more restrictive pre-treatment privacy settings (i.e., those who had opted out of either OBA, third-party data, or Apple's ATT policy)[9], we note that their WTA treatment effect point estimates are each non-significant and slightly less attenuated versus the point estimates from the less-privacy restricted populations. This pattern is consistent with privacy-sensitive users incurring a smaller utility loss from reduced personalization, possibly because they value reduced data use itself. It is also consistent with a weaker treatment: having opted out, these users already saw less personalized ads at baseline, so the LPA condition changed their ad experience less. In either case, the differences are not statistically significant within any of the three settings, and the estimates for opted-out users become significantly negative with pre-treatment controls. Hence, while there may be some differential effect by pre-treatment privacy settings, it seems the median valuation for Facebook is higher under PA for even the users with more restrictive privacy settings ex ante.

Third, each of these subpopulations is still relatively large and untargeted. We may think, for example, that different interactions of these demographics (e.g., younger women) or users identified by other demographics not listed here may exhibit heterogeneous effects. This motivates our next analysis.

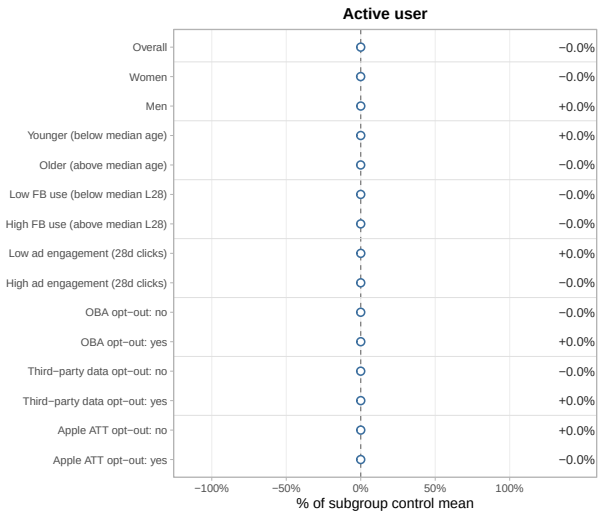
5.2.2 Estimating Heterogeneity More Flexibly

To explore heterogeneity more deeply, we take a broad range of pre-treatment covariates and flexibly estimate heterogeneous treatment effects with them using causal forests

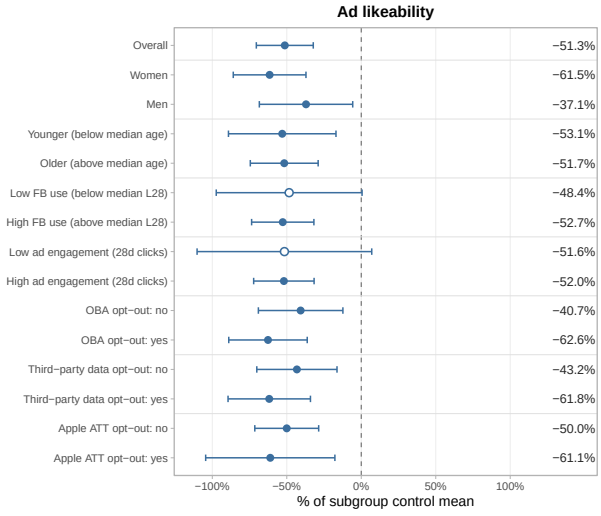
[9]These are the three most used data privacy settings in our sample. OBA opt-out refers to opting out of Online Behavioral Advertising; this was a setting the industry created through self-regulation that prevents external web browsing data from being used in online advertising. Third-party data opt out means any data collected from third parties cannot be used in ad delivery for that user (e.g., publicly available data, or data from other apps). Apple's App Tracking Transparency (ATT) was a feature launched starting in iOS 14.5 whereby opted-out iOS users would not have external data from browsers or other mobile apps shared back with Meta for ad delivery.



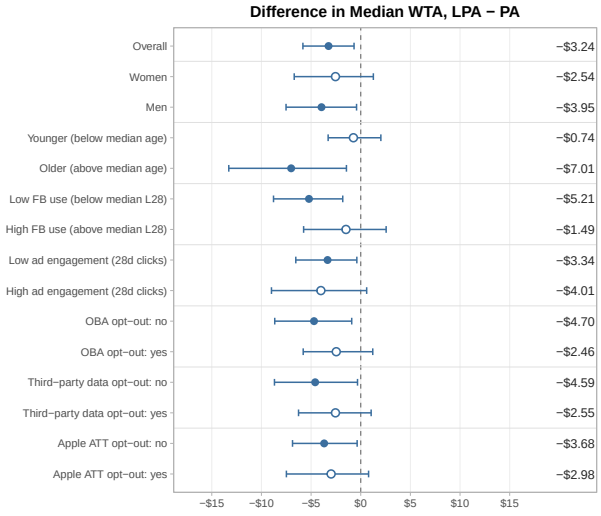
(a) Ad clicks



(b) Active user



(c) Like their ads experience (survey)



(d) WTA ($).

Figure 2: Effects of less personalized ads across different subpopulations for select outcomes. Each panel reruns our main analysis restricted to the respective subpopulation. As before, negative coefficients (Panels A-C) denote the average treatment effect of LPA was negative on that outcome variable for that subpopulation. (Filled markers denote $p < 0.05$.) Overall, we do not see evidence of substantial variability in treatment effects across these subpopulations.

Wager and Athey (2018). (The full list of 86 pre-treatment covariates is documented in Appendix A.)

A cost of such flexible heterogeneous treatment effect procedures is that sample size requirements for well-powered results increase substantially. Even with our relatively large sample of survey respondents, we found our analyses for the survey-based and WTA outcomes could not detect evidence of heterogeneity but were underpowered. We have substantially greater power for the log-based outcomes, however; we report results for each of those in the Appendix. Below, we focus on the single outcome of ad clicks but note when results differ across metrics.

We implement the causal forest by partitioning our sample into a random 50-50 split of a training and test sample. We train the forest on the former and estimate the conditional average treatment effects (CATEs) on the latter. For inference of both the overall average treatment effect and that by different subgroups, we use the augmented-inverse-propensity-weighted (AIPW) doubly-robust score of Athey et al. (2017b), which is a standard estimator in these settings.

For each causal forest we run, we report three diagnostics on the quality of the model's fit to the data. First, we compare the forest-based AIPW point estimate of the ATE to the simple OLS ATE estimate; agreement between the two confirms that the forest preserves the average treatment effect.[10] Second, we report the rank-weighted average treatment effect (RATE) using the area under the targeting operator characteristic curve (AUTOC) of Yadlowsky et al. (2025). The AUTOC summarizes how much average benefit can be captured by treating only the highest-CATE units relative to the overall ATE; a value statistically greater than zero is evidence that the forest's CATE estimates carry signal about who responds most. Finally, we report the calibration test of Chernozhukov et al. (2018). This test regresses held-out outcomes on the forest's mean prediction and on its demeaned differential prediction; if the coefficient on the mean prediction equals one, it confirms that the forest is well-calibrated for the ATE, and if the coefficient on the differential prediction equals one, it suggests that units the forest predicts to have larger treatment effects do, on average, have larger treatment effects in the held-out data.

For ad clicks, we find evidence both that the model fits the data well and that there is evidence of heterogeneity. The AIPW estimate of the ATE on ad clicks is -82.89 (s.e. $=$ 0.349), which is very close to the OLS ATE of -82.81 (s.e. $= 0.188$). The AUTOC statistic is 63.48 (s.e. $= 0.263$, $p < 0.001$): there is highly significant predictable heterogeneity. The calibration test yields a coefficient on the mean forest prediction of 1.001 (s.e. $= 0.004$) and

[10] Note our results in this section are not expressed in terms of percents of the control mean but in absolute terms instead.

a coefficient on the differential forest prediction of 1.025 (s.e. = 0.006), both very close to one. In short, the forest is well calibrated for both the level and the heterogeneity in the treatment effect.

The substantive content of the heterogeneity is reported in Figure 3. In this graph, we have ranked individuals by their individual-level estimated treatment effects and binned individuals by percentiles into 20 bins. We have then plotted the AIPW estimate of the average treatment effect for each bin. Some comments on this graph.

First, over the entire held-out sample of 765,481 users, only 14 are predicted to have a positive treatment effect on clicks. Similarly, each ventile bin lies significantly below zero. These findings suggest that while there is substantial heterogeneity in how users' ad clickiness is predicted to respond to LPA, the variation in predicted treatment effects across users is concentrated in the negative; nearly all users are predicted to click with ads less, very few are predicted to engage more.

Second, we note that the ATE for the lowest ventile is more than 5x larger than the overall ATE. That is consistent with some users being greatly impacted – much more than the ATE would suggest – although, again, all the treatment effect signs are negative.

Finally, in the Appendix we provide evidence that the users who are in the top bin tend to be more marginal users. They are users who pre-experiment spent less time on and engaged less with the platform. This is consistent with a dosage story – the more exposed users are to LPA the larger the magnitude of their response.

For other outcomes, there are some bins who are significantly different from zero and of an opposite sign to the main effect. However, when this occurs, the magnitudes of the effects are small: for example, for ad x-outs, around 25.4% of the sample are predicted to have negative treatment effects – i.e., the number of x-outs will decrease under LPA. The magnitude of the estimated treatment effects is small, however – the median CATE for individuals predicted to be below zero is -0.0072, which is about a 2.6% reduction from the control mean.

There are also metrics where a small share of users are predicted to respond strongly. For example, for ad reports, the bottom ventile is predicted to decrease ad reports by more than 100% of the baseline control under LPA. (Similarly, the top ventile is predicted to increase reports by more than 100% of the baseline.) These results are sensitive to the degree of winsorization, suggesting they are driven by outliers, but from a policy standpoint we may still care about these outliers.

