# Long or Short?
# Personalizing Ad Length and Frequency

Ali Goli
University of Rochester

David Reiley
UC Berkeley

Jonas Tungodden
Norwegian School of Economics

Wenfeng Qiu
Sirius XM Holdings Inc.

**Abstract**

We study the economics of advertising length using a large-scale randomized experiment on 18.3 million Pandora listeners. The experiment promoted a listening feature and randomly assigned users to advertising campaigns that varied in ad format (10 vs. 30 seconds) and frequency. For both users unfamiliar with the feature and those already familiar, we document that both ad formats generate incremental effects, with the relative advantage of long over short ads substantially larger among unfamiliar users. We then build a machine learning model to capture heterogeneity in both responsiveness to the campaign and realized exposure levels, and conduct off-policy evaluation of alternative campaign designs. We find that a personalized policy that optimally assigns users across the available experimental conditions, choosing the ad format and frequency for each user, increases incremental ad effects by about 30 percent. Two-thirds of this efficiency gain comes from leveraging heterogeneity in exposure rather than heterogeneity in responsiveness to the campaign.

**JEL Codes:** M31, M37, D83, C93

**Keywords:** field experiments, advertising effectiveness, ad length, personalization, heterogeneous treatment effects, machine learning, off-policy evaluation

**Acknowledgments:** We wish to thank Megan Enright, Charlie Kuller, Eli Villanueva, and Ai Xia for their collaboration on the design and implementation of the experiment. We thank Alex Corrigan, Emi Giannella, Rick Handt, Simha Mummalaneni, Erling Skancke, Semyon Tabanakov, Steve Tadelis, and Bertil Tungodden for helpful comments. Author contacts: agoli@ur.rochester.edu, david@davidreiley.com, jonas.tungodden@gmail.com, Wenfeng.Qiu@siriusxm.com.

Ad-supported streaming media platforms manage advertising exposure using two primary levers: how often ads are delivered and how long each exposure lasts. These choices determine a consumer's total ad load, a constraint that platforms manage not only for revenue but also for user experience. Each additional ad second competes with the media consumption time and can reduce engagement or increase churn. Despite the central role of ad length and frequency in practice, there is limited field evidence on how these levers interact, and in particular on whether longer ads produce larger incremental ad effects or whether multiple short exposures can substitute for a single long one.

A natural question arising from this gap is how different allocations of total exposure time—such as one 30-second ad versus three 10-second ads—affect advertising effectiveness. On the one hand, longer ads may be particularly valuable for users who are unfamiliar with the product because they convey more content per exposure. On the other hand, multiple short exposures could reinforce awareness more often and therefore be similarly or even more effective. A priori, it is unclear which allocation should perform better, and the answer may vary by user segment.[1] Empirically, these questions are difficult to study because they require detecting heterogeneity across formats and user segments, yet even estimating the average effects of advertising is challenging due to an array of issues including identity fragmentation, measurement error in impressions, noisy outcomes, and interference from concurrent campaigns.[2] These difficulties are magnified when studying heterogeneity in ad effects or comparing alternative formats, as both tasks require even more precise measurement.

This paper draws on a setting that offers the combination of precise measurement and exogenous variation needed for studying how ad length and frequency shape advertising effectiveness. We examine the promotion of a relatively unfamiliar listening feature on Pandora Internet Radio, where both ad exposure and downstream engagement occur on the same platform under persistent user identities and where the feature is neither promoted nor used outside of Pandora, eliminating cross-platform interference issues. This setting avoids many of the limitations that typically complicate advertising studies and allows us to observe the complete causal chain from ad delivery to feature

[1] Related concepts appear in classic discussions of the "informative" and "persuasive" roles of advertising (Ackerberg, 2001; Bagwell, 2007), but these frameworks do not speak directly to how format (i.e., ad length) should matter or how these effects vary with user awareness levels.

[2] For a few exmaples, please see Drèze and Hussherr (2003); Lewis and Reiley (2014); Lewis et al. (2015); Bounie et al. (2016); Shapiro et al. (2021); McGranaghan et al. (2022).

adoption. The experiment simultaneously randomizes ad length (10 versus 30 seconds) and ad frequency across more than 18 million listeners and tracks outcomes for up to two years, yielding rare field evidence on how length and frequency affect both short- and long-run advertising effects.

Our experiment yields three main findings. First, longer ads are substantially more effective on a per-ad and per-second basis: during the experimental period, long ads generate 4.6 times the response of short ads, far exceeding the 3:1 length ratio. Second, this advantage is driven primarily by uninformed users: among listeners unfamiliar with the advertised feature, long ads are 5.3 times more effective than short ads initially, growing to 7.0 times by two years later. Among experienced listeners, the relative effectiveness roughly matches the duration ratio, with effects that become statistically indistinguishable from zero within months. Third, long-term observation windows are important not only for understanding the effect of format but also for measuring the returns to advertising across user segments: evaluations limited to short-term outcomes would understate the returns to advertising among uninformed consumers while overstating them among informed ones.

To establish these findings, we build on the ghost ads framework (Johnson et al., 2017), which improves the statistical efficiency of measuring advertising effects by comparing exposed users to counterfactual "would-have-been-exposed" users identified through placebo ads. We apply an extended version of this approach that moves from the classic on/off treatment margin to the intensive margin of ad delivery (Johnson et al., 2016). Each listener is simultaneously enrolled in three parallel campaigns that mix promotional and placebo content of different lengths, which allows us to randomize both the intensity (number of exposures) and the format (short versus long ads). This design not only yields cleaner causal estimates of exposure but also allows us to study how length and frequency impact ad effectivenss, and whether tailoring these levers can reduce ad load without reducing impact.

Using this experimental variation, we conduct an off-policy analysis to understand how personalization could improve campaign performance. Our analysis highlights the importance of two distinct sources of heterogeneity in campaign optimization. First, platforms must identify which users are most responsive to advertising when exposed (heterogeneity in lift). Second, and less studied, platforms must model how much advertising each user naturally receives (heterogeneity in realized exposure intensity). This second dimension varies substantially across users: for instance,

heavy listeners encounter many ad opportunities while occasional users see few, creating systematic differences in treatment dosage even within the same campaign. We find that two-thirds of personalization gains come from modeling this exposure heterogeneity rather than from predicting response heterogeneity. This insight reframes optimal targeting strategy. Platforms should focus not only on who responds to ads but also on how intensively to treat each user.

Furthermore, personalization changes how platforms should interpret the trade-off between ad format (long vs. short) and frequency. In aggregate comparisons, short ads appear to underperform long ads on every dimension. But once the assignment is allowed to vary with user characteristics, a clearer pattern emerges. The incremental return to longer ad format is low for users who are both eligible to receive many impressions (i.e., high-exposure users) and are already familiar with the feature, making short ads the more efficient format for these segments. In contrast, unfamiliar users exhibit much larger marginal effects of longer versus shorter ads. Our personalized policy mirrors this heterogeneity: it assigns short ads primarily to high-exposure or familiar listeners while allocating long ads to unfamiliar ones, increasing the incremental ad effect by about 31 percent relative to a uniform campaign (from 10.79% to 14.12% lift). This reframes the role of short ads and shows how personalization can improve advertising efficiency without increasing user ad load.

These findings have important implications for the $40 billion audio advertising market (Statista, 2024) and speak directly to platform design and consumer welfare. Beyond the classic trade-off between reach and repetition, streaming platforms face an additional constraint: every second of ad load competes with listening time, risking lower engagement or churn. Our results demonstrate that personalizing ad length and frequency can reduce aggregate ad load while maintaining similar advertising effectiveness, potentially improving the listener experience without harming platform ad revenues. This provides a path toward reducing the tension between user experience and monetization in ad-supported platforms.

# Literature Review

Our work relates to three strands of research on advertising effectiveness, experimentation, and platform design. The discussion in this section is intended to be illustrative, highlighting representative studies that frame our contribution rather than providing a comprehensive review of the literature.

### *Advertising Frequency and Format*

Recent work examines how the delivery of advertising—through frequency, timing, and length—affects effectiveness.[3] For example, Sahni (2015) show that spacing exposures over time increases purchase response, and Sahni et al. (2019) quantify how frequency and timing shape repeat visits. Evidence on ad length is more limited in field settings. Most prior studies use laboratory experiments (Elsen et al., 2016; Holmes, 2021; Wang et al., 2020; Johnson et al., 2021) or non-experimental approaches (Peters and Bijmolt, 1997; Ge et al., 2021; Fossen et al., 2025).

Building on these insights, we directly randomize both ad length and frequency in a large-scale audio setting. Our results highlight distinct roles for different formats: long ads operate primarily as informative advertising for unfamiliar users, while short ads function more as persuasive or reminder advertising for those already familiar (Ackerberg, 2001; Bagwell, 2007). We then quantify the gains from personalizing exposure across formats to reduce overall ad load while maintaining effectiveness.

### *Experiments and Measurement in Digital Advertising*

Experiments have become essential for measuring advertising effects because observational methods often yield biased estimates (Blake et al., 2015; Gordon et al., 2019, 2023). Yet even large-scale tests face limitations. Lewis and Rao (2015) show that effects are often too small to detect reliably, and

[3] A complementary literature shows that ad content (creative, framing, personalization) itself has first-order causal effects on consumer response, for a few examples see Bertrand et al. (2010); Liaukonyte et al. (2015); Sahni et al. (2018); Biswas (2022); Morozov and Tuchman (2024); Kalyanam et al. (2025). Our long-short comparison should thus be read as a length-plus-content contrast typical of practice, rather than the effect of time alone.

Berman and Van den Bulte (2022) document high false discovery rates in industry A/B testing. Goldfarb and Tucker (2019) provide a comprehensive discussion on both the promise and the challenges of experimental approaches in digital markets.

The ghost ads framework (Johnson et al., 2017) partly addresses power concerns by moving from intent-to-treat to treatment-on-the-treated via construction of comparable counterfactual exposures, though exposure intensity remains endogenous and the comparison is still on/off. Johnson et al. (2016) first extended this approach to the intensive margin, enabling variation in treatment intensity while maintaining the ghost ads identification strategy. More recent work advances experimental design further. Hermle and Martini (2022) and Waisman and Gordon (2025) propose budget experiments or multicell experiments that recover marginal treatment effects, enabling analysis along intensive margins such as reach and spend. Our study contributes to this literature by extending the ghost-ads design in a manner similar to Johnson et al. (2016) and by creating orthogonal variation in both treatment intensity (frequency) and format (length) in a natural streaming environment. This design provides causal field evidence on the trade-offs between ad duration and frequency in driving ad effectiveness.

### *Platform Trade-Offs and Personalization*

Advertising in two-sided media markets acts as an implicit price that platforms must balance against user experience. Prior work formalizes these trade-offs (Rochet and Tirole, 2003; Anderson and Coate, 2005), and subsequent models examine how multihoming, ad-skipping technologies, and demographic variation in ad pricing shape equilibrium outcomes (Anderson et al., 2018; Tåg, 2009; Anderson and Gans, 2011; Gentzkow et al., 2024). A related empirical literature documents the negative impact of disruptive and obtrusive ads on media consumption (Goldstein et al., 2014; Goldfarb and Tucker, 2011), with Wilbur (2008) and Goli et al. (2025b) showing how ad load affects viewing and listening behavior in television and streaming audio. Beyond consumption effects, recent work also examines the welfare effects of online advertising using long-horizon experimental evidence (Brynjolfsson et al., 2025).

Personalization has been studied across many domains, including pricing (Rossi et al., 1996; Shiller et al., 2013; Dubé and Misra, 2023), promotions and email (Ansari and Mela, 2003; Sahni et

al., 2018; Hitsch et al., 2024), and mobile advertising (Rafieian and Yoganarasimhan, 2021; Rafieian, 2023). For a comprehensive review of personalization, targeting, and experimentation in marketing, see Lemmens et al. (2025); Rafieian and Yoganarasimhan (2023). In the context of streaming media and ad load, Goli et al. (2025a) show that personalizing ad load can improve subscription revenues without reducing advertising revenues. Our contribution is to demonstrate that personalization at the format level provides a new lever: platforms can lower aggregate ad load while preserving advertising effectiveness.

# Study Methodology

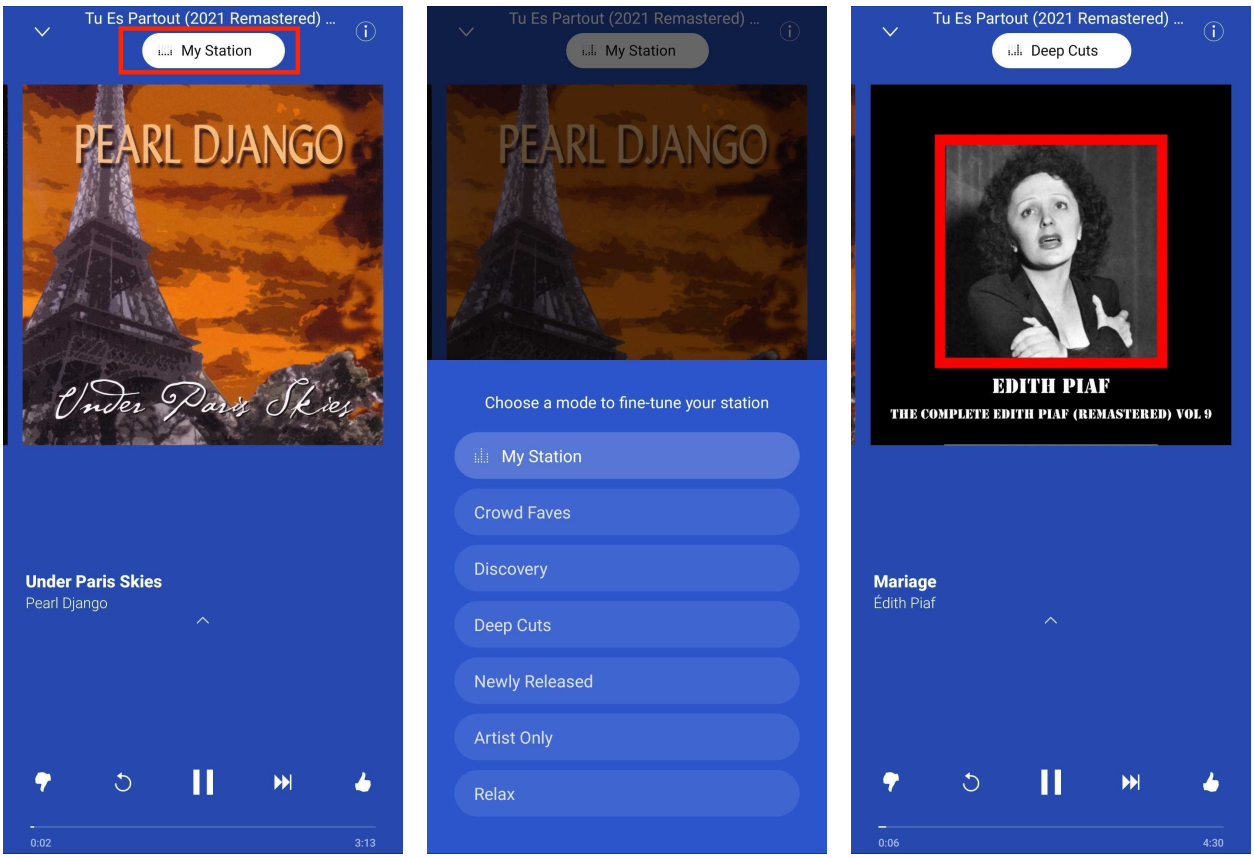
In this section we discuss (1) the setting for our study, (2) our experimental design, (3) the process of making the audio ad creatives, and (4) the implementation of the study.

### *Study Setting*

We study a large, randomized advertising campaign that ran on Pandora Internet Radio in the spring of 2020. Pandora is a music streaming service that operates exclusively in the United States, with nearly sixty million monthly active listeners in 2020. The platform generates revenue through subscriptions and advertising. Our experiment targeted ad-supported Pandora users who were eligible to receive advertisements.

To investigate the effectiveness of ad length and frequency, we designed an experimental marketing campaign promoting the Pandora Modes feature. Modes allows listeners to customize their radio stations by selecting between different listening experiences such as "Crowd Faves" (most thumbed-up songs by other listeners), "Discovery" (more artists who don't usually play on that station), "Deep Cuts" (lesser-known tracks from station artists), and "Newly Released" (newest releases from station artists). To access Modes, listeners must navigate to a station and tap the "My Station" button, which opens a pop-up menu displaying the available mode options (see Figures 1a, 1b, and 1c).