Finally, another metric to mention is timespent, where a large share of users are predicted to increase their timespent under LPA (with a median effect around 3% of the control mean). This appears to be a novelty effect: plotting the time series for the treatment effect shows

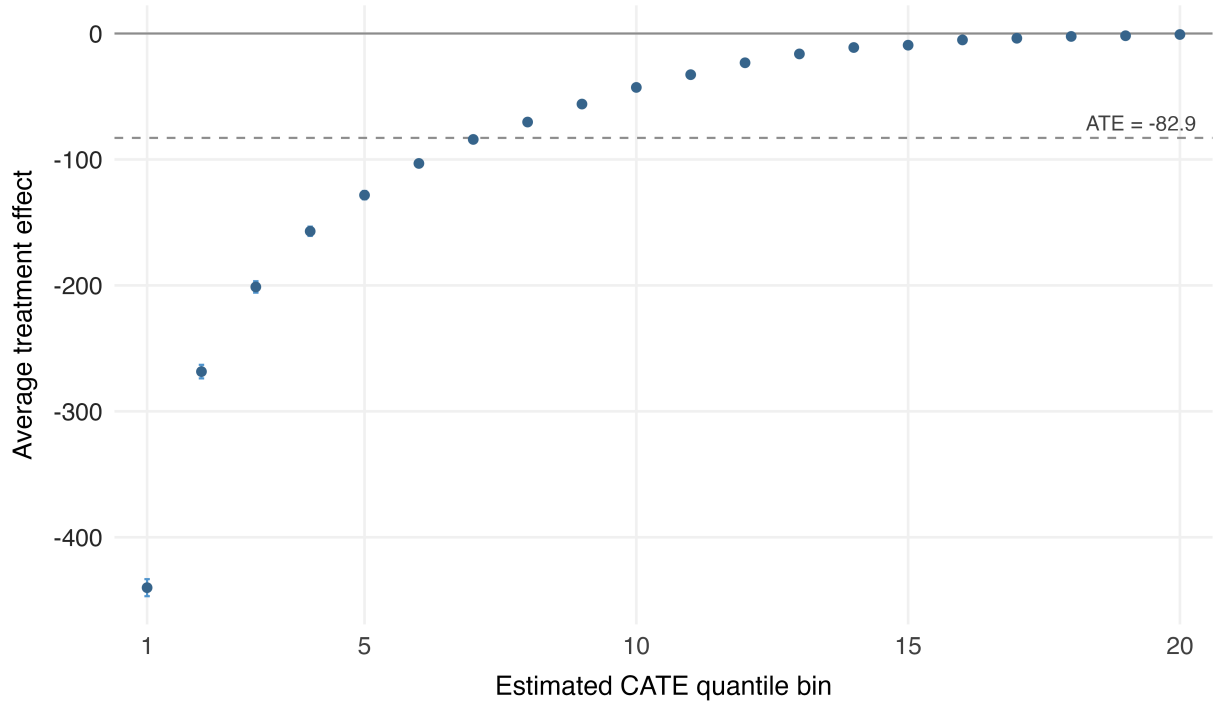


Figure 3: Estimated average treatment effect on ad clicks within out-of-bag CATE-quantile bins. The held-out test sample ($N = 765{,}481$) is partitioned into twenty equally-sized bins by predicted CATE (ventiles); the vertical axis reports the AIPW estimate of the ATE within each bin, with 95% confidence intervals. The dashed horizontal line marks the AIPW estimate of the population ATE (-82.9).

that it spikes at launch and then decreases to zero over the course of the experiment. This suggests (speculatively) that the causal forest is picking up who is likely to be surprised by LPA versus users who consistently engage more.

To conclude this section, net, across both our analyses, with our data we do not find evidence of a large subgroup of users who appear to substantially increase their positive engagement with the platform under LPA. To be clear, our results speak to a lack of evidence, not a definitive rule out, for the existence of such groups. There may be other benefits, over other horizons, with people we cannot identify with our covariates. We leave further exploration to future work.

6 Conclusion

We analyze a set of large-scale field experiments on Facebook where users were randomly assigned to receive less personalized ads (LPA). We analyze more than forty log- and survey-based outcomes to paint a holistic picture of the user experience under LPA.

Our results are consistent with ad personalization in our context conveying benefits, not costs, to consumers. The median user valuation of Facebook is significantly and substantively higher under personalized ads, and that finding is buttressed by effects we observe across a broad range of log- and survey-based outcomes. Our heterogeneity analysis detects no evidence of sizable subgroups whose engagement improves substantially under decreased personalization; in contrast, a large share of individuals are predicted to meaningfully reduce their engagement. Whether there exist additional or offsetting benefits from implementing less personalized ads - for example, from structural improvements to ad platform competition - is a question our design does not address. We leave estimating such potential benefits to future research, and we leave to policymakers how to trade off those versus the costs we document here.

References

Arielle Feger. Digital makes up over three-quarters of total ad spend in the us. Technical report, 2024. URL https://www.emarketer.com/content/digital-makes-up-over-three-quarters-total-ad-spend-us.

Avi Goldfarb and Catherine E Tucker. Privacy regulation and online advertising. *Management Science*, pages 57–71, 2011.

Malika Korganbekova and Cole Zuber. Balancing user privacy and personalization. *Working paper.*, 2023.

Nils Wernerfelt, Anna Tuchman, Bradley T Shapiro, and Robert Moakler. Estimating the value of offsite tracking data to advertisers:: Evidence from meta. *Marketing Science*, 44(2):268–286, 2025.

Susan Athey, Christian Catalini, and Catherine Tucker. The digital privacy paradox: Small money, small costs, small talk. Technical report, National Bureau of Economic Research, 2017a.

Brett R Gordon, Robert Moakler, and Florian Zettelmeyer. Close enough? a large-scale exploration of non-experimental approaches to advertising measurement. *Marketing Science*, 42(4):768–793, 2023.

Erik Brynjolfsson, Avinash Collis, Daniel Deisenroth, Haritz Garro, Daley Kutzman, Asad Liaqat, and Nils Wernerfelt. The consumer welfare effects of online ads: Evidence from a nine-year experiment. *American Economic Review: Insights*, 7(4):447–462, 2025.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Robert Donnelly, Ayush Kanodia, and Ilya Morozov. Welfare effects of personalized rankings. *Marketing Science*, 43(1):92–113, 2024.

Kevin Zielnicki, Guy Aridor, Aurélien Bibaut, Allen Tran, Winston Chou, and Nathan Kallus. The value of personalized recommendations: Evidence from netflix. *arXiv preprint arXiv:2511.07280*, 2025.

Eduardo Abraham Schnadower Mustri, Idris Adjerid, and Alessandro Acquisti. Behavioral advertising and consumer welfare: An empirical investigation. *Federal Trade Commission PrivacyCon*, pages 1–27, 2022.

George Beknazar-Yuzbashev, Rafael Jimenez-Duran, Andrey Simonov, and Mateusz Stalinski. Advertising load discrimination on social media. 2025.

Steven Tadelis, Christopher Hooton, Utsav Manjeer, Daniel Deisenroth, Nils Wernerfelt, Nick Dadson, and Lindsay Greenbaum. Learning, sophistication, and the returns to advertising: Implications for differences in firm performance. Technical report, National Bureau of Economic Research, 2023.

Peter E Rossi, Robert E McCulloch, and Greg M Allenby. The value of purchase history data in target marketing. *Marketing Science*, 15(4):321–340, 1996.

Susan Athey, Guido Imbens, Thai Pham, and Stefan Wager. Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5):278–281, 2017b.

Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association*, 120(549):38–51, 2025.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Appendix

This Appendix contains five sections. In Section A we provide a list of all our pre-treatment covariates and conduct randomization checks on our data; in Section B we provide evidence on whether we can pool our treatment and control arms for the survey and WTA analyses; in Section C we provide additional details on our surveys, the recruitment flow, response rates, and so forth; in Section D we provide the results from causal forest analyses for each of our log-based outcomes (our survey-based and WTA outcomes were underpowered); and finally in Section E we provide details on two supplementary analyses – the relative role of ad versus match quality in driving our results and who the individuals are who are predicted to respond the most vs. least to the LPA intervention. For the last analysis, we provide results for ad clicks and ad x-outs, but we see similar patterns across our other metrics as well.

A Pre-treatment Covariates and Balance Checks

A.1 Pre-treatment covariates

For each user in our data, we collected detailed pre-treatment covariates from Facebook. We use these both in our heterogeneity analyses, but also to provide evidence on randomization (next subsection) and for robustness checks (e.g., re-running analyses controlling for these covariates). Broadly, our 86 demographics can be categorized into three groups:

- **Basic demographics.**
  - Age
  - Gender (female indicator)
  - Account age (days since joining on Facebook).
  - Predicted country of residence (Belgium, Germany, Spain, France, Italy, Netherlands, Poland, Portugal, Romania, other, other)
  - Profile-picture present indicator
  - Relationship-status specified indicator
  - Primary phone operating system (Android, iOS).

  - Device RAM class (under 0.75 GB, 0.75–1 GB, 1–1.5 GB, 1.5–2 GB, 2–3 GB, 3–4 GB, 4 GB and above, other)
  - Messenger user indicator

- **Historical engagement.**
  - Social graph: friend count, subscriber count, subscription count, number of friend requests initiated.
  - Activity: Days active in the past 7 days; days active in the past 28 days; days active on web in the past 28 days; days active on mobile in the past 28 days.
  - Ad engagement in the past 7 days: impressions, clicks, likes, comments, shares, page opens, page likes, onsite conversions, onsite revenue.
  - Platform engagement in the past 7 days: app sessions, time spent.
  - Ad engagement in the past 28 days: impressions, clicks, likes, comments, shares, page opens, page likes, x-outs, x-outs reported as irrelevant, onsite conversions, onsite revenue.
  - Platform engagement in the past 28 days: app sessions, time spent, reshares, likes, comments, posts, total VPVs (video plays viewed), and post reactions by type (Love, Support, Anger, Haha, Sorry).

- **Privacy settings.** Indicators for a number of different privacy settings the user could have made:
  - Online behavioral advertising (OBA)
  - Third-party-data (3PD) opt-out
  - Off-Facebook Activity (OFA) opt-out
  - Account Association Disablement (AAD)
  - ePrivacy Directive (EPD) opt-out
  - Apple App Tracking Transparency (ATT) opt-out

For variables that could be missing for some users (e.g., age, sex, account age), we additionally include a binary indicator for whether the value was missing. After deduplicating perfectly collinear columns, this yields the 86 covariates used in the balance table (Appendix A) and in all heterogeneity analyses.

A.2 Balance Checks

We here provide evidence on whether the randomization was carried out correctly for our experimental data and whether there is evidence we can treat the broader sample as if treatment were randomized.

We note that rerunning all our main effects controlling for our pre-treatment covariates yields no material changes in the results. Further, our heterogeneity analysis relies only on the experimental data (though rerunning those analyses on the broader sample yields very similar results).

Given the two samples in our data, below we present two balance checks. The first sample, hereafter the *platform/ads sample*, restricts to the single experiment on which the log-based causal forests are estimated. The second sample, hereafter the *full sample*, pools all randomization arms used for the survey-based outcomes. For each covariate X_j we compute the normalized difference following Imbens and Rubin (2015, Section 14.2):

$$\widehat{\Delta}_j \;=\; \frac{\bar{X}_{j,1} - \bar{X}_{j,0}}{\sqrt{(s^2_{j,1} + s^2_{j,0})/2}}, \tag{1}$$

where $\bar{X}_{j,w}$ and $s^2_{j,w}$ denote the sample mean and variance of X_j in arm $w \in \{0, 1\}$.[11] We also report, for each sample, the R^2 from the linear-probability regression of the treatment indicator W_i on the full vector of demographics X_i,

$$W_i \;=\; \alpha + X_i'\beta + \varepsilon_i, \tag{2}$$

which provides a joint summary of the predictive power of pre-treatment covariates for treatment assignment.

Platform/ads sample. Across all 86 demographic covariates, the maximum absolute normalized difference is 0.0058 and the median is 0.0017. No covariate exceeds the 0.10 informal threshold of Stuart (2010), and a joint orthogonality F-test of $W \sim X$ does not reject balance ($F(84,\ 1{,}530{,}972) = 0.898$, $p = 0.737$). The R^2 of regression (2) is 4.9×10^{-5}: demographics jointly explain essentially none of the variation in treatment assignment. Table A1 reports the 20 covariates with the largest absolute normalized differences.

[11] Equation (1) is scale-free and, unlike the two-sample t-statistic, does not mechanically grow with sample size; it is therefore the standard balance diagnostic in large-N randomized experiments. Stuart (2010) suggests $|\widehat{\Delta}_j| > 0.10$ and $|\widehat{\Delta}_j| > 0.25$ as informal thresholds of meaningful imbalance.

Full sample. The full sample is large enough that small differences in covariates means become statistically detectable, but they remain small in magnitude. The maximum absolute normalized difference across the 86 demographics is 0.0445 and the median is 0.0036; no covariate exceeds the 0.10 threshold. The joint F-test rejects exact equality of means ($F(84,\ 12{,}129{,}537) = 95.66$, $p < 0.001$), but the R^2 from regression (2) is only 0.0007, again indicating that the demographics jointly carry essentially no information about treatment assignment. Table A2 reports the 20 covariates with the largest absolute normalized differences. As a robustness check, the main-effect estimates reported in the body of the paper are unchanged when the full vector of demographics is included as controls.

Table A1: Balance of demographic covariates—platform/ads sample.

Covariate	N_0	N_1	Mean_0	Mean_1	$\widehat{\Delta}_j$
d_country_prediction_NL	138,943	1,392,114	0.0370	0.0381	0.0058
d_ad_share_28d	138,943	1,392,114	0.0414	0.0402	−0.0047
d_ad_imps_7d_missing	138,943	1,392,114	0.1040	0.1055	0.0047
d_ad_share_7d	138,943	1,392,114	0.0114	0.0110	−0.0046
d_oba_optout_user_missing	138,943	1,392,114	0.1021	0.1034	0.0044
d_country_prediction_IT	138,943	1,392,114	0.1320	0.1305	−0.0042
d_att_optout_user	138,943	1,392,114	0.2292	0.2276	−0.0041
d_country_prediction_PL	138,943	1,392,114	0.0882	0.0893	0.0041
d_friend_count	138,943	1,392,114	392.41	390.06	−0.0040
d_is_messenger_user	138,943	1,392,114	0.9980	0.9979	−0.0038
d_epd_optout_user	138,943	1,392,114	0.0797	0.0806	0.0035
d_ram_class_0-0.75gb	138,943	1,392,114	0.0003	0.0003	0.0035
d_ad_comment_28d	138,943	1,392,114	0.0501	0.0492	−0.0033
d_oba_browser_optout_user	138,943	1,392,114	0.0004	0.0003	−0.0032
d_ch_90d_optout_user	138,943	1,392,114	0.0007	0.0006	−0.0032
d_country_prediction_0	138,943	1,392,114	0.2328	0.2316	−0.0030
d_oba_optout_user	138,943	1,392,114	0.5204	0.5190	−0.0029
d_ram_class_1-1.5gb	138,943	1,392,114	0.0014	0.0015	0.0029
d_ram_class_3-4gb	138,943	1,392,114	0.2963	0.2950	−0.0029
d_reaction_sorry	138,943	1,392,114	0.1352	0.1369	0.0029
Summary across all 86 demographic covariates:					
Max $\lvert\widehat{\Delta}_j\rvert$					0.0058
Mean $\lvert\widehat{\Delta}_j\rvert$					0.0019
Median $\lvert\widehat{\Delta}_j\rvert$					0.0017
Count with $\lvert\widehat{\Delta}_j\rvert > 0.10$					0/86
Count with $\lvert\widehat{\Delta}_j\rvert > 0.25$					0/86
Joint orthogonality F-test					$F(84,\ 1{,}530{,}972) = 0.898,\ p = 0.737$
R^2 from $W = \alpha + X'\beta + \varepsilon$					4.93×10^{-5}

Notes. Subscripts 0 and 1 denote the control ($W = 0$) and treatment ($W = 1$) arms. $\widehat{\Delta}_j$ is the normalized difference for covariate j defined in Equation (1), following Imbens and Rubin (2015, §14.2). The table lists the 20 covariates with the largest absolute normalized differences; summary statistics in the lower panel are computed over all 86 demographic covariates.

Table A2: Balance of demographic covariates—full sample (survey outcomes).

Covariate	N_0	N_1	Mean_0	Mean_1	$\widehat{\Delta}_j$
d_ad_imps_7d_missing	5,162,758	6,966,864	0.0922	0.1054	0.0445
d_oba_optout_user_missing	5,162,758	6,966,864	0.0904	0.1034	0.0442
d_reshares_missing	5,162,758	6,966,864	0.1266	0.1393	0.0374
d_l28	5,162,758	6,966,864	21.4583	21.1002	−0.0360
d_l7	5,162,758	6,966,864	5.3675	5.2773	−0.0351
d_mobile_l28	5,162,758	6,966,864	20.9238	20.5727	−0.0338
d_age_missing	5,162,758	6,966,864	0.0327	0.0365	0.0209
d_facebookage_missing	5,162,758	6,966,864	0.0326	0.0363	0.0207
d_country_prediction_NA	5,162,758	6,966,864	0.0332	0.0370	0.0206
d_female_missing	5,162,758	6,966,864	0.0326	0.0363	0.0206
d_ram_class_NA	5,162,758	6,966,864	0.0979	0.1040	0.0201
d_facebookage	5,162,758	6,966,864	3523.3626	3487.8436	−0.0193
d_profilepicturepresent	5,162,758	6,966,864	0.8730	0.8680	−0.0151
d_age	5,162,758	6,966,864	43.6287	43.4082	−0.0138
d_subscription_count	5,162,758	6,966,864	256.1537	251.3377	−0.0106
d_ram_class_4+gb	5,162,758	6,966,864	0.4792	0.4741	−0.0102
d_overall_app_sessions_28d	5,162,758	6,966,864	313.4411	309.9755	−0.0099
d_overall_app_sessions_7d	5,162,758	6,966,864	78.0468	77.1766	−0.0098
d_friend_count	5,162,758	6,966,864	396.2351	390.7594	−0.0093
d_num_friendships_initiated	5,162,758	6,966,864	176.4610	174.0463	−0.0090
Summary across all 86 demographic covariates:					
Max $\lvert\widehat{\Delta}_j\rvert$					0.0445
Mean $\lvert\widehat{\Delta}_j\rvert$					0.0077
Median $\lvert\widehat{\Delta}_j\rvert$					0.0036
Count with $\lvert\widehat{\Delta}_j\rvert > 0.10$ (Stuart, 2010)					0/86
Count with $\lvert\widehat{\Delta}_j\rvert > 0.25$					0/86
Joint orthogonality F-test				$F(84,\ 12{,}129{,}537) = 95.66,\ p < 0.001$	
R^2 from $W = \alpha + X'\beta + \varepsilon$					0.0007

Notes. Subscripts 0 and 1 denote the control ($W = 0$) and treatment ($W = 1$) arms. $\widehat{\Delta}_j$ is the normalized difference for covariate j defined in Equation (1), following Imbens and Rubin (2015, §14.2). The joint F-test rejects exact equality of means at this sample size, but the R^2 of the linear-probability regression of W on X indicates that demographics carry essentially no information about treatment assignment. The table lists the 20 covariates with the largest absolute normalized differences; summary statistics in the lower panel are computed over all 86 demographic covariates.