This feature provides an ideal setting for our questions because it is relatively less used and




(a) (b) (c)

Figure 1: Accessing Pandora Modes: User Interface Flow. Panel (a) shows the station interface where users must locate the "My Station" button. Panel (b) displays the pop-up menu that appears with available mode options. Panel (c) shows the interface after a mode has been selected and activated.

less discoverable, creating substantial variation in baseline awareness. This variation allows us to examine whether longer ads are particularly effective for users who start out unfamiliar with the feature, and whether shorter ads function as efficient substitutes for users who already know about it. We also expect that advertising may generate more persistent lift among previously unaware users, as learning about a new feature can produce effects that extend well beyond the ad exposure window. Our experiment design and extended observation period up to two years after the experiment allow us to examine these hypotheses.

### *Experimental Design*

To examine the effect of ad length and frequency on ad effectivness, we designed a randomized field experiment targeting Pandora's ad-supported users. Rather than running a single advertising campaign, we created multiple distinct advertising campaigns and enrolled users in different combinations of these campaigns to vary exposure patterns, which we discuss below.

#### *Ghost ads methodology*

Our experimental design follows the Ghost Ads methodology developed by Johnson et al. (2017), which solves a fundamental problem in measuring advertising effectiveness. The core challenge is that when researchers assign listeners to receive promotional ads, many never actually hear them due to limited listening time, competing advertiser demand, or ad-serving algorithms. Traditional intent-to-treat analysis compares all assigned listeners regardless of actual exposure and compares them against a control group that did not receive ads, but this reduces statistical power because unexposed listeners contribute only noise to the analysis.

The Ghost Ads solution uses placebo advertisements to identify comparable exposed listeners across treatment and control groups. We run placebo campaigns (advertising an unrelated nonprofit) alongside our promotional campaigns, allowing us to identify which control group listeners would have received ads under similar conditions. To elaborate further on the idea of placebo ads: when we book promotional ads to the treatment group, we get to specify which listeners are eligible for the ad campaign, but that does not guarantee that each targeted listener will actually receive an ad in this campaign. For example, listeners who are only active on Pandora for a short time

during the experiment are unlikely to hear any promotional ads, even if they are eligible to hear them. Similarly, listeners in high demand by other advertisers might have had our campaign's ads crowded out by ads from those other advertisers. We prefer to exclude unexposed treatment-group listeners from our analysis, because their behavior could not have been influenced by the ads, and therefore they can contribute only statistical noise, but no signal, to our analysis. But if we want to exclude these listeners, we have to exclude comparable control-group listeners. Placebo ads allow us to identify the comparable control-group listeners and thus narrow our sample to listeners who were exposed to either promotional or placebo ads, for increased statistical power.

Since the classical ghost ads design only creates variation in exposure to the campaign (on/off), the realized frequency of ads remains endogenous and depends on the factors such as user activity or advertiers' demand for that user. While we can filter comparable users in the control group based on exposure, we cannot examine how repetition affects incremental ad effects because the classic ghost ads design does not generate exogenous variation in the number of impressions served. To study the effect of frequency and format, we extend the methodology by generating exogenous variation in both format and realized number of ads. As illustrated in Figure 2, instead of enrolling each listener in a single campaign, we assign each listener to a bundle of three parallel campaigns, and vary the composition of promotional and placebo campaigns across treatment arms. Because each campaign follows the same delivery rules, changing the mix of campaigns changes the expected number of promotional impressions, thereby generating exogenous variation in realized frequency while still preserving the ghost-ads logic for identifying comparable exposed listeners. We describe the mechanics of this multi-campaign design in detail below; the key idea is that varying the composition of campaign bundles introduces the variation needed to study how ad length and frequency interact.

*Campaign types and enrollment structure*

We created multiple parallel advertising campaigns, and each user in the experiment was enrolled in exactly three of these campaigns simultaneously. All campaigns used identical targeting criteria, frequency caps, and delivery parameters, targeting listeners who had been active on the platform at least once in the 60 days before the start of the experiment, excluding only listeners who were

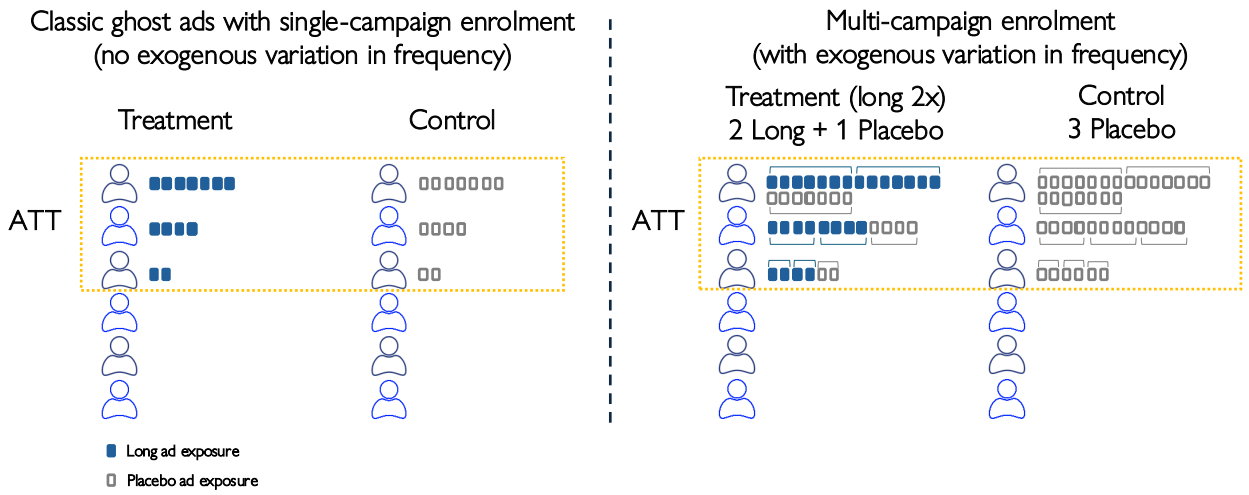



Figure 2: Classical ghost ads versus multi-campaign ghost ads. The left panel shows the classical ghost ads design, which uses placebo ads to identify comparable exposed listeners, those who would have received an impression under either condition, allowing estimation of the Average Treatment Effect on the Treated (ATT). However, this design does not generate exogenous variation in ad frequency: conditional on exposure, the number of impressions remains endogenous and is determined by user activity and advertisers' demand. The right panel shows our extension, in which each listener is enrolled in a bundle of three parallel campaigns (e.g., two long-ad campaigns plus one placebo vs. three placebos). This design generates exogenous variation in realized frequency while preserving the ghost-ads logic for identifying the comparable exposed set.

part of Pandora's existing marketing holdout groups (totaling 3% of the user base).[4]

We designed two distinct types of advertising campaigns. *Promotional campaigns* advertised the Pandora Modes feature, providing content about how to access and use the modes functionality. *Placebo campaigns* advertised Innovations for Poverty Action (IPA), a nonprofit organization completely unrelated to Pandora's features, soliciting donations for global poverty reduction efforts. The key assumption underlying our use of placebo campaigns is that IPA advertisements will not affect listeners' usage of Pandora Modes, allowing us to isolate the effects of promotional content.

Both campaign types were produced in two lengths: *short ads* lasting 10-11 seconds and *long ads* lasting 27-30 seconds. For Modes promotional campaigns, we created three different short ad creatives and three different long ad creatives. For IPA placebo campaigns, we created two short and two long ad creatives. Multiple creatives within each category were rotated randomly by the ad server to ensure our results were not driven by a single unusually effective or ineffective advertisement. All ads included complementary display banners that appeared on mobile screens

[4]All campaigns were configured with identical frequency caps: maximum 1 ad per hour, 5 ads per day, and 15 ads per week.

during audio playback, please see Web Appendix A for more details.

*Treatment arm structure*

The experimental manipulation works by varying which combination of the campaigns, e.g., short/long and promotional or placebo, each listener is enrolled in. For example, a listener assigned to the "Long" treatment is enrolled in one long promotional campaign and two placebo campaigns. A listener in the "Short 2x" treatment is enrolled in two short promotional campaigns and one placebo campaign. A listener in the control group is enrolled in three placebo campaigns.

The experimental structure is summarized in Table 1. We randomized ten percent of listeners into the control group and fifteen percent into each of six treatment arms. For exposition, we denote x as the target number of promotional ads in the simplest treatment arm, recognizing that realized delivery varies across listeners due to differences across users. However, our design ensures that the same type of user is exposed to the same number of ads (3x in total) across all experiment conditions.

Table 1: Experimental design. The table shows the intended number of ads per type (short, long, placebo) listeners would receive in each of the 7 experimental arms.

| Arm | Placebo ads | Long ads | Short ads |
|---|---|---|---|
| Control | 3x | 0 | 0 |
| Long | 2x | 1x | 0 |
| Long 2x | 1x | 2x | 0 |
| Short | 2x | 0 | 1x |
| Short 2x | 1x | 0 | 2x |
| Short 3x | 0 | 0 | 3x |
| Combine | 1x | 1x | 1x |

As presented in Table 1, the Control arm receives only placebo ads (3x per person). The Long arm receives one long promotional campaign plus two placebo campaigns (1x + 2x = 3x total). The Long 2x arm receives two long promotional campaigns plus one placebo campaign (2x + 1x = 3x total). We have three short ad arms: Short (1x promotional + 2x placebo), Short 2x (2x promotional + 1x placebo), and Short 3x (3x promotional + 0x placebo). Lastly, users assigned to the Combine arm are enrolled in one long campaign, one short campaign, and one placebo campaign. In the next section, we discuss the implementation details and present the realized experimental variation in

the data.

*Creating audio ad creatives*

We worked with the domain experts in Pandora's marketing department to produce ad creatives for this experiment. Our objective was to make the most effective audio ad possible within each time limit (30 seconds for long ads, 10 seconds for short ads). The final audio creatives ended up ranging from 10 to 11 seconds for short audio ads, and 27 to 30 seconds for long audio ads. Of course, changing the length of an ad necessarily changes its content. In practice, however, advertisers face this exact trade-off, and our creative process reflected industry norms by asking Pandora's staff to generate the best possible copy within the assigned time constraint. We return to this issue in the results discussion below.

For the Modes promotional campaigns, we created three different long ads and three different short ads. The different creatives were "rotated" randomly by the ad server within the campaign, so that each listener got a random mixture of ad creatives. The advantage of creating multiple ads of each type was to reduce the risk of our results being driven by a single unusually effective (or ineffective) ad. We also created display banner ads which appeared on the mobile screen while the audio ads played (if the listener happened to be looking at the Pandora app while the ad played). We used the same display banners for both short and long ads.

For placebo ads, we cooperated with Innovations for Poverty Action (IPA), a charity devoted to randomized experimental evaluation of development economics policies. Pandora's design team worked in conjunction with IPA to design two short and two long ad creatives to advertise their charity and solicit donations. We note that a key assumption for the use of the placebo ads is that the placebo ads will not impact listeners' engagement with the Pandora features that are promoted. This assumption seems reasonable as the ads for Innovations for Poverty Action are completely unrelated to the promoted Pandora features. Transcripts of the audio ads and pictures of the display banners can be found in Web Appendix A.

*Implementation*

We implemented the experiment in spring 2020, capitalizing on reduced advertiser demand during the COVID-19 pandemic to book extensive experimental advertising inventory. We randomized the majority of Pandora's ad-supported listeners into experimental treatment arms, targeting listeners who had been active on the platform at least once in the 60 days before the experiment launch. We only excluded listeners who were part of Pandora's existing marketing holdout groups (totaling 3% of the user base). We randomized 10% of eligible listeners into the control group and 15% into each of the six treatment arms, making the treatment arms slightly larger than the control as we anticipated needing more statistical power to detect differences between treatment arms than between each treatment arm and control.

The experiment ran from May 31 to June 16, 2020, during which we delivered approximately 220 million experimental ads promoting the Pandora Modes feature across all treatment arms. The campaign reached 18.3 million listeners total, with each of the six treatment arms reaching approximately 2.75 million listeners and the control arm reaching 1.84 million listeners.

Table 2 presents the average realized number of ads delivered across experimental arms. As expected, the average delivery closely matched our intended experimental design from Table 1. Listeners in the control group received an average of 11.95 placebo ads over the experimental period. In the promotional arms, realized delivery aligned with our targeted ratios: listeners in the Long arm received an average of 4.02 long ads and 7.98 placebo ads (approximately a 1:2 ratio as intended), while those in the Long 2x arm received 7.98 long ads and 4.00 placebo ads (approximately a 2:1 ratio). Similarly, the Short arms showed proportional delivery, with the Short 3x arm receiving an average of 12.04 short ads and no placebo ads, exactly as designed.[5]

While the average delivery matched our experimental design, there was substantial variation in the number of ads received by individual listeners within each treatment arm. Figure 3 illustrates this variation by showing the distribution of total promotional ads received by listeners in the Long 2x arm. The average listener in this arm received approximately 8 long ads, but the distribution shows considerable heterogeneity: some listeners received as few as 1-2 ads, whereas others received

[5]Web Appendix E presents the full distribution of total advertising exposure (promotional plus placebo) across arms.

Table 2: Number of users reached and average realized number of ads across experimental arms. The table reports listener counts and the average number of placebo, short, and long ads delivered. Ad intensity varied by arm (1x, 2x, 3x), and the "Combine" arm split exposure evenly across ad types, while the control group received only placebo ads.

| Arm | Listeners | Short ads | Long ads | Placebo ads |
|---|---|---|---|---|
| Control | 1837838 | 0.00 | 0.00 | 11.95 |
| Short | 2742029 | 3.98 | 0.00 | 8.03 |
| Short 2x | 2745781 | 7.99 | 0.00 | 4.03 |
| Short 3x | 2746495 | 12.04 | 0.00 | 0.00 |
| Long | 2751667 | 0.00 | 4.02 | 7.98 |
| Long 2x | 2747568 | 0.00 | 7.98 | 4.00 |
| Combine | 2748471 | 4.02 | 4.00 | 4.00 |

20 or more. This variation stems from multiple factors including differences in listening patterns and competition from other advertisers for particular audience segments.

This variation in realized impressions has important implications for campaign effectiveness. The substantial variation in realized impressions creates heterogeneity in both the impact and efficiency of advertising across listeners. Optimal advertising strategies should account for both heterogeneity in the returns to advertising (i.e., which listeners might respond more to the campaign conditional on their current exposure level) and heterogeneity in exposure patterns across listeners. Traditionally, advertisers have used frequency caps to limit overexposure to heavy listeners, but this approach does not fully leverage the variation in both effectiveness and exposure patterns across the entire listener distribution. We return to this issue in the "Personalization" Section, where we explore how content platforms might personalize ad length and frequency strategies to better account for these sources of heterogeneity.

Lastly, to verify successful randomization, Table 3 presents balance tests across the seven experimental arms. Panel A reports pre-experimental Modes usage metrics for each arm. To preserve commercial confidentiality per our agreement with Pandora, all usage metrics in Panel A have been multiplied by an undisclosed constant, though this transformation preserves relative differences across arms and statistical relationships. Across all arms, baseline Modes usage was relatively modest. Although we cannot disclose the exact usage rates, we can report that less than 10% of listeners had used the feature in the two months before the experiment. The similarity of usage patterns across arms provides initial evidence of successful randomization. Panel B presents a formal bal-

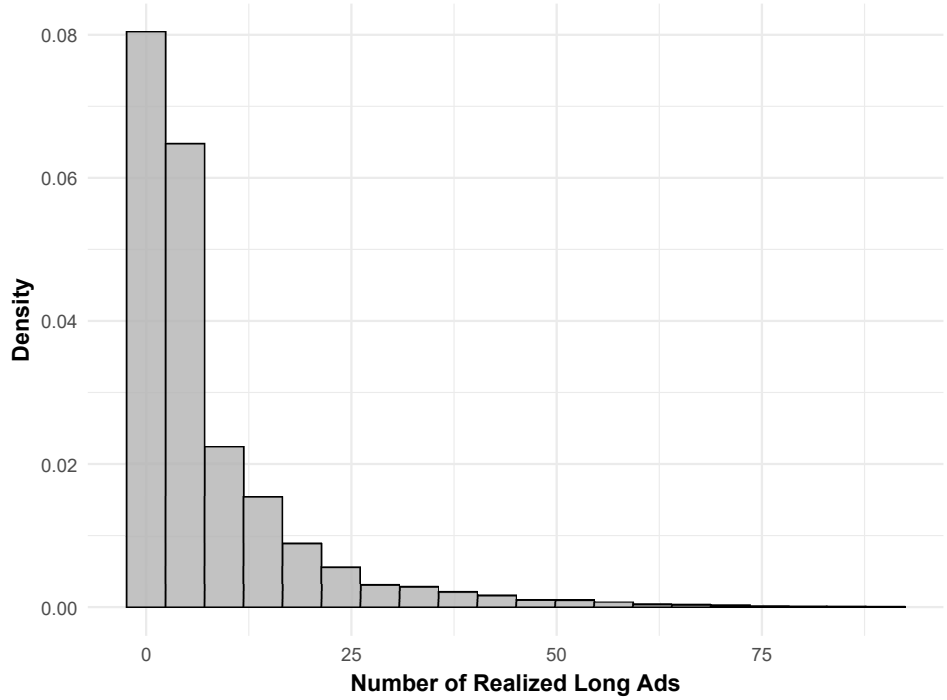



Figure 3: Distribution of Realized Long Ad Impressions in the Long 2x Treatment Arm.

ance test across arms. For both continuous measures of usage duration, F-tests yield p-values above 0.05 (p=0.063 for 2-month usage, p=0.072 for 3-week usage), indicating no significant differences across arms. Similarly, chi-squared tests for differences in the share of users who tried Modes yield p-values of 0.169 and 0.378 for the two-month and three-week periods, respectively, confirming balanced randomization.

# Results

In this section, we present results from our field experiment. We first show that the advertising treatments have a causal impact on usage of the promoted feature, with effects that persist for up to two years. Second, using an instrumental-variable approach, we show that long ads are substantially more effective than short ads on a per-ad and per-second basis. Third, we demonstrate that this differential effectiveness varies across users: long ads are particularly more effective (relative to short ones) for users without prior feature experience, while the advantage is much smaller among

Table 3: Pre-treatment Modes Usage and Balance Tests. Panel A reports means with standard errors in parentheses. The usage metrics are multiplied by an undisclosed constant to preserve commercial confidentiality, though this preserves relative patterns and statistical relationships. Panel B reports F-tests for continuous variables and chi-squared tests for binary variables, testing the null hypothesis of no differences across the seven experimental arms.

| **Panel A: Pre-treatment Modes Usage by Experimental Arm** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Control | Short | Short 2x | Short 3x | Long | Long 2x | Combine |
| Pretreatment Modes usage | 3431.8 | 3525.2 | 3457.5 | 3509.2 | 3412.5 | 3443.9 | 3499.0 |
| (2 months, sec × constant) | (36.4) | (30.0) | (29.4) | (29.8) | (29.3) | (29.0) | (30.9) |
| Pretreatment Modes usage | 1471.1 | 1498.8 | 1482.2 | 1508.3 | 1448.9 | 1477.7 | 1490.5 |
| (3 weeks, sec × constant) | (16.8) | (13.9) | (13.8) | (14.1) | (13.5) | (13.7) | (14.3) |
| Used Modes once, pre-treatment | 0.149 | 0.150 | 0.150 | 0.150 | 0.149 | 0.150 | 0.150 |
| (2 months, share of users × constant) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Used Modes once, pre-treatment | 0.082 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 | 0.083 |
| (3 weeks, share of users × constant) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| N | 1,837,838 | 2,742,029 | 2,745,781 | 2,746,495 | 2,751,667 | 2,747,568 | 2,748,471 |

| **Panel B: Balance Tests Across Arms** | | | |
|---|---|---|---|
| Variable | Test | Statistic | P-value |
| Pretreatment Modes usage (2 months, sec) | F | 1.99 | 0.063 |
| Pretreatment Modes usage (3 weeks, sec) | F | 1.931 | 0.072 |
| Used Modes once, pre-treatment (2 months) | $\chi^2$ | 9.075 | 0.169 |
| Used Modes once, pre-treatment (3 weeks) | $\chi^2$ | 6.421 | 0.378 |

users already familiar with the feature. Finally, we examine aggregate treatment effects over the first year and show that mixing short and long ads in a uniform, non-targeted way generates little incremental benefit, thereby motivating the personalized approach developed in the next section.

### *Overall Treatment Effects on Modes Usage*

We begin by examining the overall impact of our advertising treatments on Modes usage across experimental arms.[6] To preserve confidentiality while measuring treatment effects, we normalize outcomes relative to the control group average, expressing results as percentage deviations from control rather than absolute usage levels:

$$\tilde{Y}_{it} = 100 \cdot \frac{Y_{it}}{\frac{\sum_{j \in \mathcal{C}} Y_{jt}}{\mathcal{N}_{\mathcal{C}}}}, \tag{1}$$

where $Y_{it}$ is a binary indicator equal to 1 if listener $i$ used Modes in period $t$, $\mathcal{C}$ denotes the set of listeners in the control group, and $\mathcal{N}_{\mathcal{C}}$ is the number of control group listeners. We then estimate the following reduced-form regression:

[6]We note that Appendix D confirms treatment assignment did not differentially affect listening behavior itself, validating that our treatment effects capture advertising impact on feature adoption rather than changes in activity.

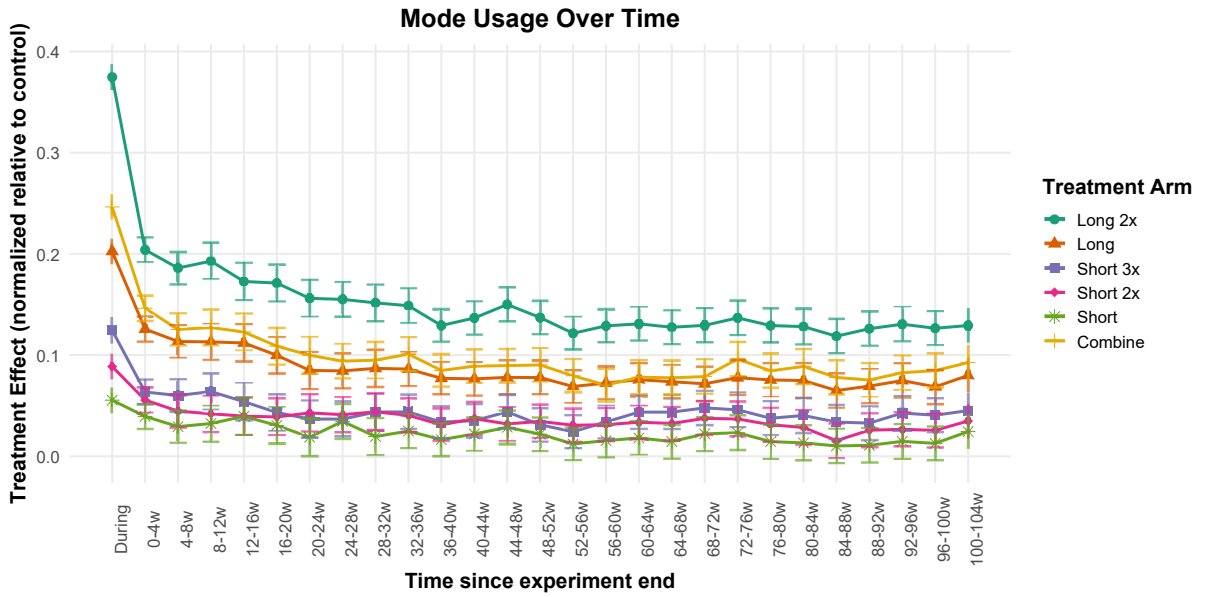



Figure 4: Treatment Effects on Modes Usage Over Time.

$$\tilde{Y}_{it} = \alpha_t + \sum_k \beta_{kt} \cdot \mathbb{I}_{\{\text{Treatment}_i = k\}} + \epsilon_{it}, \tag{2}$$

where $\mathbb{I}_{\{\text{Treatment}_i=k\}}$ are indicator variables for each of the six treatment arms (Short, Short 2x, Short 3x, Long, Long 2x, Combine), with the control group serving as the omitted baseline. We estimate this regression separately for each time period $t$: during the experimental period and in four-week blocks for up to 104 weeks (two years) post-experiment. Figure 4 displays the estimated coefficients $\beta_{kt}$ over time, showing how each treatment arm's effect on Modes usage evolves relative to control.

We first begin by examining the lift during the experimental period. All promotional treatment arms show immediate and substantial increases in Modes usage, with effects ranging from 5.5% (Short arm, $p = 6.01 \times 10^{-18}$) to 37.5% (Long 2x arm, $p \approx 0$)[7] relative to control. The ranking by effectiveness during the experimental period is: Long 2x (37.5%, $p \approx 0$), Combine (24.6%, $p = 3.36 \times 10^{-322}$), Long (20.3%, $p = 1.79 \times 10^{-218}$), Short 3x (12.4%, $p = 1.06 \times 10^{-83}$), Short 2x (8.9%, $p = 2.46 \times 10^{-43}$), and Short (5.5%, $p = 6.01 \times 10^{-18}$). All treatment effects are highly statistically significant.

Treatment effects persist well beyond the experimental period, remaining elevated for the full

[7]This p-value is machine zero.

two-year observation window, though they exhibit gradual decay over time. We observe clear and persistent differences in effectiveness across treatment arms. The Long 2x arm consistently shows the largest treatment effects throughout the observation period, while the Short arm shows the smallest effects among promotional treatments. This ranking remains remarkably stable over time. By week 104, treatment effects range from 2.4% (Short, $p = 0.004$) to 12.9% (Long 2x, $p = 2.74 \times 10^{-51}$), demonstrating substantial persistence of the advertising effects. Even after two years, all treatment arms maintain statistical significance: Long 2x ($p = 2.74 \times 10^{-51}$), Combine ($p = 3.66 \times 10^{-27}$), Long ($p = 1.22 \times 10^{-20}$), Short 3x ($p = 1.43 \times 10^{-7}$), Short 2x ($p = 4.83 \times 10^{-5}$), and Short ($p = 0.004$).

The persistence of treatment effects indicates that the advertising campaign served an informative rather than a simple reminder function for at least some listeners. For these users, exposure to the promotional content provided new information about a feature they would not have otherwise discovered or used. This interpretation is consistent with the low baseline Modes usage (less than 10% of listeners had used the feature in the two months before the experiment), suggesting limited organic discovery of this functionality. Once informed about the feature's existence and access method through the advertisements, these listeners incorporated Modes into their ongoing listening behavior without continued promotional exposure. This persistence contrasts with reminder advertising effects, which typically exhibit rapid decay following the cessation of exposure. In the next step, we employ an instrumental variable approach to further investigate the relative effectiveness of long versus short ads and the underlying mechanisms through which they operate.

### *Instrumental Variable Analysis: Per-Ad Effectiveness of Short versus Long Ads*

We now pool across experimental conditions to formally test the relative effectiveness of short versus long ads on a per-ad basis. We employ an instrumental variable approach that leverages the randomized experimental variation in ad delivery across treatment arms. Similar to the previous section, we normalize the outcome variable as in equation (1), so the coefficients measure the percentage increase in Modes usage per short or long ad relative to control. We estimate the following instrumental variable model:

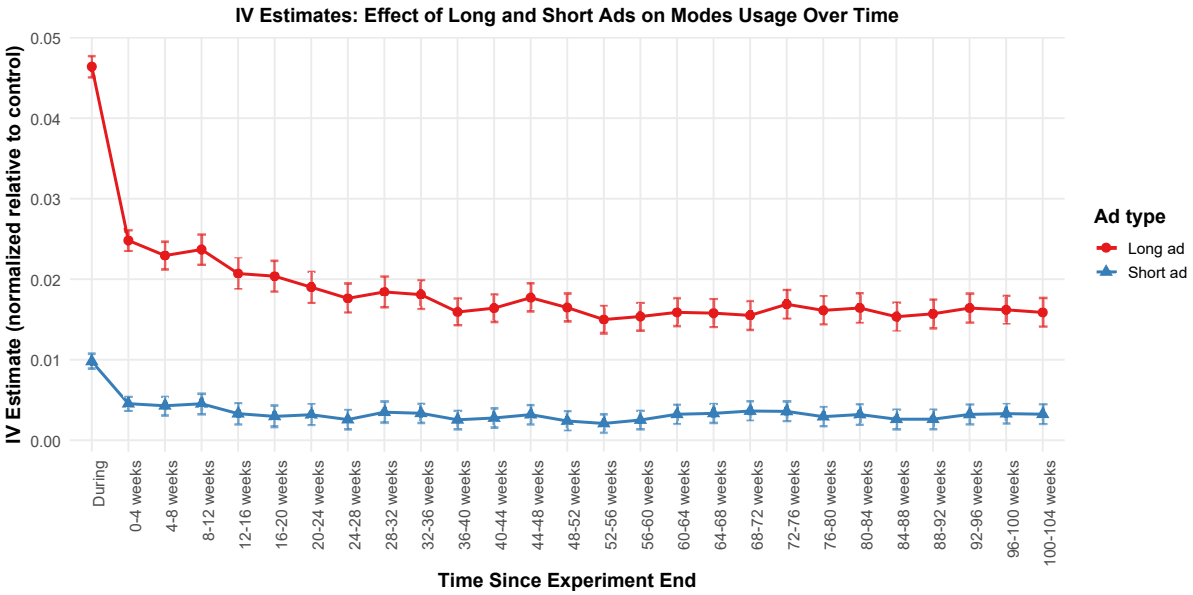



Figure 5: Per-Ad Effectiveness of Long versus Short Ads Over Time. The figure displays estimated coefficients from the instrumental variable regression in equation (3), showing the percentage point increase in Modes usage per ad relative to control. Error bars represent 95% confidence intervals.

$$\tilde{Y}_{it} = \beta_{0t} + \beta_{St}\widehat{S_i} + \beta_{Lt}\widehat{L_i} + \varepsilon_{it}, \tag{3}$$

where $S_i$ represents the number of short promotional ads received by listener $i$, $L_i$ represents the number of long promotional ads received, and $\widehat{S_i}$ and $\widehat{L_i}$ denote the fitted values from the first-stage regressions. The coefficients $\beta_{St}$ and $\beta_{Lt}$ represent the per-ad effectiveness of short and long ads, respectively, in period $t$. The first-stage regressions are:

$$S_i = \gamma_{S0} + \sum_k \gamma_{Sk} \cdot \mathbb{I}_{\{\text{Treatment}_i = k\}} + \eta_{Si}, \tag{4}$$

$$L_i = \gamma_{L0} + \sum_k \gamma_{Lk} \cdot \mathbb{I}_{\{\text{Treatment}_i = k\}} + \eta_{Li}, \tag{5}$$

where $\mathbb{I}_{\{\text{Treatment}_i = k\}}$ are indicator variables for the six treatment arms, with control serving as the omitted category. Our identification relies on the randomized assignment of listeners to treatment conditions.

Figure 5 displays the estimated coefficients $\beta_{St}$ and $\beta_{Lt}$ from the instrumental variable regression in equation (3) over time. Several key patterns emerge from this analysis. First, long ads consistently outperform short ads on a per-ad basis throughout the entire observation period. During the

experimental period (week 0), each long ad generates a 4.6 percentage point increase in Modes usage relative to control ($p \approx 0$), compared to 1.0 percentage point for each short ad ($p = 1.93 \times 10^{-100}$). This represents a 4.7:1 effectiveness ratio, substantially higher than the 3:1 length ratio of the ads.

Second, the per-ad effectiveness of both ad types declines over time but remains statistically significant throughout the two-year observation window. By week 104, long ads generate a 1.6 percentage point increase per ad ($p = 1.66 \times 10^{-69}$) while short ads generate a 0.3 percentage point increase ($p = 1.51 \times 10^{-7}$), maintaining an approximately 5:1 effectiveness ratio. The persistence of these effects indicates that, for at least some users, the ads provided information they acted on over time rather than functioning solely as a temporary reminder or a persuasive nudge.

Third, the ratio of long to short ad effectiveness remains relatively stable over time, ranging between 4.5:1 and 6:1 across most periods. This consistency, together with the low share of users who had used Modes prior to the experiment, suggests that long ads may serve better as informative advertising while short ads function more as reminders. The superior performance of long ads likely reflects their greater capacity to convey information about feature functionality and access methods. To further examine this possibility, we next repeat the analysis separately for listeners who had previously used Modes and those who had not used it in the two months before the experiment. This distinction allows us to examine how the relative effectiveness of long and short ads varies with prior awareness, providing a clearer picture of how the two formats perform across user segments.

### *The Role of Prior Experience in Ad Effectiveness*

We first introduce a simple illustrative model to interpret how treatment effects might vary with listeners' prior experience. The model is not intended for estimation but serves two purposes: it motivates why comparing the relative effectiveness of long and short ads across previously informed and uninformed users is informative, and it highlights the type of heterogeneity that our machine-learning algorithm ultimately uses when personalizing ad length. It also provides a conceptual lens for evaluating how the personalized policy allocates formats across users.