B Tests of Pooling Across Treatment and Control Arms

The survey outcomes are collected across five treatment arms and four control arms (denoted by different `version_ids`). The platform/ads outcomes are observed in a single (treatment, control) randomization, so no pooling test is required for those metrics. For each survey outcome Y we estimate the arm-saturated regression

$$Y_i = \sum_{v \in \mathcal{V}} \gamma_v \cdot \mathbf{1}[\texttt{version_id}_i = v] + u_i, \tag{3}$$

where $\mathcal{V}$ is either the set of treatment arms (for the treatment-pooling test) or the set of control arms (for the control-pooling test), and report the F-statistic for the null hypothesis $\gamma_v = \gamma_{v'}$ for all $v, v' \in \mathcal{V}$.

Pooling the treatment arms. Table A3 reports the F-tests of equality of conditional means across the five LPA treatment arms for each of the 9 survey outcomes. None of the outcomes reject. We note that with per-arm survey response counts typically in the hundreds, these tests are low powered. Our additional robustness checks (e.g., rerunning all main analyses with the pre-treatment covariates) help mitigate concerns.

Pooling the control arms. Table A4 reports the analogous F-tests across the four control arms. None of the 9 survey outcomes rejects at the nominal 5% level.

Willingness to accept (WTA). The WTA-elicitation sample is split across the same five LPA arms and four PA arms. We test equality of the arm-specific median WTA values within each survey-arm condition via a bootstrap χ^2 Wald test (1,000 replications). For the LPA arms (total N = 11,817), the test does not reject (W = 0.403, df = 4, p = 0.982). For the PA arms (total N = 11,727), the test again does not reject (W = 1.962, df = 3, p = 0.580).

Table A3: Test of pooling across treatment arms—survey outcomes.

Outcome	$\hat{\tau}_{\text{pooled}}$	F-stat	p-value
`satisfied`	−0.1186	1.9416	0.1006
`entertaining_ad`	−0.0279	1.3919	0.2345
`pad_load`	−0.0209	1.1518	0.3306
`data_control`	0.0203	0.9830	0.4157
`dissatisfied`	0.0254	0.4790	0.7512
`disruptive`	0.0259	0.3732	0.8278
`relevant_ad`	−0.0841	0.3328	0.8560
`comfort_data`	−0.0453	0.1982	0.9393
`like_ad`	−0.1396	0.1902	0.9436
Summary:			
Count with $p < 0.05$			0 / 9
Count with $p < 0.05/9$ (Bonferroni)			0 / 9

Notes. For each survey outcome, we report the F-statistic and p-value for the joint test of equality of the treatment-arm-specific means in Equation (3), $H_0 \colon \gamma_v = \gamma_{v'}$ for all v, v' in their version IDs. $\hat{\tau}_{\text{pooled}}$ is the pooled treatment effect (treatment versus pooled control). Rows ordered by F-statistic. The total sample size is $N = 12{,}129{,}622$ (5,162,758 control, 6,966,864 treated).

Table A4: Test of pooling across control arms—survey outcomes.

Outcome	Active arms	$\bar{Y}_{\text{ctrl}}$	F-stat	p-value
`entertaining_ad`	3	0.2392	1.9159	0.1484
`relevant_ad`	3	0.3279	1.6279	0.1976
`comfort_data`	3	0.1632	1.5635	0.2106
`satisfied`	4	0.6794	0.7060	0.5483
`data_control`	3	0.4334	0.6971	0.4985
`dissatisfied`	4	0.1030	0.6283	0.5967
`disruptive`	3	0.7566	0.4106	0.6638
`pad_load`	3	0.7139	0.1739	0.8405
`like_ad`	3	0.2719	0.1214	0.8857
Summary:				
Count with $p < 0.05$				0 / 9
Count with $p < 0.05/9$ (Bonferroni)				0 / 9

Notes. For each survey outcome, we report the F-statistic and p-value for the joint test of equality of the control-arm-specific means in Equation (3), $H_0 \colon \gamma_v = \gamma_{v'}$ for all $v, v' \in \{0, 53551, 53826, 53998\}$. "Active arms" counts the number of control arms in which at least one user responded to the item; for seven of the nine retained items, control arm 0 did not collect a response and is excluded from the joint test (see column note in the underlying R output). $\bar{Y}_{\text{ctrl}}$ is the overall control mean of the outcome. Rows ordered by F-statistic. Total N across the four control arms is 5,162,758.

C Survey Materials and Scope

Our research comprises survey responses from three separate sets of surveys sent to Facebook users, including: (i) a survey to elicit users' willingness to accept (WTA) payment to deactivate Facebook for a month, (ii) a survey to capture users' perceptions about their ad environment on Facebook, and (iii) a survey to obtain users' satisfaction after having made a purchase from an ad on Facebook.

C.1 Survey Invitations

The invitations to participate in both the WTA and ad perceptions surveys were shown to respondents at the top of their Facebook newsfeeds. The respondent user could choose to open the survey invitation, and thereafter begin the survey, by clicking on the "Start Survey" option.

For the WTA survey, the survey invitation specifically conveyed to users that they had an opportunity to receive a payment if they took the survey. For the ad perceptions survey, there was no such payment incentive and, instead, users were simply shown the survey invitation asking if they would like to share their opinion with Meta.

The post-purchase satisfaction survey had no separate survey invitation and was delivered to users directly on the Facebook newsfeed.

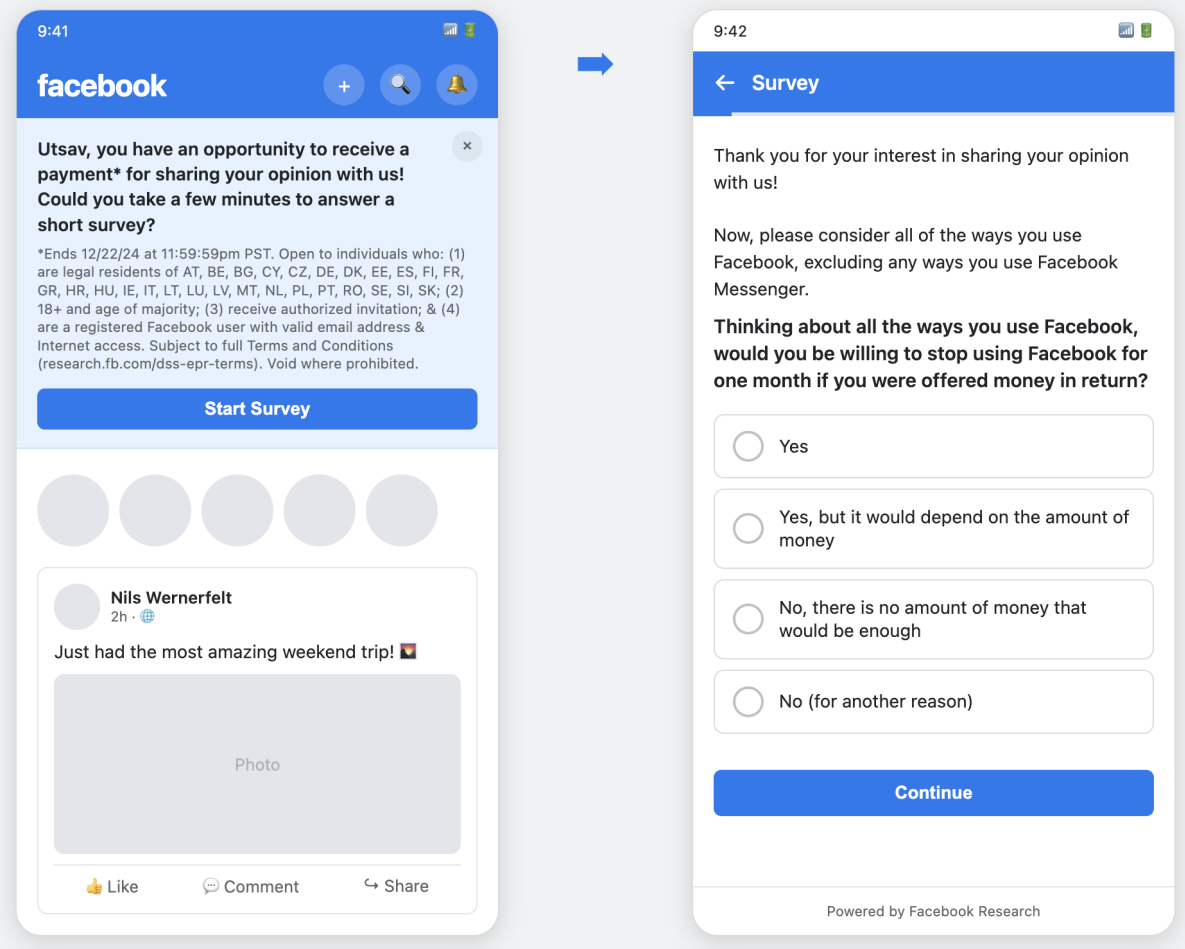


Figure A1: Survey invitation for the Willingness to Accept survey with payment incentive (left) and the first page of the survey shown after the user chooses to "Start Survey" (right).