Let $i$ index listeners and let $A_i \in \{0, 1\}$ indicate whether listener $i$ is already aware of the Modes feature ($A_i = 1$) or not ($A_i = 0$). Each listener is exposed to an ad format $f \in \{0, S, L\}$ (no promotional ad, short ad, long ad). If $A_i = 0$, format $f$ introduces Modes into the listener's

choice set with probability $a_f$, where $a_0 = 0$ and $a_S \leq a_L \leq 1$; if $A_i = 1$, Modes is always available. Conditional on availability, the utility from using Modes after format $f$ is given by:

$$U_{if} = v_i + \rho_f + \varepsilon_{if},$$

where the utility of not using Modes is normalized to zero and the shocks $\varepsilon_{if}$ are i.i.d. type-I extreme value. Thus the probability of use conditional on availability follows the standard logit form:

$$p_f(v_i) = \frac{\exp(v_i + \rho_f)}{1 + \exp(v_i + \rho_f)}, \qquad \rho_0 = 0.$$

Overall usage probabilities are therefore equal to:

$$P_f^{\text{inf}} = p_f(v_i), \qquad P_f^{\text{uninf}} = a_f \, p_f(v_i).$$

To summarize the relative effectiveness of long versus short ads, we consider the ratio:

$$R^g = \frac{P_L^g}{P_S^g}, \qquad g \in \{\text{inf}, \text{uninf}\}.$$

For informed listeners,

$$R^{\text{inf}} = \frac{p_L(v_i)}{p_S(v_i)} = \frac{\exp(\rho_L)}{\exp(\rho_S)},$$

because the common match component $v_i$ cancels out under the logit structure. For uninformed listeners,

$$R^{\text{uninf}} = \frac{a_L p_L(v_i)}{a_S p_S(v_i)} = R^{\text{inf}} \cdot \frac{a_L}{a_S}.$$

Hence the excess long-short ratio among previously uninformed relative to informed listeners,

$$\frac{R^{\text{uninf}}}{R^{\text{inf}}} = \frac{a_L}{a_S}.$$

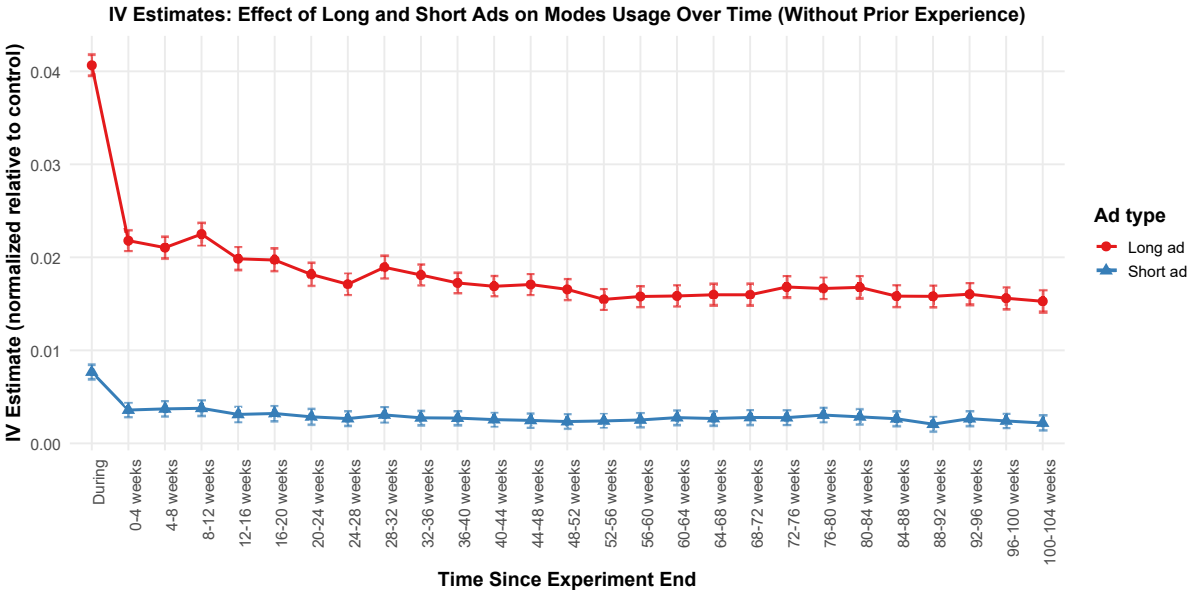
This decomposition identifies the informative advantage of long ads through the ratio $a_L/a_S$, while $R^{\text{inf}}$ captures a persuasive (reminder) effect. Our goal in presenting this structure is not to test a theory of informative versus persuasive (reminder) mechanisms or to take a position

on their relative importance. The model is intended only to highlight why prior experience is a natural dimension along which the performance of long and short ads may differ. We refer to the persuasive component in this model as a reminder effect and remain agnostic about its deeper interpretation. The key point is that this simple structure helps organize the heterogeneity in our data and highlights why such heterogeneity is useful for personalization.

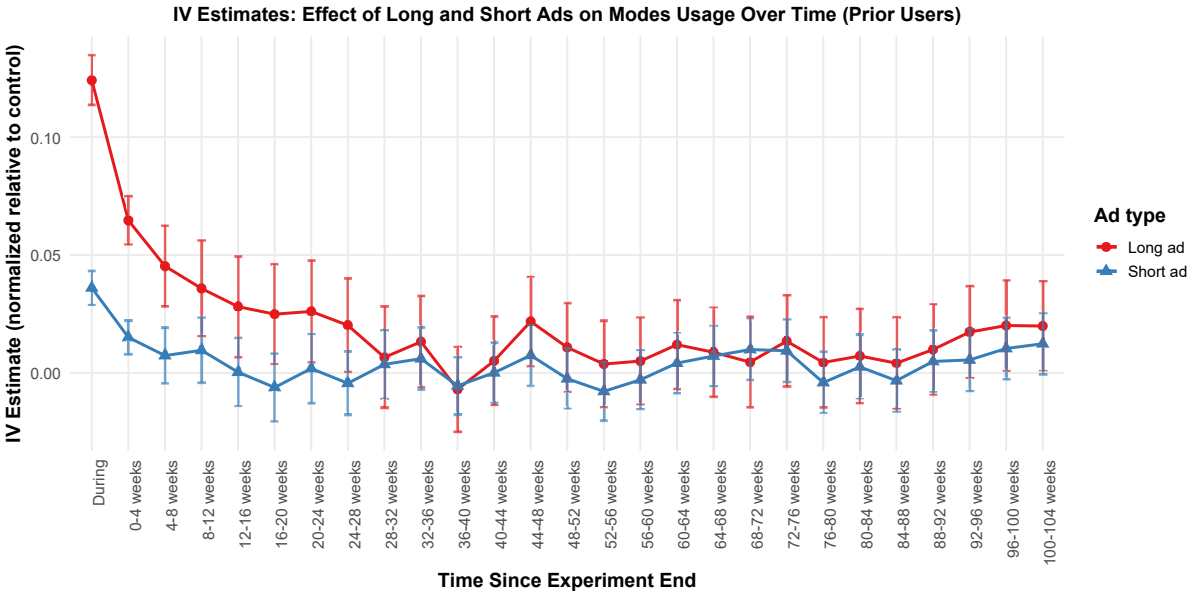
Motivated by this framework, we now examine how advertising effectiveness of long versus short ads varies across listeners with and without prior experience using the Modes feature. We identify prior users as listeners who used Modes at least once in the two months before the experiment. We then estimate the instrumental variable regression from equation (3) separately for users without prior experience and users with prior experience. Figure 6 displays the results of this analysis, with panel (a) showing per-ad effectiveness for users without prior experience and panel (b) showing results for users with prior experience. The contrast between these panels highlights clear differences in advertising effectiveness across the two groups.

Among users without prior experience (Panel a), long ads demonstrate substantial and persistent superiority over short ads. During the experimental period (week 0), each long ad generates a 4.1 percentage point increase in Modes usage relative to control ($p \approx 0$), compared to 0.8 percentage point for each short ad ($p \approx 0$). This represents a 5.3:1 effectiveness ratio, substantially exceeding the 3:1 duration ratio. The effects remain significant throughout the two-year observation window, with the relative advantage of long ads actually increasing over time. By week 104, long ads still generate a 1.5 percentage point increase per ad ($p = 2.77 \times 10^{-144}$), while short ads generate a 0.2 percentage point increase ($p = 8.33 \times 10^{-8}$), yielding a 7.0:1 effectiveness ratio.

Among users with prior experience (Panel b), the pattern is very different. During the experimental period, long ads generate a 12.4 percentage point increase in Modes usage, while short ads generate a 3.6 percentage point increase. This represents a 3.4:1 effectiveness ratio, much closer to the 3:1 duration ratio. Furthermore, the effectiveness of both ad types decays rapidly, and the effects shrink toward zero. By week 8, short ads produce an effect of only 0.7 percentage points ($p = 0.219$), while long ads show 4.5 percentage points ($p = 2.66 \times 10^{-7}$). By week 104, long ads generate 2.0 percentage points ($p = 0.040$) while short ads generate 1.2 percentage points ($p = 0.060$), with the short ad effect no longer statistically significant at conventional levels.




(a) Users Without Prior Modes Experience




(b) Users With Prior Modes Experience

Figure 6: Per-Ad Effectiveness by Prior Experience: Long versus Short Ads Over Time. Panel (a) displays estimated coefficients from the instrumental variable regression for listeners without prior Modes experience (95.2% of sample). Panel (b) shows results for listeners with prior Modes experience, defined as having used Modes at least once in the two months before the experiment. Both panels show the percentage point increase in Modes usage per ad relative to control, with error bars representing 95% confidence intervals.

The key empirical finding is the difference in the relative effectiveness of long ads to short ads across user types. Following our illustrative model, the ratio of ratios $R^{\text{uninf}}/R^{\text{inf}} = 5.3/3.4 \approx 1.6$ during the experimental period provides a measure of the relative informative advantage of long ads ($a_L/a_S$). These patterns reveal substantial heterogeneity in how different users respond to advertising formats, with suggestive evidence that prior experience shapes ad effectiveness. Among users without prior experience with Modes, the relative effect of long ads to short ones was larger and effects persisted over time.

These contrasting temporal patterns have important implications for targeting strategies as well. If we evaluated effectiveness only during the experimental period, experienced users would appear particularly valuable with initial responses of 12.4 percentage points for long ads and 3.6 for short ads. However, our extended observation window provides a more complete picture. While retargeting experienced users yeilds a larger initial response, these effects decline over time. In contrast, targeting uninformed users yields a smaller responses but appear to be more persistent.

These differential patterns in relative effectiveness help interpret what our design captures. Changes in ad length necessarily imply changes in ad content, i.e., it is impossible to extend or shorten an advertisement without adjusting the script, tone, and pacing. In practice, advertisers aim to produce the best possible creative given the allotted time. Our process mirrored this industry norm: Pandora's marketing staff were instructed to generate the highest-quality copy feasible within the assigned length constraints (see Web Appendix A for more details). Thus, while length and content are intertwined, this confounding reflects the actual decision environment facing advertisers. What our experiment identifies is therefore not the effect of "time alone," but the effectiveness of realistic campaigns designed for different durations.

Nevertheless, we believe the comparisons across arms, over time, and across user experience levels are informative. Among listeners who were unfamiliar with the feature, advertising in general has persistent effects, but the relative advantage of long ads is especially pronounced in this group. Among prior users, by contrast, effects for both formats dissipate rapidly and the gap between long and short ads is much smaller. Taken together, these patterns support the interpretation that long ads act primarily as informative advertising, providing new knowledge to unfamiliar users, while short ads operate more as persuasive reminders once basic awareness has been established.

This distinction between information and reminder roles provides a coherent lens through which to interpret the heterogeneous treatment effects.

### *Aggregate Treatment Effects: First-Year Analysis*

We now examine treatment effects aggregated across the experimental period and the subsequent twelve months to address two methodological concerns. First, this aggregation establishes that our findings are robust and do not arise from heterogeneous response patterns across listener types or from our specific temporal aggregation choices. The stability of treatment arm rankings across different time windows demonstrates that the relative effectiveness of advertising strategies reflects systematic differences in campaign performance rather than measurement artifacts. Second, by focusing on cumulative effects over a full year, we provide a policy-relevant measure that captures the total economic value of different campaign designs for advertisers making budget allocation decisions.

Figure 7 reports results from the reduced-form regressions introduced above (equation (2)), now applied to outcomes aggregated over the experiment period and the subsequent first year. All promotional arms yield statistically significant lifts in Modes usage, with $p$-values ranging from $p = 4.66 \times 10^{-9}$ (Short) to $p \approx 3.9 \times 10^{-222}$ (Long 2x). The ranking of arms is consistent with our results in the previous section: Long 2x produces the largest effect (13.9%), followed by Combine (9.1%), Long (8.0%), Short 3x (4.6%), Short 2x (3.6%), and Short (2.6%).

Two comparisons help evaluate the economic mechanisms underlying our findings. First, we compare strategies with similar total advertising duration but different format compositions. The Long treatment (9.2%) substantially outperforms Short 3x (5.4%) despite both involving approximately the same total advertising time per listener. This reinforces our per-ad analysis: the superior effectiveness of long ads cannot be explained solely by their greater duration, consistent with long ads serving primarily an informational function while short ads provide reminders. Second, we examine whether combining formats yields complementary effects. Given that long ads appear more effective at informing users about feature functionality while short ads serve as reminders, one might expect that mixing formats would allow campaigns to both educate uninformed listeners and reinforce usage among those already aware of the feature. The Combine treatment delivers 1x

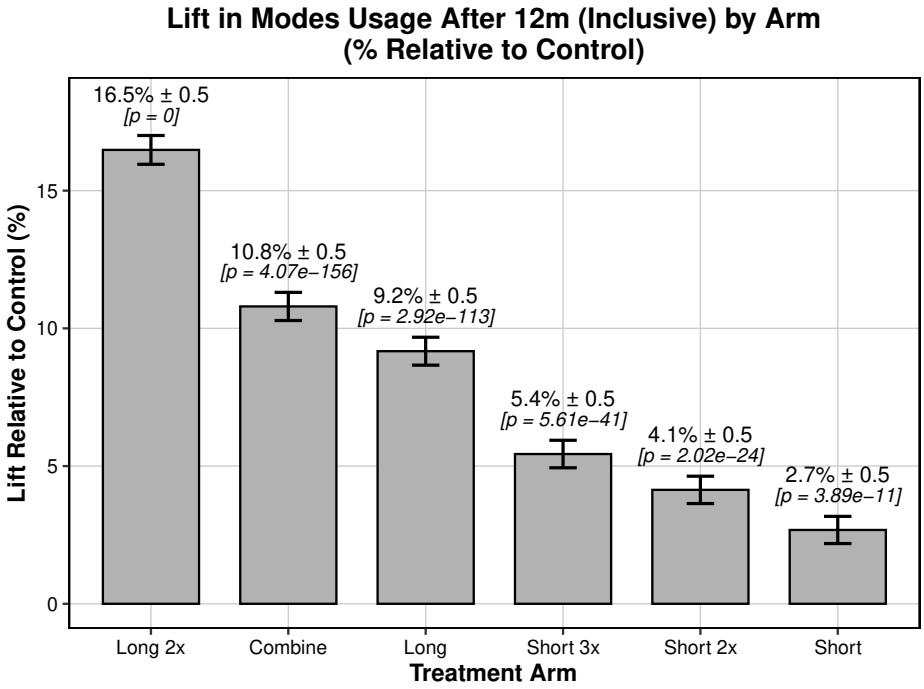



Figure 7: Aggregate Treatment Effects on Modes Usage, One Year Post-Experiment. The figure reports normalized treatment effects relative to control, aggregated over the experimental period and twelve months after.

long and 1x short ads per listener, which in terms of lenght of advertising corresponds to the average exposure of the Long 2x and Short 2x treatments. If formats were simply additive, we would expect Combine to achieve the average performance of these two treatments ($\frac{16.5+4.1}{2} = 10.3\%$). However, Combine achieves 10.8%, only modestly exceeding this benchmark. This suggests that naïve approaches to mixing ad formats have limited success in capturing complementarities. The finding points to the potential value of more sophisticated targeting approaches that account for user heterogeneity rather than uniformly mixing ad formats across all listeners.