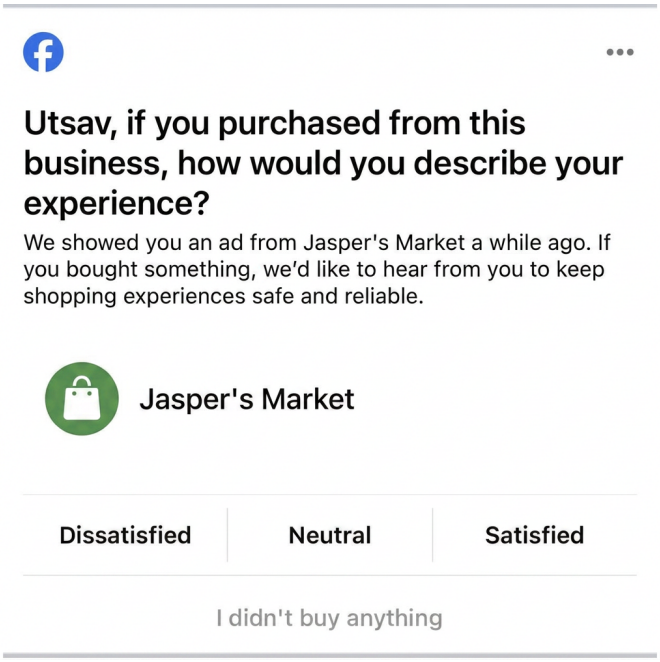
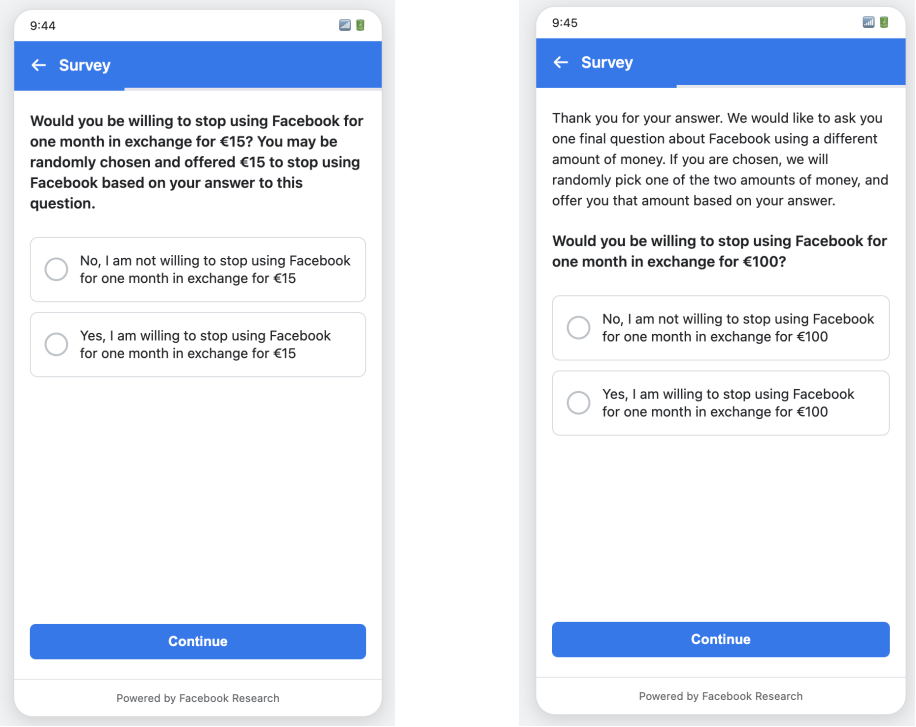


Figure A2: Post-purchase satisfaction survey delivered to users directly on their Facebook newsfeed.

C.2 Survey Questions

Our WTA survey consisted of two core questions and follows a similar structure as Brynjolfsson et al. (2025). These core questions were shown to users after the user had opened the survey invitation and completed the screener question (see Figure A1 above).

First, the user was asked if they would be willing to stop using Facebook for one month in exchange for X amount of money (see Figure A3 below). X was chosen randomly from a set of nine monetary values ranging from \$5 to \$100. Then, if the user answered "yes," they were asked about their willingness to accept Y to deactivate Facebook, where Y was less than X. If the user instead answered "no," they were asked about their willingness to accept Z to deactivate Facebook, where Z was higher than X. The initial set of offers ranged from \$5 to \$100, whereas the second offers ranged from \$1 to \$500.



(a) Initial WTA question. **(b)** Follow-up WTA question.

Figure A3: The two core questions of the Willingness to Accept survey, shown to users after they completed the screener question. The left panel shows the initial question with a randomly drawn offer; the right panel shows the follow-up question, in which the offer is adjusted upward (if the user rejected the initial offer) or downward (if the user accepted it).

As with Brynjolfsson et al. (2025), the respondents were given offers in their own currencies. Moreover, as these offers were incentivized, respondents had to agree to a set of terms and conditions consistent with local requirements in their countries. Each of the two choices

made by respondents to reject or accept an offer had a 2% chance of being selected. The actual probability was not known to participants.

For our ad perceptions survey, we asked active Facebook users in our experiment the questions reported in Table A5.

Finally, as we describe above, our purchase satisfaction survey consists of a single question where we ask respondents: *"[Respondent Name], if you purchased from this business, how would you describe your experience?"* with the following response options: Dissatisfied, Neutral, Satisfied, I didn't buy anything.

C.3 Survey Response Rates

Our core incentivized willingness-to-accept estimates are obtained using single-binary discrete-choice experiments using survey responses from 23,544 users. The survey sample consisted of Facebook users that were in one of the two experimental arms of this research and were active in December 2024 when the survey was conducted.

The observed response rates across different stages of the survey funnel suggest that respondents' behavior was consistent across the LPA and PA arms. At the upper end of the funnel, of those who were invited to take the survey, 7.38% and 7.21% of users opened the survey in LPA and PA, respectively. Similarly, of those who opened the survey, 69.27% and 68.46% of users started the survey in LPA and PA, respectively. While the differences are small in magnitude, they are statistically significant, likely due to the large sample sizes. Finally, of those who started the survey, 40.0% and 40.1% of users completed the survey in LPA and PA, respectively. Reassuringly, this difference is not statistically significant ($p = 0.70$).

For our ad-perceptions survey, 2,111 users answered at least one question, with the highest number of responses being recorded for the ad-recall question, which was the first question shown to users who chose to start the survey.

As with the WTA survey, respondents' behavior was consistent across LPA and PA. At the upper end of the funnel, of those who were invited to take the survey, 5.48% and 5.58% of users opened the survey in LPA and PA, respectively. Similarly, of those who opened the survey, 72.01% and 69.86% of users started the survey in LPA and PA, respectively. Finally, of those who started the survey, 58.9% and 61.6% of users completed the survey in LPA and PA, respectively. In each of these comparisons, the differences in the rates across LPA and PA were not statistically significant ($p = 0.62$, 0.23, and 0.25, respectively).

Finally, for the purchase satisfaction survey, we obtained responses from 14,732 users across LPA and PA. Unlike the aforementioned surveys, we cannot estimate full-funnel re-

sponse rates for this survey as users are not required to open a survey invitation. Rather, the survey appears as an "impression" on their Facebook newsfeed. We find that the share of users who responded to the survey conditional on viewing the impression was about 8.0% in LPA and 8.6% in PA, and the difference was not statistically significant ($p = 0.39$), assuaging any concerns of differential response rates.

Table A5: Ad-perceptions survey items: measure, question text, and response options.

Measure	Question	Response options
Ad disruptiveness	Thinking about your last visit to Facebook, how often did ads interrupt what you were trying to do?	Never; Rarely; Occasionally; Often; Very Often; I can't remember.
Perceived ad load	Thinking about your last visit to Facebook, how frequently did you see ads?	Never; Rarely; Occasionally; Frequently; Very Frequently; I can't remember.
Ad likeability	How much do you like or dislike this ad? *(The question included an image of a randomly chosen ad that the user had recently seen on Facebook.)*	Strongly dislike; Somewhat dislike; Neither like nor dislike; Somewhat like; Strongly like.
Ad relevance	How relevant to you is this ad? *(The question included an image of a randomly chosen ad that the user had recently seen on Facebook.)*	Not at all relevant; Slightly relevant; Moderately relevant; Very relevant; Extremely relevant.
Ad entertainability	How entertaining do you find this ad? *(The question included an image of a randomly chosen ad that the user had recently seen on Facebook.)*	Not at all entertaining; Slightly entertaining; Moderately entertaining; Very entertaining; Extremely entertaining.
Comfort with ad data use	Facebook uses some of your personal information to show you ads. How comfortable or uncomfortable are you with the information Facebook may have used to select this ad for you? *(The question included an image of a randomly chosen ad that the user had recently seen on Facebook.)*	Very uncomfortable; Somewhat uncomfortable; Neither comfortable nor uncomfortable; Somewhat comfortable; Very comfortable.
Control over ad-targeting data	How much influence do you feel like you have over the personal information Facebook uses to show you ads?	None at all; Not much; A moderate amount; A lot; A great deal.

D Individual-Metric Causal Forests

Here we provide the output from causal forest analyses for each of our log-based outcome variables. Given their sample sizes, our survey-based and WTA measures were underpowered for such analyses. For each metric, we report three sets of outcomes:

1. The AUTOC statistic of Yadlowsky et al. (2025), computed with the forest's CATE estimates as the prioritization rule. A two-sided z-test of H_0: AUTOC $= 0$ tests whether the forest's prioritization carries any signal about realized heterogeneity.

2. The differential-prediction coefficient from the calibration test of Chernozhukov et al. (2018). A coefficient indistinguishable from zero indicates no detectable heterogeneity captured by the forest; a coefficient indistinguishable from one indicates that the forest's CATE estimates are correctly scaled to the realized heterogeneity in the held-out half of the data.

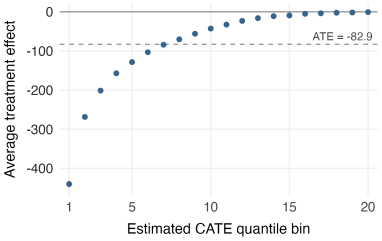
3. The 20-bin CATE-quantile plot: the held-out test sample is sorted by predicted CATE and split into twenty equally-sized bins. Within each bin we plot the AIPW estimate of the average treatment effect with 95% confidence intervals. The solid horizontal line marks zero; the dashed horizontal line marks the AIPW estimate of the pooled ATE.

Two facts to keep in mind when reading the panels. First, every panel is rendered on the metric's own scale, so the vertical axes are not comparable across panels. The within-panel question—"does the CATE distribution lie on one side of zero, and how much does it spread?"—is the comparable one. Second, the AUTOC z-test is highly powered at our sample sizes; small but precisely estimated differences across bins yield p-values close to zero even when the practical magnitude of heterogeneity is modest. The differential-prediction calibration coefficient is therefore a more useful summary of the importance of heterogeneity: values close to one indicate that the predicted ranking translates cleanly to a proportionally ranked held-out effect.