We further document heterogeneity in these aggregate effects by examining how treatment responses vary across listener characteristics. Appendix B presents logit regression results comparing select treatment arms, confirming that long ads are particularly effective for listeners with no prior experience with the feature, while the advantage diminishes substantially among familiar users. These heterogeneous treatment effects underscore the limits of uniform campaign designs and motivate the personalized targeting approach we explore next.

# Personalization

Building on the evidence that uniform advertising strategies have limited ability to capture complementarities across ad formats, we now study the value of personalizing ad length and frequency. Using pre-treatment listener characteristics, we construct assignment rules that map each user to a treatment arm in order to maximize expected performance. We then evaluate the performance of these personalized rules using inverse probability weighting (IPW), an off-policy evaluation method that leverages the randomization in our experiment.

To construct optimal personalized policies, we must predict, for each listener and each treatment arm, two key components: expected causal lift in Modes usage ("returns") and expected total ad length delivered, measured as the number of realized impressions multiplied by ad duration ("cost"). With these predictions in hand, a personalized assignment rule can then allocate each listener to the treatment arm that maximizes predicted value per unit of advertising budget.

To understand the sources of improvement from personalization, we compare three counterfactual rules: (i) full personalization, which uses heterogeneous treatment effects (HTEs) for both returns and costs; (ii) returns-only personalization, which ignores heterogeneity in cost by replacing predicted cost with its average (ATE for cost); and (iii) cost-only personalization, which ignores heterogeneity in returns by replacing predicted lift with its average (ATE for returns). Comparing these rules decomposes the gains into contributions from learning heterogeneity in returns and heterogeneity in cost.

### *HTE for Returns: Modes Usage*

We estimate heterogeneous treatment effects for Modes usage using a multi-headed neural network architecture. Let $\mathbf{x}_i$ denote pre-treatment listener characteristics including demographics, historical listening patterns, pre-experiment ad load levels, and prior Modes usage. We model the probability of listener $i$ using Modes under treatment arm $k \in \{0, 1, ..., 6\}$ as:

$$P(\text{Modes}_i = 1|\mathbf{x}_i, k) = \sigma(\theta_k(\mathbf{x}_i)), \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function and $\theta_k(\mathbf{x}_i)$ is the output of treatment-specific head $k$.

The architecture follows the TARNet design (Johansson et al., 2016; Shalit et al., 2017), with shared representation layers $\phi(\mathbf{x}_i)$ followed by treatment-specific output heads. This structure is similar to dueling networks in reinforcement learning (Wang et al., 2016). The shared layers learn features relevant across all treatment conditions, while separate heads capture treatment-specific responses.[8]

Our experimental design randomized 10% of listeners to control and 15% to each of the six treatment arms. This imbalanced design creates different sample sizes across conditions: $N_0 \approx 1.84$ million for control and $N_k \approx 2.75$ million for each treatment arm $k \in \{1, ..., 6\}$. To prevent the optimization from over-fitting to larger treatment groups, we minimize an inverse propensity weighted cross-entropy loss:

$$\mathcal{L}_{\text{returns}} = -\frac{1}{N} \sum_{i=1}^{N} w_{k_i} \left[ y_i \log \hat{p}_{k_i} + (1 - y_i) \log(1 - \hat{p}_{k_i}) \right], \tag{7}$$

where $y_i$ indicates whether listener $i$ used Modes during the 12-month post-experiment period (including the experimental period), $\hat{p}_{k_i} = P(\text{Modes}_i = 1 | \mathbf{x}_i, k_i)$ is the predicted probability under their assigned treatment $k_i$, and $w_{k_i} = \frac{N}{N_k}$ is the inverse propensity weight that balances the effective sample size across treatment conditions.

### *HTE for Costs: Total Ad Duration*

As established earlier, assigning listeners to a treatment arm does not guarantee uniform ad delivery across all users. The number of realized impressions for each indivdiual varies substantially based on listening patterns, competing advertiser demand, and other user-specific factors. Figure 3 illustrates this variation, showing that while the average listener in the Long 2x arm received approximately 8 long ads, some received as few as 1-2 ads while others received 20 or more. This heterogeneity in realized exposure represents an important component of advertising costs that must be accounted

[8]Neural-network approaches have been shown to perform well for estimating heterogeneous treatment effects in high-dimensional settings (Farrell et al., 2021). Another key advantage of this multi-headed structure is that extending it to multiple treatment arms is straightforward, since each arm corresponds to an additional output head trained over the shared representation. By contrast, extending causal forests (Wager and Athey, 2018) beyond binary treatment typically requires more specialized methodology, such as generalized random forests or multivariate R-learner formulations (Athey et al., 2019).

for in personalized targeting. We therefore estimate heterogeneous treatment effects on total ad duration delivered, representing the expected cost of reaching each listener under different treatment assignments. We model the expected total ad duration (in seconds) as:

$$\mathbb{E}[\text{AdDuration}_i|\mathbf{x}_i, k] = \mu_k(\mathbf{x}_i). \tag{8}$$

Using an analogous multi-headed architecture with shared base layers and treatment-specific heads, we minimize a weighted mean squared error loss:

$$\mathcal{L}_{\text{cost}} = \frac{1}{N}\sum_{i=1}^{N} w_{k_i}\,(d_i - \hat{\mu}_{k_i})^2 \tag{9}$$

where $d_i$ is the realized total ad duration for listener $i$, computed as:

$$d_i = 10 \times \text{ShortAds}_i + 30 \times \text{LongAds}_i, \tag{10}$$

and $\hat{\mu}_{k_i} = \mu_{k_i}(\mathbf{x}_i)$ is the predicted duration under their assigned treatment. The weights $w_{k_i} = \frac{N}{N_k}$ are identical to those used in the returns model, ensuring balanced learning across the imbalanced treatment groups.

### *Training and Off-Policy Evaluation*

We randomly split the data into training (50%) and holdout (50%) sets, stratified by treatment assignment to maintain the experimental proportions in each split. The training set is further divided 80/20 for training and validation, also using the same stratification approach. Both neural networks are trained using AdamW optimizer (Loshchilov and Hutter, 2017) with early stopping based on validation loss. Importantly, the holdout set is never used during model training or selection, reserving it exclusively for off-policy evaluation.

For evaluating personalized policies, we employ off-policy evaluation using inverse probability weighting (IPW) on the holdout sample. Given a targeting policy $\pi$ that assigns listener $i$ with features $\mathbf{x}_i$ to treatment $\pi(\mathbf{x}_i)$, we need to estimate two quantities: the expected Modes usage and the expected total ad duration under this policy.

For each outcome, we use IPW estimation restricted to listeners whose experimental assignment matches the policy recommendation. For Modes usage:

$$\hat{V}_{\text{usage}}(\pi) = \frac{1}{N_H} \sum_{i \in \mathcal{H}} \frac{\mathbb{I}_{\{k_i = \pi(\mathbf{x}_i)\}}}{p_{k_i}} \cdot y_i^{\text{modes}}, \tag{11}$$

For total ad duration:

$$\hat{V}_{\text{duration}}(\pi) = \frac{1}{N_H} \sum_{i \in \mathcal{H}} \frac{\mathbb{I}_{\{k_i = \pi(\mathbf{x}_i)\}}}{p_{k_i}} \cdot d_i, \tag{12}$$

where $\mathcal{H}$ denotes the holdout set with $N_H$ listeners, $k_i$ is the treatment assigned to listener $i$ in the experiment, $\mathbb{I}_{\{k_i = \pi(\mathbf{x}_i)\}}$ indicates whether the policy recommendation matches the treatment assignment in the experiment, $p_{k_i}$ is the propensity score (0.10 for control, 0.15 for each treatment), $y_i^{\text{modes}}$ is the observed Modes usage (0/1), and $d_i = 10 \times \text{ShortAds}_i + 30 \times \text{LongAds}_i$ is the realized total ad duration in seconds. These estimators weight matched observations by the inverse of their assignment probability to obtain unbiased estimates of the realized outcomes under a given policy (Dudík et al., 2014; Swaminathan and Joachims, 2015).

### *Personalized Policy Construction*

Given a budget constraint on the total length of ads delivered, our goal is to maximize Modes usage across all listeners by optimally allocating different ad treatments. Using the neural network models learned above, where $\sigma(\theta_k(\mathbf{x}_i))$ represents the treatment-specific network output for returns and $\phi_k(\mathbf{x}_i)$ represents the treatment-specific network output for costs, we formulate this as:

$$\max_{\pi} \sum_{i=1}^{N} \underbrace{\sigma(\theta_{\pi(\mathbf{x}_i)}(\mathbf{x}_i))}_{\substack{\text{expected Modes} \\ \text{usage under policy } \pi}} \quad \text{s.t.} \quad \sum_{i=1}^{N} \underbrace{\phi_{\pi(\mathbf{x}_i)}(\mathbf{x}_i)}_{\substack{\text{expected total ad} \\ \text{duration under policy } \pi}} \leq B, \tag{13}$$

where policy $\pi$ assigns each listener to one of our seven experimental conditions (Control, Short, Short 2x, Short 3x, Long, Long 2x, or Combine), and $B$ represents the total advertising duration budget.

The optimization problem in (13) is a discrete optimization problem that is intractable to solve

directly. We therefore apply Lagrangian relaxation:

$$\mathcal{L}(\pi, \lambda) = \sum_{i=1}^{N} \sigma(\theta_{\pi(\mathbf{x}_i)}(\mathbf{x}_i)) - \lambda \left( \sum_{i=1}^{N} \phi_{\pi(\mathbf{x}_i)}(\mathbf{x}_i) - B \right). \tag{14}$$

For a fixed $\lambda \geq 0$, the problem decouples across listeners, yielding independent subproblems:

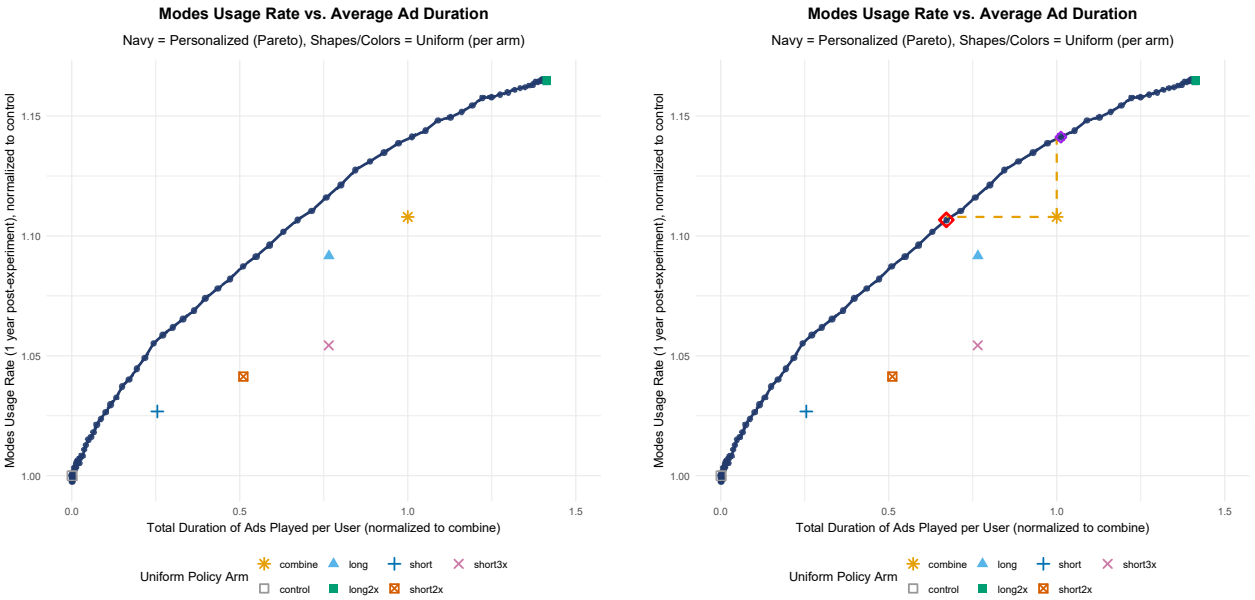
$$\pi_\lambda(\mathbf{x}_i) = \arg \max_{k \in \{0,1,\ldots,6\}} \{\sigma(\theta_k(\mathbf{x}_i)) - \lambda \cdot \phi_k(\mathbf{x}_i)\}. \tag{15}$$

The parameter $\lambda$ represents the shadow price of ad duration, that is, the marginal value of an additional unit of advertising budget in terms of increasing Modes usage. At $\lambda = 0$, the policy maximizes Modes usage without considering ad duration, potentially assigning most listeners to Long 2x (the most intensive treatment). As $\lambda$ increases, the policy shifts toward less ad-intensive treatments, balancing effectiveness against resource consumption.

By varying $\lambda$, we generate a frontier of policies ranging from maximum effectiveness (high Modes usage, high ad duration) to minimum cost (control assignment for all). Each point on this frontier represents an optimal allocation for a specific budget constraint, with both the expected Modes usage lift and expected advertising load per listener estimated using the IPW approach described in Equations (11) and (12) on the holdout sample. This setup builds directly on the welfare-based treatment choice literature under resource constraints (Kitagawa and Tetenov, 2018; Athey and Wager, 2021), with randomized experimental variation enabling us to trace feasible policy frontiers for allocating advertising exposure.

Figure 8 displays the Pareto frontier comparing personalized policies to uniform strategies. Because of our confidentiality agreement with Pandora, all measures of advertising load are normalized relative to the Combine strategy rather than reported in absolute seconds. The left panel shows the Pareto frontier (solid navy line) traced out by varying $\lambda$, with uniform strategy performance points plotted using distinct shapes and colors for each treatment arm. The right panel annotates the efficiency gains from personalization by showing two personalized counterparts to the Combine strategy (star shape): a red diamond to the left shows a personalized policy that achieves the same 10.8% lift in Modes usage with approximately one-third less advertising load (a 33% reduction), while a purple diamond positioned vertically above shows a personalized policy

that uses the same total advertising duration but achieves a 31% higher lift (from 10.8% to 14.1%). These results demonstrate that accounting for heterogeneity in treatment effects allows for more efficient resource allocation: personalized assignment of ad frequency and length based on listener characteristics can either maintain campaign effectiveness while substantially reducing ad load, or increase effectiveness while holding ad load constant.




(a) Pareto Frontier and Uniform Strategies

(b) Personalized Counterpart to Combine Strategy

Figure 8: Personalized vs. Uniform Advertising Strategies. The left panel displays the Pareto frontier of personalized policies (solid navy line) alongside the performance of uniform strategies (distinct shapes and colors). The right panel illustrates two personalized counterparts to the Combine strategy: a diamond marker (to the left) shows a policy achieving the same lift with one-third less ad duration, while a purple diamond (vertically above) shows a policy using the same ad duration but achieving 31% higher lift. For confidentiality, outcomes are normalized such that Control equals 1 on the usage axis (y) and Combine equals 1 on the total ad duration axis (x).

### *Sources of Personalization Gains*

The personalized targeting approach described above exploits the heterogeneous treatment effects learned by our neural network models along two dimensions: listener responsiveness to advertising treatments (returns) and realized advertising delivery conditional on treatment assignment (costs). To understand the relative contribution of each source of heterogeneity to the observed efficiency gains, we construct two alternative policies that use only one dimension while holding the other constant at its average treatment effect.

Let $\mu_k = \mathbb{E}[\sigma(\theta_k(\mathbf{x}_i))]$ denote the average treatment effect of assignment $k$ on Modes usage, and $\tau_k = \mathbb{E}[\phi_k(\mathbf{x}_i)]$ denote the average treatment effect on advertising duration. Our baseline *full personalization* policy, presented in Equation (15), uses individual predictions from both neural networks. The counterparts to this approach are the following restricted policies:

The *returns-only personalization* policy uses individual-level predictions for returns but average effects for costs:

$$\pi_\lambda^{\text{returns}}(\mathbf{x}_i) = \arg\max_{k \in \{0,1,\ldots,6\}} \left[\sigma(\theta_k(\mathbf{x}_i)) - \lambda \cdot \tau_k\right], \tag{16}$$

where $\tau_k$ represents the average treatment effect of assignment $k$ on advertising duration.

The *costs-only personalization* policy uses individual-level predictions for costs but average effects for returns:

$$\pi_\lambda^{\text{costs}}(\mathbf{x}_i) = \arg\max_{k \in \{0,1,\ldots,6\}} \left[\mu_k - \lambda \cdot \phi_k(\mathbf{x}_i)\right], \tag{17}$$

where $\mu_k$ is the average treatment effect of assignment $k$ on Modes usage.