What the panels show. The patterns documented in Section 5.2 for ad clicks and ad x-outs generalize. For the ad-engagement-good outcomes (clicks, likes, comments, shares, page opens, page likes, onsite conversions), every CATE bin lies below zero and the differential-prediction coefficient is close to one: there is real, well-calibrated heterogeneity, but no subset of users responds favorably. For the ad-engagement-bad outcomes (reports, x-outs, irrelevant x-outs, repetitive x-outs), at most the bottom one or two bins fall below zero, and the magnitude of the bottom bins is small relative to the top bins; loosely, LPA reduces

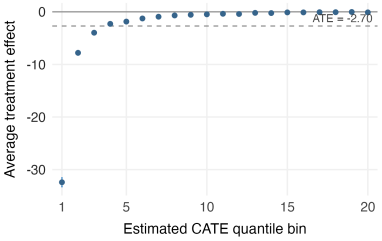
dismissals for a minority of users, but the absolute size of the improvement is small. For the platform-engagement outcomes the picture is qualitatively different: the bin point estimates straddle zero and the per-panel AUTOC magnitudes are modest—consistent with the near-zero average effects documented in Section 5.1.

Ad clicks



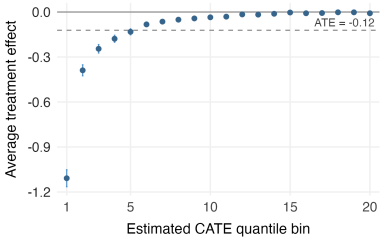
AUTOC = 63.48 (0.263), p = <0.001
Cal. = 1.03 (0.0058), p = <0.001

Ad likes



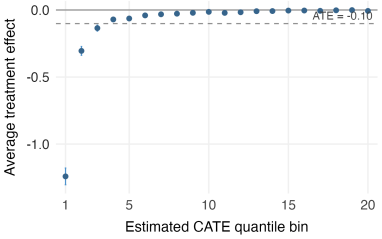
AUTOC = 2.37 (0.026), p = <0.001
Cal. = 1.06 (0.013), p = <0.001

Ad comments



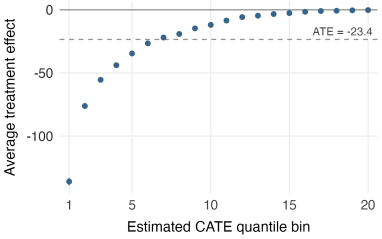
AUTOC = 0.101 (0.0021), p = <0.001
Cal. = 1.14 (0.026), p = <0.001

Ad shares



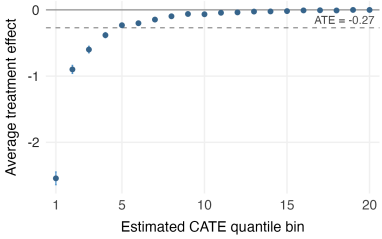
AUTOC = 0.089 (0.0021), p = <0.001
Cal. = 1.10 (0.026), p = <0.001

Ad page opens



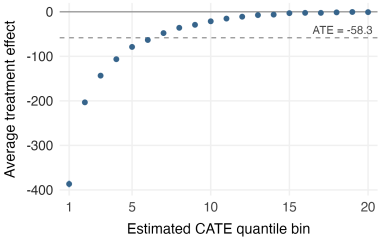
AUTOC = 18.18 (0.098), p = <0.001
Cal. = 1.03 (0.0074), p = <0.001

Ad page likes



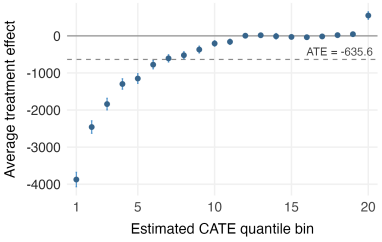
AUTOC = 0.232 (0.0035), p = <0.001
Cal. = 1.16 (0.020), p = <0.001

Onsite conversions



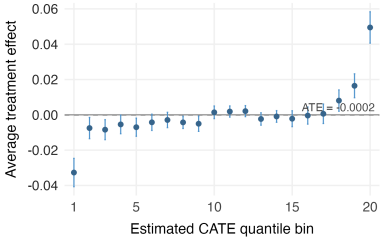
AUTOC = 47.22 (0.238), p = <0.001
Cal. = 1.03 (0.0066), p = <0.001

Ad impressions



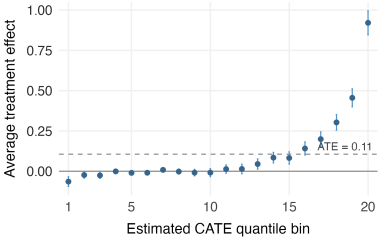
AUTOC = 665.3 (13.85), p = <0.001
Cal. = 1.13 (0.020), p = <0.001

Ad reports



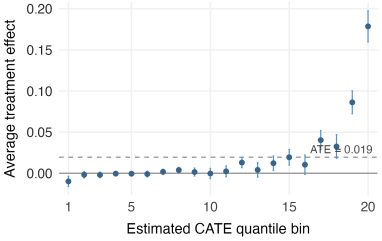
AUTOC = 0.014 (0.0010), p = <0.001
Cal. = 1.53 (0.108), p = <0.001

Ad x-outs



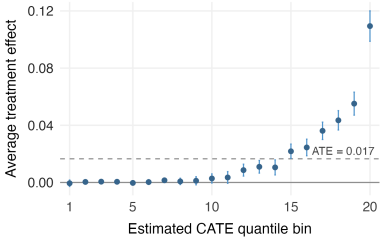
AUTOC = 0.225 (0.0068), p = <0.001
Cal. = 1.38 (0.050), p = <0.001

Irrelevant x-outs



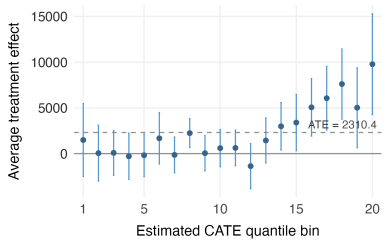
AUTOC = 0.041 (0.0019), p = <0.001
Cal. = 1.35 (0.070), p = <0.001

Repetitive x-outs



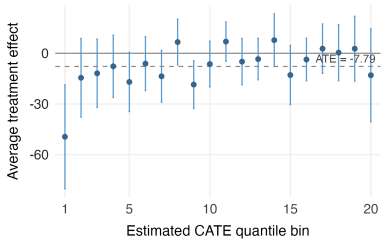
AUTOC = 0.027 (0.0010), p = <0.001
Cal. = 1.23 (0.050), p = <0.001

Time spent



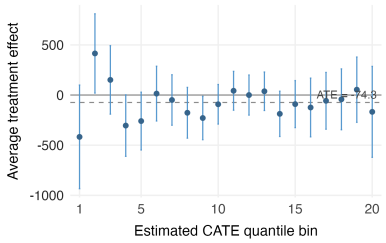
AUTOC = 2553.7 (579.7), p = <0.001
Cal. = 0.361 (0.073), p = <0.001

Sessions



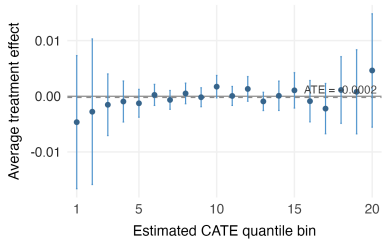
AUTOC = 2.68 (2.87), p = 0.350
Cal. = 0.045 (0.075), p = 0.272

Total VPVs



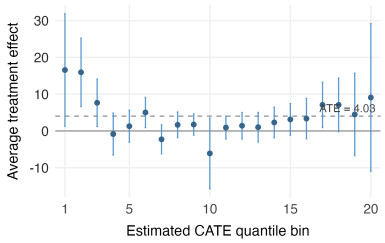
AUTOC = 8.54 (46.19), p = 0.853
Cal. = 0.074 (0.077), p = 0.167

Active user



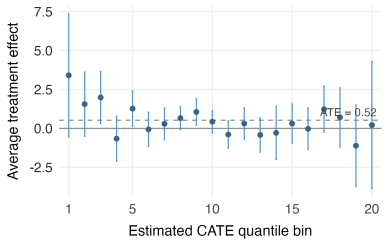
AUTOC = 0.0009 (0.0011), p = 0.389
Cal. = -0.180 (0.289), p = 0.734

Likes



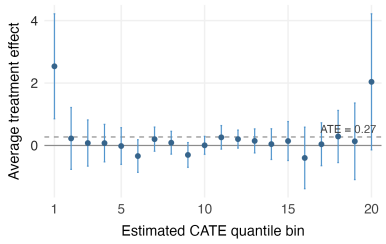
AUTOC = 0.696 (1.99), p = 0.727
Cal. = 0.263 (0.112), p = 0.010

Comments



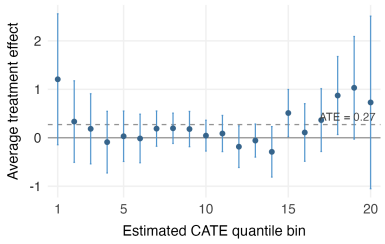
AUTOC = -0.370 (0.461), p = 0.422
Cal. = 0.015 (0.118), p = 0.449

Reshares



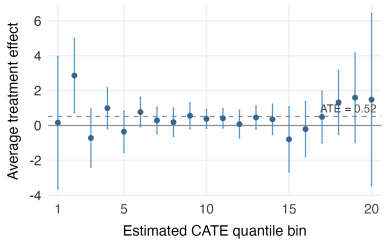
AUTOC = 0.138 (0.209), p = 0.509
Cal. = 0.043 (0.132), p = 0.371

Posts



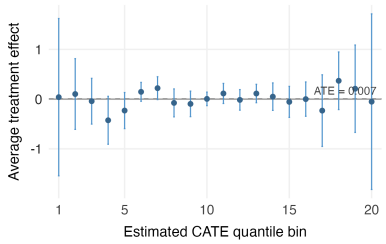
AUTOC = 0.032 (0.173), p = 0.855
Cal. = -0.215 (0.118), p = 0.966

Reactions: Love



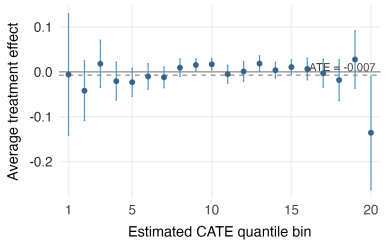
AUTOC = 0.217 (0.498), p = 0.663
Cal. = -0.030 (0.129), p = 0.592

Reactions: Haha



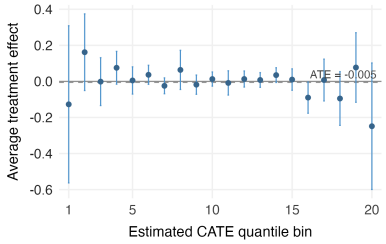
AUTOC = 0.069 (0.184), p = 0.707
Cal. = -0.012 (0.153), p = 0.530

Reactions: Anger



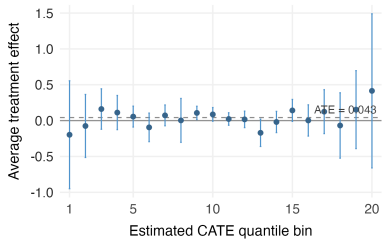
AUTOC = -0.021 (0.014), p = 0.121
Cal. = 0.060 (0.259), p = 0.408

Reactions: Sorry



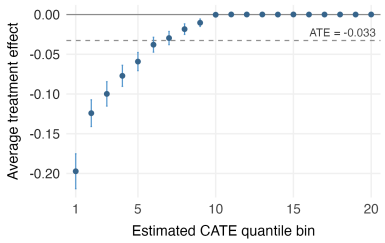
AUTOC = -0.042 (0.039), p = 0.281
Cal. = -0.082 (0.157), p = 0.700

Reactions: Support



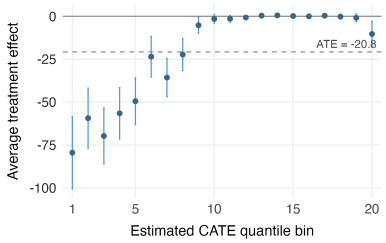
AUTOC = 0.049 (0.102), $p = 0.631$
Cal. = -0.025 (0.147), $p = 0.568$

Long-term purchases



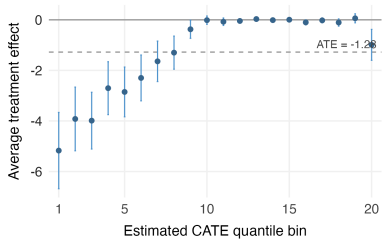
AUTOC = 0.028 (0.0009), $p = <0.001$
Cal. = 1.12 (0.041), $p = <0.001$

LT conv. (attributed)



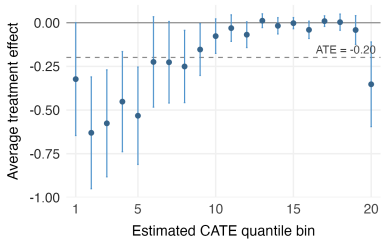
AUTOC = 14.80 (1.17), $p = <0.001$
Cal. = 0.647 (0.055), $p = <0.001$

LT purchases (attributed)



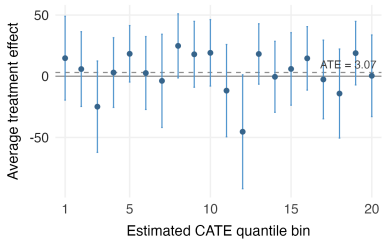
AUTOC = 0.818 (0.096), $p = <0.001$
Cal. = 0.596 (0.063), $p = <0.001$

LT purchases (all)



AUTOC = 0.076 (0.028), $p = 0.007$
Cal. = 0.582 (0.097), $p = <0.001$

LT avg. price per good



AUTOC = -0.481 (3.99), $p = 0.904$
Cal. = -2.54 (1.02), $p = 0.994$

E Supplemental Analyses

E.1 Match Quality versus Ad Quality

Average treatment effects on ad engagement under LPA could in principle reflect two distinct mechanisms. The first is a change in *match quality*: with less personal information available, the platform is less able to select the ad that best matches a given user's preferences from the pool of available ads, even if the underlying pool of ad creatives is unchanged. The second is a change in *ad quality*: if LPA users are systematically shown ads with different qualities, the documented engagement differences could reflect this compositional shift rather than a personalization mechanism.

To assess the second channel, we evaluate the entire population of ads shown during the experimental window using a panel of automated image-quality scores. The scores are produced by a model that assigns each ad image a continuous score on dimensions including aesthetics, lighting, compositional structure, color harmony, depth of field, and a Mean Opinion Score (MOS) trained on human ratings. For each feature, we compute the impression-weighted mean over the ads that ran in that arm and the absolute standardized difference between arms. Table A6 reports the results.

Given the large sample size, the differences are all statistically significant. However, the magnitudes of the differences are very small: no LPA-vs-PA difference exceeds about one-tenth of a pooled standard deviation. The largest absolute standardized difference across the nine features is 0.070 (lighting quality score); all other features have $|d| < 0.06$. By the conventional balance thresholds of Imbens and Rubin (2015) and Stuart (2010), neither $|d| > 0.10$ nor $|d| > 0.25$ is exceeded for any feature: the mean ad quality is very similar across arms.

This suggests that the differences in user responses we observe are not due to differences in ad quality but differences in match quality (i.e., personalization).

Table A6: Ad image-quality features by treatment arm.

Feature	Coverage (%)	N (ads)	Range (p10 → p90)	Mean LPA	Mean PA	Diff	$\lvert d \rvert$
Aesthetic score	46.94	6.90M	2.16 → 2.57	2.353	2.355	−0.002	0.015
Lighting quality score	46.94	6.90M	0.85 → 0.97	0.916	0.912	0.005	0.070
Rule of thirds composition score	46.94	6.90M	0.85 → 0.95	0.906	0.904	0.002	0.060
Compositional balance of visual elements score	46.94	6.90M	0.89 → 0.95	0.925	0.924	0.001	0.056
Color vividness and saturation score	46.86	6.88M	0.23 → 0.63	0.451	0.446	0.006	0.036
Color harmony score	46.94	6.90M	0.56 → 0.77	0.670	0.669	0.002	0.021
Depth of field score	39.78	5.85M	0.04 → 0.34	0.179	0.177	0.002	0.017
Image quality	4.37	641K	0.06 → 0.25	0.149	0.146	0.003	0.037
Mean Opinion Score	4.35	639K	60 → 79	70.150	70.410	−0.255	0.034

Notes. Each row is an automated image-quality feature scored on the population of ads that ran during the experimental window in the treated and control arms of experiment 12263. *Coverage (%)* is the share of ads in the experimental window for which the feature could be computed (some features require specific image attributes); *N (ads)* is the corresponding ad count. *Range (p10 → p90)* reports the interdecile range of the feature pooled across arms, to convey the scale of the score. *Mean LPA* and *Mean PA* are the per-arm means. *Diff* is the LPA arm mean minus the PA arm mean; $|d|$ is the absolute standardized mean difference (Cohen's d), computed as |Mean LPA − Mean PA| divided by the pooled standard deviation of the feature. By the convention of Imbens and Rubin (2015) and Stuart (2010), an absolute standardized difference below 0.10 is regarded as evidence that the two distributions are balanced; every feature in the table satisfies this threshold, with the largest $|d|$ equal to 0.070 (lighting quality score).

Below we provide the variable names and definitions. Given these are the outputs of internal classifiers, we do not want to lean heavily on these findings. We take them as suggestive that ad quality did not change across LPA and PA conditions.

- **Aesthetic score**: Overall aesthetic-quality rating from a model-scored aesthetic pipeline;

combines the sub-scores below into a holistic number (ranging from 2.0–2.6 in our data; scale is model-specific).

- **Compositional balance of visual elements score**: Compositional balance of visual elements within the frame; how evenly weight is distributed across the image (ranges from 0 to 1; higher means better balance).
- **Color harmony score**: Color-harmony score capturing how well the ad's colors complement each other on a perceptual color wheel (ranges from 0 to 1).
- **Color vividness and saturation score**: Color vividness and saturation score (0–1; higher means more vivid colors).
- **Depth of field score**: Depth-of-field score capturing foreground and background separation and focus (0–1; higher means more pronounced shallow depth of field, "professional photo" look).
- **Lighting quality score**: Lighting-quality score capturing exposure, contrast, and lighting cleanliness (0–1; higher means better lighting).
- **Rule of thirds composition score**: Rule-of-thirds composition score; how well the image follows classical rule-of-thirds positioning of subjects (0–1).
- **Image quality**: Mean of 24 perceptual-quality components. This score is output from an image-quality model that captures naturalness and sharpness; lower values indicate higher perceived quality.
- **Mean Opinion Score**: a learned predictor of perceived image quality on a roughly 0–100 scale (higher means better perceived quality).

E.2 Who responds most, and who responds least?

We use the individual-metric causal forests from the platform/ads sample to characterize the users whose estimated treatment effects sit at the two extremes of the CATE distribution. Following the procedure of Chernozhukov et al. (2018), we partition the held-out test sample into 20 equally-sized bins according to the out-of-bag CATE estimate, and then plot, separately for the bottom and top bins, the empirical distribution of the covariates that the forest selects as the strongest predictors of heterogeneity (the variables entering the best-linear projection of the CATE). Figure A4 shows the distributions for ad clicks; Figure A5 shows the analogous distributions for ad x-outs (i.e., user-initiated dismissals of ads).