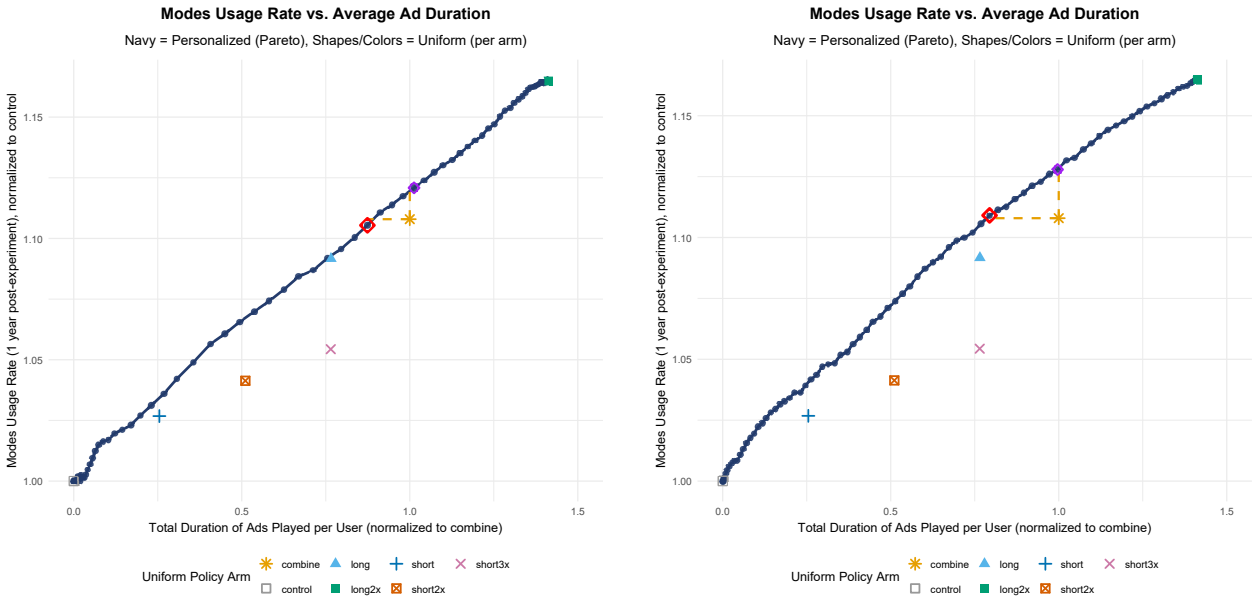
We construct Pareto frontiers for each policy by varying $\lambda$ and evaluate performance using IPW estimation as described in Equations (11) and (12). The performance gap between each restricted policy and the full personalization baseline quantifies the contribution of each heterogeneity dimension to overall efficiency gains.

Figure 9 displays the Pareto frontiers for our two restricted personalization policies, with the full personalization approach presented earlier in Figure 8 serving as the benchmark. The left panel shows the returns-only personalization frontier, which accounts for heterogeneous treatment effects but uses the average treatment effect for realized advertising delivery. The right panel shows the costs-only personalization frontier, which accounts for heterogeneous delivery patterns but uses the average treatment effect for treatment responses.

The results reveal that both sources of heterogeneity contribute meaningfully to personalization gains. Returns-only personalization achieves moderate efficiency improvements, reducing ad load by about 13% relative to the uniform Combine strategy. Costs-only personalization delivers larger gains, cutting ad load by about 20%. In comparison, full personalization (Figure 8) achieves the same performance with roughly one-third less ad load, a 33% reduction.

These findings indicate that modeling heterogeneity in ad exposure intensity accounts for ap-

proximately two-thirds of the total personalization benefit (20 percentage points of the 33 percentage point total reduction), while modeling heterogeneous treatment responses contributes the remaining one-third. The larger contribution from exposure intensity heterogeneity suggests that while much attention in advertising research focuses on whether to treat users at all (the extensive margin), substantial efficiency gains arise from optimizing how intensively to treat each user (the intensive margin).




(a) Returns-Only Personalization

(b) Costs-Only Personalization

Figure 9: Sources of Personalization Gains. The left panel displays the Pareto frontier for returns-only personalization, which accounts for heterogeneous treatment effects but uses the average delivery costs across conditions. The right panel shows the Pareto frontier for costs-only personalization, which accounts for heterogeneous delivery patterns but uses the average treatment effect across conditions. Both panels include the uniform Combine strategy performance for comparison. For confidentiality, outcomes are normalized such that Control equals 1 on the usage axis and Combine equals 1 on the total ad duration axis.

### *Understanding the Personalized Allocation Strategy*

To better understand how the personalized targeting policy operates in practice, we examine how treatment assignments vary across two key dimensions of heterogeneity identified in our previous analysis: (1) prior experience with the Modes feature, and (2) expected advertising exposure intensity. Specifically, we analyze the personalized policy that achieves equivalent effectiveness to the uniform Combine strategy while reducing total advertising load by 33% (the red diamond point in Figure 8). This analysis reveals the economic logic underlying the algorithm's allocation decisions

and connects our personalization results to the underlying treatment heterogeneity documented earlier.

*Segmenting users by exposure intensity*

Our ghost ads experimental design provides a unique opportunity to observe heterogeneity in realized advertising exposure that is comparable across all treatment arms. Recall that every listener in our experiment is simultaneously enrolled in exactly three campaigns (some combination of promotional and placebo campaigns depending on their treatment assignment). As shown in Table 2, treatment arms achieve similar average total advertising exposure (approximately 12 ads per listener across all three campaigns).

We construct our exposure intensity measure by summing the total number of ads that each listener received across all three of their enrolled campaigns. This total number of realized ads reveals each listener's underlying exposure level to the campaigns: some users naturally receive many ads due to their listening patterns, platform engagement, and advertiser demand for their demographic segments, while others receive relatively few ads even when targeted by the same campaigns.[9]

The ghost ads framework ensures this exposure measure is comparable across treatment arms because all listeners face the same underlying advertising delivery environment and receive the same total number of campaigns. The total number of ads received therefore captures each listener's individual exposure level independent of their experimental assignment, allowing us to identify high-exposure and low-exposure user types across all conditions.[10]

We segment listeners into exposure deciles by ranking all participants based on their total realized ad impressions across the three campaigns and dividing them into ten equal-sized groups. Users in lower deciles represent "light exposure" types who would receive relatively few ads under any treatment strategy, while those in higher deciles represent "heavy exposure" types who would receive many impressions regardless of treatment assignment. This segmentation captures

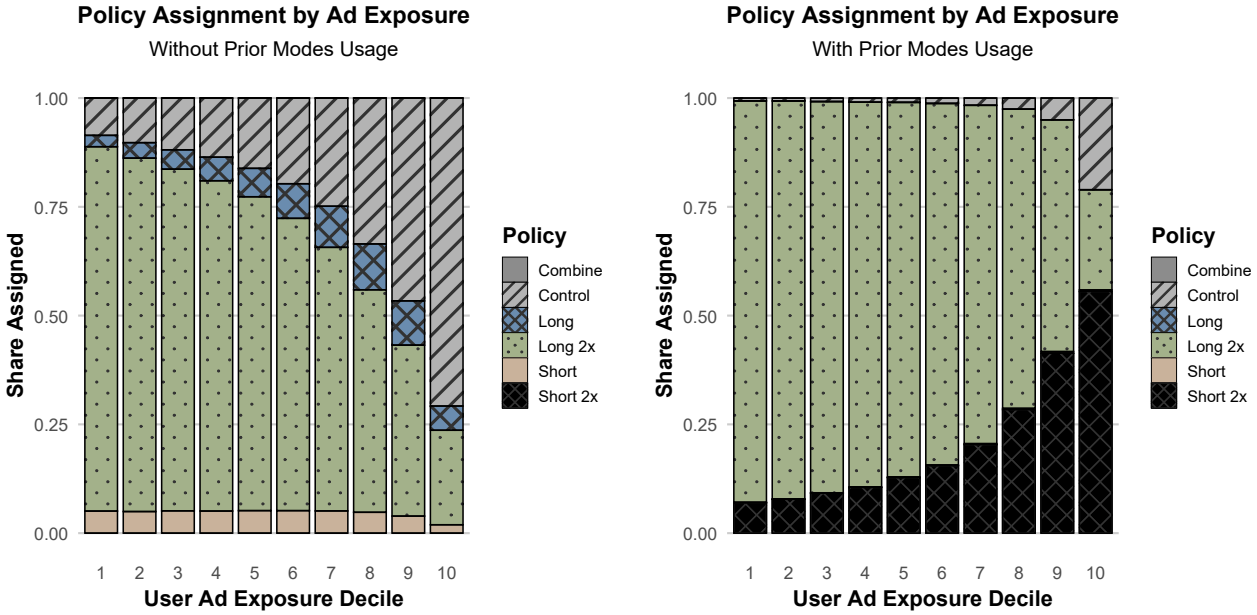
[9] As demonstrated earlier, this variation in total advertising exposure is substantial even within treatment arms. Figure 3 illustrates this heterogeneity by showing the distribution of long ads received within just the Long 2x treatment arm, where some listeners received as few as 1-2 long ads while others received over 20.

[10] Appendix E documents that these total exposure distributions are indeed similar across all experimental arms.

the intensive margin of advertising delivery that our personalization approach exploits to improve efficiency.

*Allocation patterns across user segments*

Figure 10 displays how the personalized policy assigns users across treatment conditions, segmented by exposure decile and prior Modes experience. Two interesting patterns emerge from examining how the algorithm allocates users across these dimensions.




(a) Users Without Prior Modes Experience

(b) Users With Prior Modes Experience

Figure 10: Personalized Policy Assignment by Ad Exposure Decile and Prior Experience. The figure shows how the personalized targeting algorithm assigns users across treatment conditions based on their expected advertising exposure intensity (deciles 1-10) and prior experience with the Modes feature. This policy achieves equivalent effectiveness to the uniform Combine strategy while reducing total advertising load by 33%. Panel (a) shows assignments for users without prior Modes experience, while Panel (b) shows assignments for users with prior experience.

First, consider users without prior Modes experience (left panel). The most interesting pattern is how the algorithm strategically reallocates campaign resources across different user types. Rather than uniformly distributing advertising exposure, the personalized strategy concentrates resources toward users with lower expected ad exposure while reducing allocation to users who would receive high ad exposure. For users in the highest exposure deciles, the algorithm frequently assigns them to control conditions (receiving no promotional ads). For users without prior Modes experience in

the highest exposure decile (decile 10), about 70% are assigned to control, enabling more efficient allocation of limited campaign resources across the entire user base. This resource reallocation strategy enables the same advertising effectiveness with significantly reduced total advertising load.

The second key pattern concerns how the algorithm treats users with prior Modes experience differently. For high-exposure familiar users, rather than assigning them to control (as it does for high-exposure unfamiliar users), the algorithm continues to extract value from these listeners by shifting toward more efficient ad formats. In the highest exposure deciles for users with prior experience, Short 2x becomes the dominant assignment, accounting for over 55% of allocations in decile 10. This demonstrates that the algorithm recognizes these familiar users still provide positive returns, but optimizes by serving them shorter rather than the intensive long-format campaigns. Short ads thus play an important role in the personalized strategy, delivering targeted reminder content where it performs most efficiently while freeing up campaign resources for informational advertising to unfamiliar users.

These allocation patterns reveal how personalized targeting achieves equivalent effectiveness while at the same time reducing advertising load. The algorithm operates simultaneously across three optimization margins: selecting which users to treat (extensive margin), determining how to treat them (format margin), and deciding treatment intensity (frequency margin). The contrast in the policy assignment rules across familiar and unfamiliar users demonstrates the value of matching advertising strategy to user information needs, while the systematic assignment of high-exposure users to control or more efficient formats illustrates how modeling heterogeneous delivery costs led to efficiency gains. By systematically identifying users who would consume many impressions and either removing them from promotional targeting or switching them to formats that improve lift per second of advertising, the algorithm balances heterogeneous treatment returns against heterogeneous delivery costs across the user population.

# Discussion and Conclusion

Our experiment highlights substantial heterogeneity in how different users respond to ads of varying lengths and frequencies. For listeners without prior exposure to the promoted feature, both short and long ads generate persistent effects lasting nearly two years after the experiment, with long ads generating 4.6 times the response despite being only three times longer. This persistence for both formats is consistent with an informational role for advertising among unfamiliar users, and the superior effectiveness of long ads among this segment suggests they may convey more useful content. In contrast, for listeners with prior experience with Modes, the effects of all advertising formats dissipate within weeks and the relative advantage of long ads diminishes, which aligns with a more reminder-like pattern. This heterogeneity demonstrates that advertising effectiveness depends fundamentally on the interaction between ad format and audience familiarity.

A natural limitation of our design is that changes in ad length necessarily imply changes in ad content. It is impossible to extend or shorten an advertisement without adjusting the script, tone, and pacing, and in practice advertisers aim to produce the best possible creative given the allotted time. Our process mirrored this industry norm: Pandora's marketing staff were instructed to generate the highest-quality copy feasible within the assigned length constraints. To ensure our results were not driven by a single exceptionally effective or ineffective creative, we created multiple ad copies within each format category (three short and three long Modes ads) that were randomly rotated by the ad server. Thus, while length and content are intertwined, this confounding reflects the actual decision environment facing advertisers. What our experiment identifies is therefore not the effect of "time alone," but the effectiveness of realistic campaigns designed for different durations. This interpretation is supported by the distinct patterns we observe: long ads appear particularly effective at conveying informational content to unfamiliar users, while both formats serve similar reminder functions for familiar users.

We also contribute by showing how personalization can substantially improve efficiency. A personalized policy that tailors both length and frequency can achieve the same ad effectiveness as a uniform mixed-format strategy (Combine) while reducing the ad load by one-third. The per-

sonalized policy works through three key mechanisms. First, it systematically assigns control (no promotional ads) to heavy-exposure unfamiliar users, those in the highest exposure deciles who would consume disproportionate campaign resources. Second, for heavy-exposure familiar users, rather than removing them from targeting entirely, the algorithm shifts them to more efficient formats like Short 2x, recognizing they still provide positive returns but optimizing the format to their reminder needs. Third, the algorithm concentrates intensive treatments (Long 2x) on light-to-moderate exposure unfamiliar users, where the informational content has the highest impact per second of advertising delivered.

Importantly, most of the efficiency gain comes from understanding heterogeneity in the realized number of impressions across users. The number of impressions varies widely across listeners, even within the same assignment arm, because of differences in listening patterns or competing advertiser demand for each segment. By predicting which listeners would receive many impressions and either removing them from promotional targeting or shifting them to more efficient formats, the platform can reduce total ad load while maintaining advertising effectiveness. Roughly two-thirds of the efficiency gain from personalization comes from capturing this heterogeneity in realized impressions. The remaining third comes from leveraging heterogeneous treatment responses, identifying which users are most responsive to different advertising formats. Taken together, these results show that effective personalization requires accounting for realized difference in number of impressions across users; without doing so, improvements from identifying responsive users alone would be much smaller.

Our experimental design builds on existing methodological approaches and allows us to create exogenous variation in both advertising intensity and formats. Standard ghost ads implementations randomize at the impression opportunity: in treatment, the ad server plays the promotional creative, while in control, it replaces the impression with a placebo or public-service ad. This creates extensive-margin variation but leaves exposure intensity uncontrolled, as some users naturally receive many impressions while others receive few depending on listening behavior or competing advertiser demand. We apply the intensive-margin extension of this framework by enrolling each listener in three parallel campaigns with identical targeting and frequency caps, and varying whether those campaigns are promotional or placebo as well as the format of the promotional ads. This

structure generates exogenous variation simultaneously on the extensive margin (whether a listener sees promotional ads), the intensive margin (how many impressions they receive), and the format of advertising (long versus short), enabling us to study how ad effectiveness varies along these dimensions.

In sum, this paper provides field evidence on the role of ad length and frequency for ad effectiveness, where the majority of research has come from surveys and lab experiments. For practitioners, the findings suggest that advertisers should be willing to pay a premium for longer ad spots when targeting unfamiliar audiences, as the effectiveness gain exceeds the length ratio. For reminder campaigns or to retarget familiar users, shorter formats may be more cost-effective given the smaller differential. More generally, advertisers and platforms can design targeting strategies that better match ad format to user information needs, allowing them to achieve the same campaign objectives with substantially lower advertising load. Personalizing the allocation of ad formats therefore provides a path toward less intrusive advertising on ad-supported platforms.

# References


**Ackerberg, Daniel A.**, "Empirically Distinguishing Informative and Prestige Effects of Advertising," *RAND Journal of Economics*, 2001, *32* (2), 316–333.

**Anderson, Simon P and Joshua S Gans**, "Platform siphoning: Ad-avoidance and media content," *American Economic Journal: Microeconomics*, 2011, *3* (4), 1–34.

_ **and Stephen Coate**, "Market provision of broadcasting: A welfare analysis," *The review of Economic studies*, 2005, *72* (4), 947–972.

_ **, Øystein Foros, and Hans Jarle Kind**, "Competition for advertisers and for viewers in media markets," *The Economic Journal*, 2018, *128* (608), 34–54.

**Ansari, Asim and Carl F Mela**, "E-customization," *Journal of marketing research*, 2003, *40* (2), 131–145.

**Athey, Susan and Stefan Wager**, "Policy learning with observational data," *Econometrica*, 2021, *89* (1), 133–161.

_ **, Julie Tibshirani, and Stefan Wager**, "Generalized Random Forests," *Annals of Statistics*, 2019, *47* (2), 1148–1178.

**Bagwell, Kyle**, "The economic analysis of advertising," *Handbook of industrial organization*, 2007, *3*, 1701–1844.

**Berman, Ron and Christophe Van den Bulte**, "False discovery in A/B testing," *Management Science*, 2022, *68* (9), 6762–6782.

**Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman**, "What's advertising content worth? Evidence from a consumer credit marketing field experiment," *The quarterly journal of economics*, 2010, *125* (1), 263–306.

**Biswas, Shirsho**, "Investigating the effects of including discount information in advertising," *Available at SSRN 4049330*, 2022.

**Blake, Thomas, Chris Nosko, and Steven Tadelis**, "Consumer heterogeneity and paid search effectiveness: A large-scale field experiment," *Econometrica*, 2015, *83* (1), 155–174.

**Bounie, David, Martin Quinn, and Morrisson Valérie**, "Advertising viewability in online branding campaigns," *Available at SSRN 2969891*, 2016.

**Brynjolfsson, Erik, Avinash Collis, Daniel Deisenroth, Haritz Garro, Daley Kutzman, Asad Liaqat, and Nils Wernerfelt**, "The Consumer Welfare Effects of Online Ads: Evidence from a 9-Year Experiment," *American Economic Review: Insights*, 2025.

**Drèze, Xavier and François-Xavier Hussherr**, "Internet advertising: Is anybody watching?," *Journal of interactive marketing*, 2003, *17* (4), 8–23.

**Dubé, Jean-Pierre and Sanjog Misra**, "Personalized pricing and consumer welfare," *Journal of Political Economy*, 2023, *131* (1), 131–189.

**Dudík, Miroslav, Dumitru Erhan, John Langford, and Lihong Li**, "Doubly robust policy evaluation and optimization," *Statistical science*, 2014, *29* (4), 485–511.

**Elsen, Millie, Rik Pieters, and Michel Wedel**, "Thin slice impressions: how advertising evaluation depends on exposure duration," *Journal of Marketing Research*, 2016, *53* (4), 563–579.

**Farrell, Max H, Tengyuan Liang, and Sanjog Misra**, "Deep neural networks for estimation and inference," *Econometrica*, 2021, *89* (1), 181–213.

**Fossen, Beth L, Philip Kim, and Inyoung Chae**, "EXPRESS: The Impact of Ad Length on Ad Effectiveness: Do Micro Ads Work?," *Journal of Marketing (forthcoming)*, 2025.

**Ge, Jiaoju, Yuepeng Sui, Xiaofeng Zhou, and Guoxin Li**, "Effect of short video ads on sales through social media: the role of advertisement content generators," *International Journal of Advertising*, 2021, *40* (6), 870–896.

**Gentzkow, Matthew, Jesse M Shapiro, Frank Yang, and Ali Yurukoglu**, "Pricing power in advertising markets: Theory and evidence," *American Economic Review*, 2024, *114* (2), 500–533.

**Goldfarb, Avi and Catherine Tucker**, "Online display advertising: Targeting and obtrusiveness," *Marketing Science*, 2011, *30* (3), 389–404.

_ **and** _ , "Digital economics," *Journal of Economic Literature*, 2019, *57* (1), 3–43.

**Goldstein, Daniel G, Siddharth Suri, R Preston McAfee, Matthew Ekstrand-Abueg, and Fernando Diaz**, "The economic and cognitive costs of annoying display advertisements," *Journal of Marketing Research*, 2014, *51* (6), 742–752.

**Goli, Ali, David H Reiley, and Hongkai Zhang**, "Personalizing ad load to optimize subscription and ad revenues: Product strategies constructed from experiments on pandora," *Marketing Science*, 2025, *44* (2), 327–352.

_ **, Jason Huang, David Reiley, and Nickolai M Riabov**, "Measuring consumer sensitivity to audio advertising: a long-run field experiment on Pandora internet radio," *Quantitative Marketing and Economics*, 2025, pp. 1–31.

**Gordon, Brett R, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky**, "A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook," *Marketing Science*, 2019, *38* (2), 193–225.

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

**Hermle, Johannes and Giorgio Martini**, "Valid and unobtrusive measurement of returns to advertising through asymmetric budget split," *arXiv preprint arXiv:2207.00206*, 2022.

**Hitsch, Günter J, Sanjog Misra, and Walter W Zhang**, "Heterogeneous treatment effects and optimal targeting policy evaluation," *Quantitative Marketing and Economics*, 2024, *22* (2), 115–168.

**Holmes, Todd A**, "Effects of self-brand congruity and ad duration on online in-stream video advertising," *Journal of Consumer Marketing*, 2021, *38* (4), 374–385.

**Johansson, Fredrik, Uri Shalit, and David Sontag**, "Learning representations for counterfactual inference," in "International conference on machine learning" PMLR 2016, pp. 3020–3029.

**Johnson, Garrett A, Randall A Lewis, and Elmar I Nubbemeyer**, "Ghost ads: Improving the economics of measuring online ad effectiveness," *Journal of Marketing Research*, 2017, *54* (6), 867–884.

**Johnson, Garrett, Randall A Lewis, and David Reiley**, "Location, location, location: Repetition and proximity increase advertising effectiveness," *Available at SSRN 2268215*, 2016.

**Johnson, Vinith, Zhen Zhu, Roger Anguera, Jacob Bollinger, Jonathan Eccles, David Hardtke, Maria Breza, and Theodore P Zanto**, "Increasing brand awareness: Memory for short audio ads," *Psychology & Marketing*, 2021, *38* (11), 1960–1972.

**Kalyanam, Kirthi, Raphael Thomadsen, and Nan Zhao**, "The Impact of Advertising Content on Customer Acquisition and Retention for Subscriptions of Physical Goods: Insights from a Field Experiment," *Available at SSRN 5338643*, 2025.

**Kitagawa, Toru and Aleksey Tetenov**, "Who should be treated? empirical welfare maximization methods for treatment choice," *Econometrica*, 2018, *86* (2), 591–616.

**Lemmens, Aurélie, Jason Roos, Sebastian Gabel, Eva Ascarza, Hernán Bruno, Brett Gordon, Ayelet Israeli, Elea McDonnell Feit, Carl Mela, and Oded Netzer**, "Personalization and targeting: How to experiment, learn & optimize," *International Journal of Research in Marketing*, 2025.

**Lewis, Randall A and David H Reiley**, "Online ads and offline sales: measuring the effect of retail advertising via a controlled experiment on Yahoo!," *Quantitative Marketing and Economics*, 2014, *12* (3), 235–266.

**Lewis, Randall A. and Justin M. Rao**, "The Unfavorable Economics of Measuring the Returns to Advertising," *Quarterly Journal of Economics*, 2015, *130* (4), 1941–1973.

**Lewis, Randall, Justin M Rao, and David H Reiley**, "Measuring the effects of advertising: The digital frontier," in "Economic Analysis of the Digital Economy," University of Chicago Press, 2015, pp. 191–218.

**Liaukonyte, Jura, Thales Teixeira, and Kenneth C Wilbur**, "Television advertising and online shopping," *Marketing Science*, 2015, *34* (3), 311–330.

**Loshchilov, Ilya and Frank Hutter**, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

**McGranaghan, Matthew, Jura Liaukonyte, and Kenneth C Wilbur**, "How viewer tuning, presence, and attention respond to ad content and predict brand search lift," *Marketing Science*, 2022, *41* (5), 873–895.

**Morozov, Ilya and Anna Tuchman**, "Where does advertising content lead you? We created a bookstore to find out," *Marketing Science*, 2024, *43* (5), 986–1001.

**Peters, Rik GM and Tammo HA Bijmolt**, "Consumer memory for television advertising: A field study of duration, serial position, and competition effects," *Journal of Consumer Research*, 1997, *23* (4), 362–372.

**Rafieian, Omid**, "Optimizing user engagement through adaptive ad sequencing," *Marketing Science*, 2023, *42* (5), 910–933.

_ **and Hema Yoganarasimhan**, "Targeting and privacy in mobile advertising," *Marketing Science*, 2021, *40* (2), 193–218.

_ **and** _ , "AI and personalization," *Artificial intelligence in marketing*, 2023, pp. 77–102.

**Rochet, Jean-Charles and Jean Tirole**, "Platform competition in two-sided markets," *Journal of the european economic association*, 2003, *1* (4), 990–1029.

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

**Sahni, Navdeep S**, "Effect of temporal spacing between advertising exposures: Evidence from online field experiments," *Quantitative Marketing and Economics*, 2015, *13* (3), 203–247.

**_ , S Christian Wheeler, and Pradeep Chintagunta**, "Personalization in email marketing: The role of noninformative advertising content," *Marketing Science*, 2018, *37* (2), 236–258.

**_ , Sridhar Narayanan, and Kirthi Kalyanam**, "An experimental investigation of the effects of retargeted advertising: The role of frequency and timing," *Journal of Marketing Research*, 2019, *56* (3), 401–418.

**Shalit, Uri, Fredrik D Johansson, and David Sontag**, "Estimating individual treatment effect: generalization bounds and algorithms," in "International conference on machine learning" PMLR 2017, pp. 3076–3085.

**Shapiro, Bradley T, Günter J Hitsch, and Anna E Tuchman**, "TV advertising effectiveness and profitability: Generalizable results from 288 brands," *Econometrica*, 2021, *89* (4), 1855–1879.

**Shiller, Benjamin Reed et al.**, *First degree price discrimination using big data*, Brandeis Univ., Department of Economics, 2013.

**Statista**, "Ad Spending," 2024.

**Swaminathan, Adith and Thorsten Joachims**, "Counterfactual risk minimization: Learning from logged bandit feedback," in "International conference on machine learning" PMLR 2015, pp. 814–823.

**Tåg, Joacim**, "Paying to remove advertisements," *Information Economics and Policy*, 2009, *21* (4), 245–252.

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

**Waisman, Caio and Brett R Gordon**, "Multicell Experiments for Marginal Treatment Effect Estimation of Digital Ads," *Management Science*, 2025.

**Wang, Bingcheng, Man Wu, Pei-Luen Patrick Rau, and Qin Gao**, "Influence of native video advertisement duration and key elements on advertising effectiveness in mobile feeds," *Mobile Information Systems*, 2020, *2020* (1), 8836195.

**Wang, Ziyu, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas**, "Dueling network architectures for deep reinforcement learning," in "International conference on machine learning" PMLR 2016, pp. 1995–2003.

**Wilbur, Kenneth C**, "A two-sided, empirical model of television advertising and viewing markets," *Marketing science*, 2008, *27* (3), 356–378.

# Contents of Web Appendices

# Web Appendix A: Audio stimuli

As audio stimuli for the experiment we created six ads for Modes (three short and three long) and four placebo ads (two short and two long). The directive for the marketing department was to create the most effective ads possible given the ad length constraints. We provide the complete transcripts below. For every audio ad, a display banner was shown alongside the audio content. We used identical display banners for short and long ads within each campaign type. The display banners are shown in Figure A1.

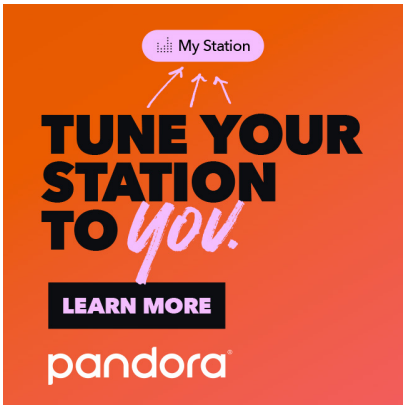
- Modes ads:
  - Short ad 1 (10s): *Want more control over your music? Now you can customize your favorite stations with Modes. Tap the "My Station" button to get started.*
  - Short ad 2 (11s): *Did you know you can customize your favorite stations' music? Switch Modes on your "My Station" menu to make your stations even better. Try Pandora Modes today!*
  - Short ad 3 (11s): *It is time to enjoy Pandora à la Mode. To try Modes and start customizing your stations, tap the "My Station" button at the top of the now playing screen.*
  - Long ad 1 (28s): *Wish you had more control over the music playing on your favorite stations? We have got a way to switch it up. Tap the "My Station" menu to change your mode to "Deep cuts", and we will play you lesser-known tracks from the albums you like. Or maybe you want to hear songs from artist or albums you have never heard on this station before? That's "Discovery mode". Switch to different modes on your "My Station" menu to make your top station even better. Even more you. Try it today!*
  - Long ad 2 (27s): *Are you ready to take control over your music and hear what you want to hear? Now you can switch it up with Modes. It is a new feature that lets you customize the music playing on your station. Choose to play only the hits with "Crowd faves", songs from a single artist with "Artist only", or lesser-known tracks with "Deep cuts". Ready to try a different mode? Just tap the "My Station" button and hit something new.*
  - Long ad 3 (27s): *It is time to enjoy Pandora à la Mode. In any artist station tap the "My Station" button at the top of the now playing screen to reveal the Modes menu and customize the music playing on your station. "Discovery mode" will spin a different mix of music. "Artist*

*only" will play songs exclusively from the artist you used to start the station. And "Deep cuts" will give you lesser-known tracks from the albums you love. It is like having Pandora with ice cream on top.*

- Placebo ads (Innovations for Poverty Action):
  - Short ad 1 (11s): *I am Dean Karlan, and I founded Innovations for Poverty Action to find the best ways to help families get out of poverty. You can help too, take action at poverty-action.org.*
  - Short ad 2 (11s): *I am Dean Karlan, and I founded Innovations for Poverty Action to find the best ways to help families get out of poverty. You can help too, please donate at poverty-action.org.*
  - Long ad 1 (27s): *Hi, my name is Dean Karlan. I am an Economist and founder of Innovations for Poverty Action. I have been working with poor households around the world for twenty years to find the best ways to help. Even before this crisis hit, many poor families didn't have enough food to eat, and now they have lost their jobs as well. At Innovations for Poverty Action we are working in twenty-two countries to find the best solutions. You can join too, take action at poverty-action.org.*
  - Long ad 2 (26s): *Hi, my name is Dean Karlan. I am an Economist and founder of Innovations for Poverty Action. I have been working with poor households around the world for twenty years to find the best ways to help. Even before this crisis hit, many poor families didn't have enough food to eat, and now they have lost their jobs as well. At Innovations for Poverty Action we are working in twenty-two countries to find the best solutions. You can join too, please donate at poverty-action.org.*

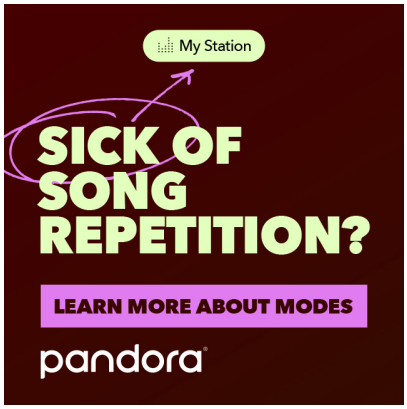
Figure A1: Display banners shown with audio ads

Modes ads




(a) Version 1


(b) Version 2




(c) Version 3

Placebo ads (Innovations for Poverty Action)

# Web Appendix B: Reduced-form comparison across arms

To understand the mechanisms underlying our main findings, we estimate logit regressions comparing select treatment arms while interacting treatment indicators with pre-experimental listener characteristics. The dependent variable is a binary indicator for any Modes usage during the 12-month post-experiment period (including the experimental period).