Ad clicks. The forest predicts that the treatment-induced change in clicks is most negative for users with high pre-treatment ad clicks, ad-page opens, ad impressions, onsite revenue, time spent, and total VPVs, and is closest to zero for users with low values of these same covariates (Figure A4). The Low-CATE group—the users whose clicks fall the most under LPA—is a high-engagement, high-revenue group of intensive ad-clickers. The High-CATE group—the users for whom LPA has essentially no effect on clicks—is a low-engagement group: their distributions of pretreatment ad clicks, ad-page opens, ad impressions, on-site revenue, and platform time spent are heavily concentrated near zero. These users are marginal: they click on essentially no ads in the pretreatment period, so the absolute change induced by switching off personalization is mechanically bounded near zero.

Ad x-outs. A symmetric pattern emerges for ad x-outs (Figure A5). The High-CATE group—the users whose x-outs fall the least under LPA, i.e., for whom LPA *reduces* x-outs by the smallest amount—is again a low-engagement group, with pretreatment ad clicks, ad page opens, time spent, onsite revenue, and subscription counts concentrated near zero. Because these users x-out very few ads to begin with, the mechanical scope for LPA to reduce x-outs is small. Two additional covariates load on heterogeneity in x-outs but not in clicks: account age and user age, both of which are higher in the High-CATE group. The Low-CATE group—the users for whom LPA reduces x-outs the most—are younger, more recently-joined, higher-engagement users with substantial pretreatment ad activity.

Implication. Taken together, the two figures suggest the heterogeneity in treatment effects on both clicks and x-outs is concentrated among engaged, "intensive-margin" users, while the less engaged users tend to have smaller treatment effects.

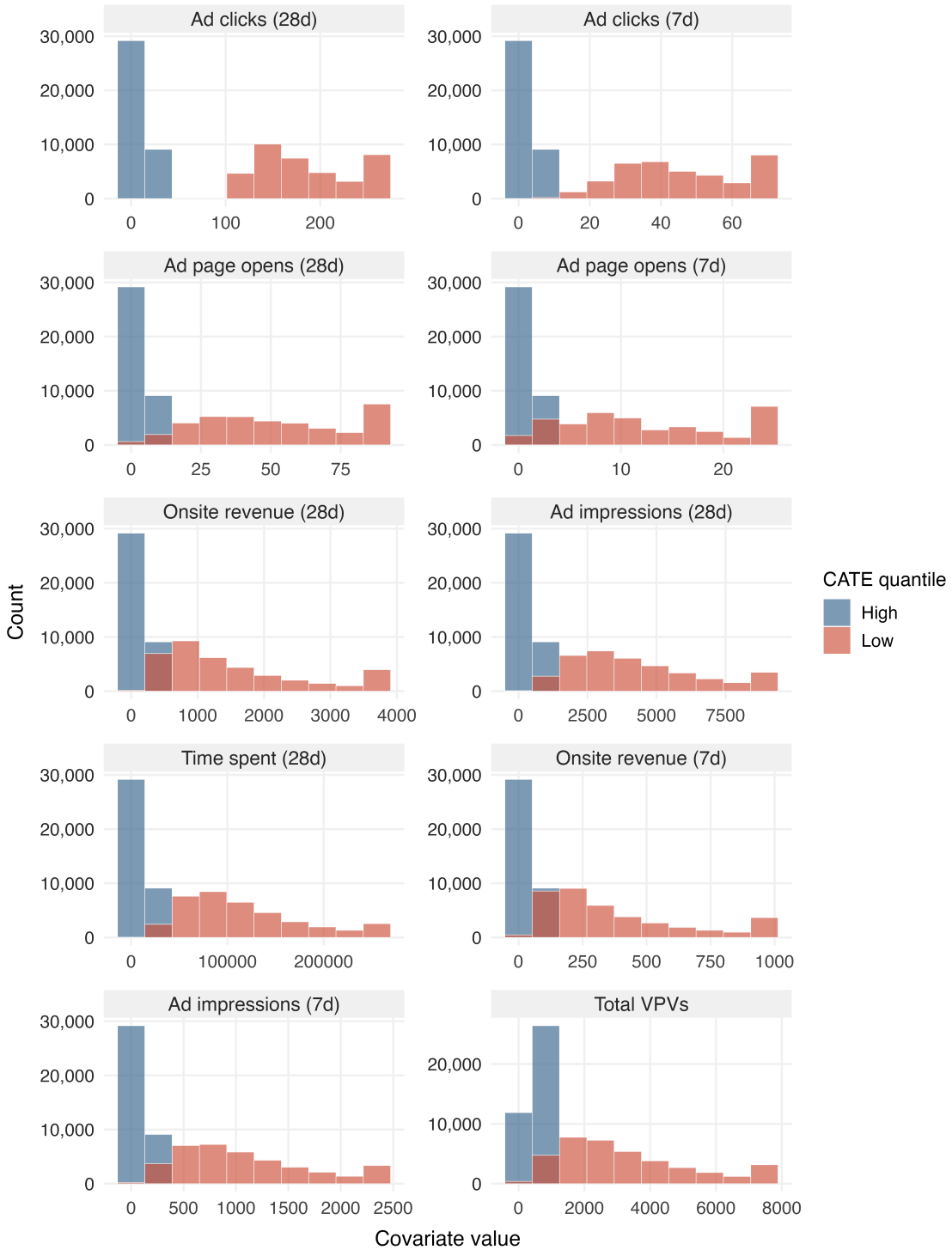


Figure A4: Distribution of the top forest-selected covariates among users in the bottom ("Low") and top ("High") CATE quantile bins for the outcome *ad clicks*. Each panel restricts to the bottom and top 1/20 of the held-out test sample, ranked by the out-of-bag CATE estimate from a causal forest fit on the platform/ads sample.

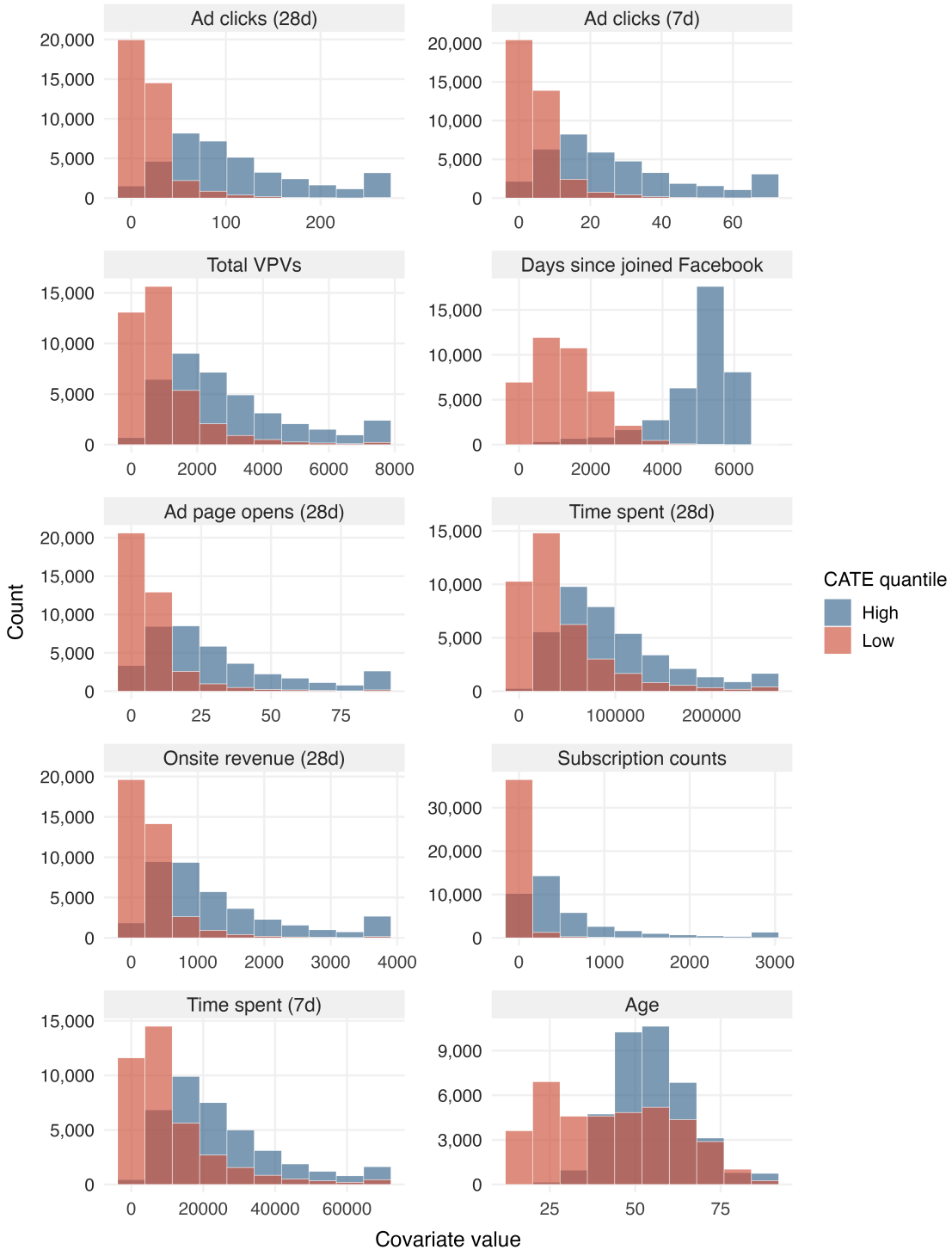


Figure A5: Distribution of the top forest-selected covariates among users in the bottom ("Low") and top ("High") CATE quantile bins for the outcome *ad x-outs*. Construction is identical to Figure A4.