We estimate the following logit model:

$$\begin{aligned}\text{logit}(P(\text{Modes}_i = 1)) = \beta_0 &+ \beta_1 \cdot \text{Treatment}_i + \beta_2 \cdot \text{Treatment}_i \times \text{Familiar}_i \\ &+ \beta_3 \cdot \text{Treatment}_i \times \text{Male}_i + \beta_4 \cdot \text{Treatment}_i \times \log(\text{Income})_i \\ &+ \beta_5 \cdot \text{Treatment}_i \times \text{Age}_i + \beta_6 \cdot \text{Treatment}_i \times \text{Tenure}_i \\ &+ \beta_7 \cdot \text{Treatment}_i \times \log(\text{PreAdLoad})_i + \varepsilon_i \end{aligned} \tag{18}$$

where $\text{Familiar}_i$ indicates listeners who used Modes at least once in the two months before the experiment. The control variables include: $\text{Male}_i$ (gender indicator), $\log(\text{Income})_i$ (logarithm of median household income in the listener's ZIP code), $\text{Age}_i$ (user age in years), $\text{Tenure}_i$ (account age since registration), and $\log(\text{PreAdLoad})_i$ (logarithm of pre-experimental advertising intensity, measured as (ads received + 1)/(listening hours + 1) during the three weeks before the experiment). All continuous variables are standardized (z-scored) for interpretability.

Table 4 presents the results. The interaction coefficients reveal how treatment effectiveness varies across listener characteristics, connecting to our main findings about advertising's dual informational and reminder functions. For Short 3x versus Control, the main treatment effect is 0.079 log-odds, but this drops by 0.043 log-odds for familiar users, reducing effectiveness by 54%. For Long 2x versus Control, the main effect is 0.233 log-odds, falling by 0.184 log-odds for familiar users, a 79% reduction. This pattern confirms that long ads derive their effectiveness primarily from educating unfamiliar listeners rather than serving as reminders.

The direct comparison of Long 2x versus Short 2x shows a 0.168 log-odds advantage that is

almost entirely erased for familiar users (reduction of 0.155 log-odds, or 92%). Among unfamiliar listeners, long intensive campaigns substantially outperform short intensive ones, but this advantage nearly vanishes for those already aware of the feature. This supports the interpretation that both short and long ads function similarly as reminders for informed users.

The male interaction coefficients reveal that longer ad formats are differentially effective for men. While male listeners show no significant additional response to intensive short advertising (Short 3x vs Control: 0.008, not significant), they respond significantly more to long ad campaigns (Long 2x vs Control: 0.064, $p < 0.001$). The direct comparison confirms this pattern, with men showing a 0.066 log-odds greater benefit from Long 2x versus Short 2x campaigns ($p < 0.001$). These results corroborate our main findings and provide additional evidence that heterogeneous returns to advertising represent a key source of potential gains from personalized targeting strategies.

Table 4: Heterogeneous Treatment Effects: Logit Regression Results

| | Dependent variable: Any Modes Usage (12 months) | | |
|---|---|---|---|
| | Short 3x vs. Control | Long 2x vs. Control | Long 2x vs. Short 2x |
| | (1) | (2) | (3) |
| Treatment | 0.079*** | 0.233*** | 0.168*** |
| | (0.007) | (0.007) | (0.006) |
| Treatment × Familiar | −0.043*** | −0.184*** | −0.155*** |
| | (0.011) | (0.011) | (0.009) |
| Treatment × Male | 0.008 | 0.064*** | 0.066*** |
| | (0.010) | (0.009) | (0.008) |
| Treatment × Log Income (z-score) | 0.004 | 0.007 | 0.006 |
| | (0.005) | (0.005) | (0.004) |
| Treatment × Age (z-score) | −0.012*** | −0.054*** | −0.040*** |
| | (0.005) | (0.005) | (0.004) |
| Treatment × Tenure (z-score) | −0.008* | 0.021*** | 0.033*** |
| | (0.005) | (0.005) | (0.004) |
| Treatment × Log Pre-Ad Load (z-score) | −0.010** | −0.011** | −0.005 |
| | (0.005) | (0.005) | (0.004) |
| Observations | 4,584,333 | 4,585,406 | 5,493,349 |
| Log Likelihood | −753,886.900 | −805,897.400 | −967,876.600 |
| Akaike Information Criterion | 1,507,802.000 | 1,611,823.000 | 1,935,781.000 |

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

# Web Appendix C: Validating heterogeneous treatment effect predictions

Recall that our personalization analysis relies on neural network models to predict heterogeneous treatment effects for both Modes usage and advertising duration, as described in the personalization section. To validate the predictive accuracy of these models, we conduct an out-of-sample validation exercise using the holdout sample that was reserved exclusively for evaluation and never used during model training or selection.

Using the trained models, we generate predictions for all listeners in the holdout set and compare predicted treatment effects against realized outcomes. For each listener $i$ in the holdout set, we predict the lift in Modes usage from assignment to Long 2x versus Control as:

$$\text{Predicted Usage Lift}_i = \sigma(\theta_{\text{Long 2x}}(x_i)) - \sigma(\theta_{\text{Control}}(x_i)). \tag{19}$$

Similarly, we predict the lift in advertising duration delivered as:

$$\text{Predicted Duration Lift}_i = \mu_{\text{Long 2x}}(x_i) - \mu_{\text{Control}}(x_i), \tag{20}$$

where $\sigma(\theta_k(x_i))$ and $\mu_k(x_i)$ represent the neural network outputs for treatment condition $k$ given listener characteristics $x_i$, as defined in the personalization section above.

Note that while we can generate these predictions for all listeners, each individual listener was assigned to only one experimental condition and therefore has only one realized outcome. We cannot directly observe the counterfactual outcome (what would have happened under the alternative treatment) for any individual. To validate our predictions, we therefore bin listeners into deciles based on their predicted lift.

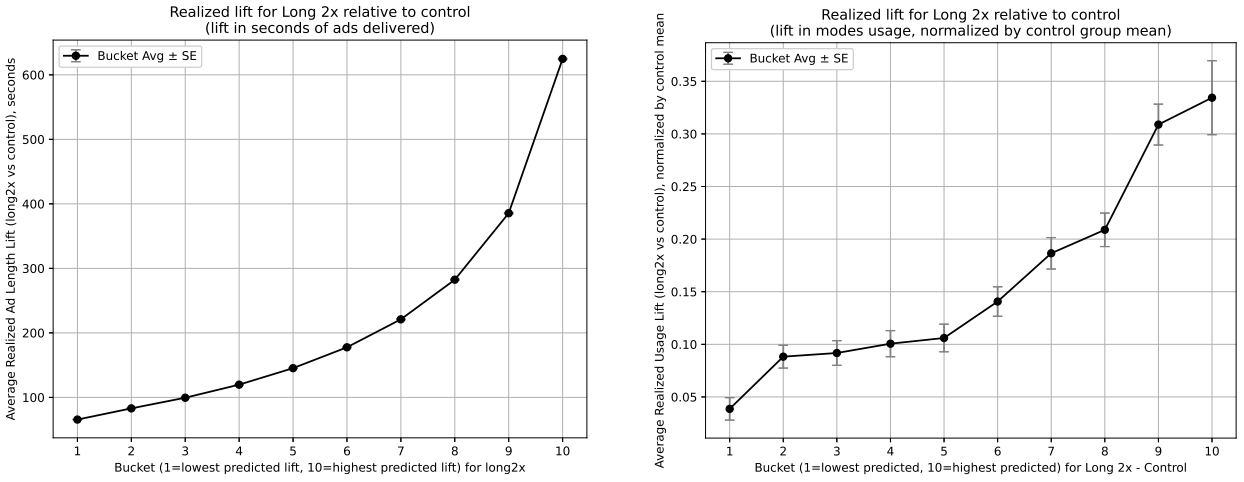
This binning approach works by sorting all holdout listeners by their predicted lift and dividing them into ten equal-sized groups (deciles). Each decile contains listeners with similar predicted treatment effects, but importantly, within each decile we have listeners who were randomly assigned to both Long 2x and Control conditions. For example, the top decile contains the 10% of listeners

predicted to have the highest treatment effects—some of these were assigned to Long 2x and others to Control. By comparing the average realized outcomes between Long 2x and Control listeners within this decile, we can estimate the realized treatment effect for high-predicted-response listeners. We repeat this process across all deciles to test whether our model's predictions align with realized heterogeneity in treatment effects. If the model is able to sort individuals based on treatment heterogeneity, we should observe larger treatment effects in deciles predicted to have higher lift, creating a monotonic relationship between predicted and realized effects.

Within each predicted lift decile, we estimate the average realized lift by running the following regression using only listeners assigned to Long 2x ($k_i = \text{Long 2x}$) or Control ($k_i = \text{Control}$):

$$Y_i = \alpha + \beta \cdot \mathbf{1}_{k_i=\text{Long 2x}} + \varepsilon_i, \tag{21}$$

where $Y_i$ represents the outcome of interest for listener $i$. For Modes usage, we normalize by the control group mean as in equation (1), while for advertising duration we use the seconds of promotional ads delivered. The coefficient $\beta$ captures the average treatment effect within that decile, representing the realized lift from Long 2x versus Control assignment.




(a) Advertising Duration Delivered

(b) Modes Usage Lift

Figure C1: Out-of-sample validation of heterogeneous treatment effect predictions. Panel (a) shows realized advertising duration differences (Long 2x - Control) across deciles of predicted advertising duration lift. Panel (b) shows realized Modes usage differences across deciles of predicted usage lift, with usage normalized by the control group mean. Both panels demonstrate strong monotonic relationships between predicted and realized treatment effects.

Figure C1 presents the validation results. Panel (a) shows that listeners predicted to receive more

advertising duration under Long 2x versus Control indeed receive substantially more advertising exposure, with realized differences ranging from approximately 100 seconds in the lowest decile to over 500 seconds in the highest decile. This demonstrates that our cost model successfully captures heterogeneity in ad delivery patterns.

Panel (b) validates our returns model by showing realized Modes usage lift across predicted lift deciles. The monotonic relationship confirms that listeners predicted to be more responsive to Long 2x versus Control advertising actually exhibit higher usage increases, with realized lift ranging from approximately 5% in the lowest decile to over 30% in the highest decile (relative to control group mean). For context, the overall Long 2x treatment achieved a 16.5% lift in Modes usage relative to control in the aggregate 12-month analysis (Figure 7), and the validation exercise demonstrates that our personalization models can accurately predict which listeners will exhibit treatment effects above or below this average.

The monotonic relationships between predicted and realized effects in both panels demonstrate that our neural network models successfully rank-ordered listeners by their treatment responsiveness. This out-of-sample validation supports the reliability of our personalization analysis, which depends critically on accurate predictions of heterogeneous treatment effects. The models correctly identify both which listeners are most responsive to different advertising treatments and which listeners are likely to receive more advertising exposure under different assignment strategies.

# Web Appendix D: Active users across treatment arms

We examine whether treatment assignment affected platform engagement directly, which could confound our estimates of advertising effectiveness on Modes usage. We define a user as active if they consumed any content on the platform (more than zero listening hours) during a given period.

Following a similar specification to our main analysis in equation (2), we estimate treatment effects on platform engagement, with outcomes normalized relative to control. Figure D1 displays the estimated coefficients over the 104-week observation period. The treatment effects on active user rates are economically ($< 0.1\%$) and statistically insignificant throughout the entire panel. This pattern holds from the experimental period through the full two-year observation window, confirming no meaningful differential impact on platform engagement across treatment arms.

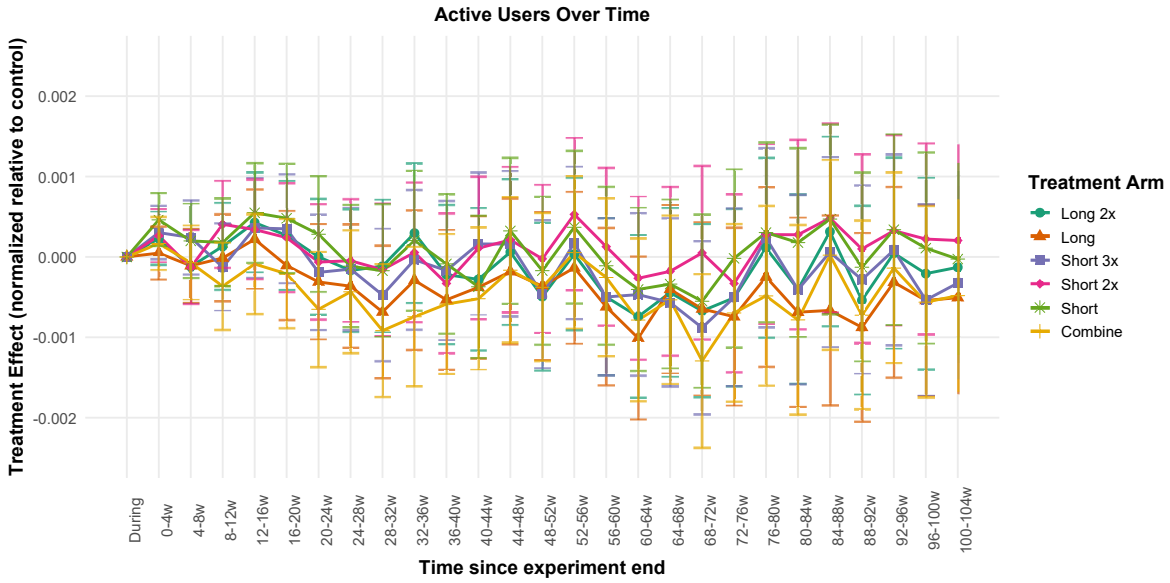



Figure D1: Treatment Effects on Platform Engagement Over Time. The figure shows estimated treatment coefficients for active user rates (those with non-zero listening hours) relative to control, from the experimental period through 104 weeks post-experiment. The near-zero coefficients indicate that advertising treatments did not differentially affect platform consumption.

The absence of consumption effects validates that the observed differences in Modes usage reflect advertising effectiveness rather than differential platform exposure opportunities.

# Web Appendix E: Total ad exposure distribution across treatment arms

Our experimental design enrolled each listener in exactly three campaigns, with varying combinations of promotional and placebo campaigns depending on treatment assignment. Figure E1 presents the distribution of total ads received (promotional plus placebo combined) across all experimental arms.

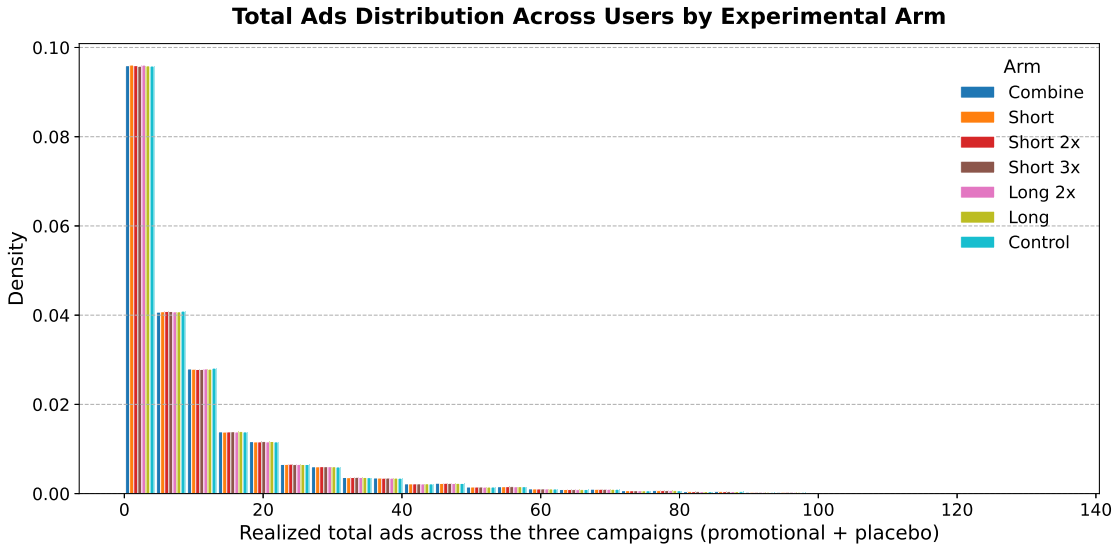



Figure E1: Distribution of Total Realized Ad Impressions Across Experimental Arms. The histogram displays the distribution of total ads received (promotional plus placebo combined) for listeners across all seven experimental arms, with each color representing a different treatment condition.

The median listener received 6 total ads across all arms. However, there is substantial heterogeneity in realized exposure: some listeners received as few as 1-2 ads while others received 30 or more ads. The distributions show similar patterns across all seven experimental conditions. This similarity in total ad exposure distributions across arms is relevant for our personalization analysis, where we examine how the personalized policy allocates treatments across different user types. Since all listeners face the same advertising delivery environment and receive the same number of campaigns (three), the variation in total impressions reflects individual-level heterogeneity in exposure patterns that is comparable across experimental conditions. This comparability validates our

approach of segmenting users by exposure intensity deciles when analyzing the allocation patterns under the personalized policy.