Measuring Consumer Sensitivity to Audio Advertising: A Field Experiment on Pandora Internet Radio

Jason Huang
David H. Reiley
Nickolai M. Riabov *

April 21, 2018

Abstract

A randomized experiment with almost 35 million Pandora listeners enables us to measure the sensitivity of consumers to advertising, an important topic of study in the era of ad-supported digital content provision. The experiment randomized listeners into nine treatment groups, each of which received a different level of audio advertising interrupting their music listening, with the highest treatment group receiving more than twice as many ads as the lowest treatment group. By keeping consistent treatment assignment for 21 months, we are able to measure long-run demand effects, with three times as much ad-load sensitivity as we would have obtained if we had run a month-long experiment. We estimate a demand curve that is strikingly linear, with the number of hours listened decreasing linearly in the number of ads per hour (also known as the price of ad-supported listening). We also show the negative impact on the number of days listened and on the probability of listening at all in the final month. Using an experimental design that separately varies the number of commercial interruptions per hour and the number of ads per commercial interruption, we find that neither makes much difference to listeners beyond their impact on the total number of ads per hour. Lastly, we find that increased ad load causes a significant increase in the number of paid ad-free subscriptions to Pandora, particularly among older listeners.

JEL codes: C93, D12, L82

*Huang: Stanford University, Uber , jason.yoshi.huang@gmail.com. Reiley: Pandora Media Inc., and University of California at Berkeley, david@davidreiley.com. Riabov: Brown University, Netflix, Inc., nriabov@netflix.com.
This work was completed while Huang and Riabov were completing internships at Pandora. We wish to recognize Chris Irwin and Jack Krawczyk for their leadership and vision in getting this experiment implemented. We are grateful to Pandora for allowing us to present these results to the scientific community; our agreement with management is to present our scientific results without discussing policy implications for the company. We thank Neil Buckman, Alexa Maturana-Lowe, Isaac Hill, Ashley Bischof, Carson Forter, and Puya Vahabi for considerable help in producing and organizing the data. We also thank Michael Hochster, David Hardtke, Oliver Bembom, Adam McCloskey, Andrew Asman, Sam Lendle, Hongkai Zhang, Zhen Zhu, Garrett Johnson, Randall Lewis, Jura Liaukonyte, and seminar participants at Brown University, UC Berkeley, Central European Institute, and University of Tennessee for helpful comments.

1 Introduction

In the Internet era, two-sided markets have taken on considerable importance with the offerings of companies as diverse as eBay, Match.com and Uber. Free, advertising-supported content is a particularly large and interesting class of two-sided markets (Rysman (2009)), including websites as diverse as Yahoo, Google, Facebook, YouTube, CNN.com, and a variety of other online news and entertainment sites. Platforms producing ad-supported content, like other two-sided markets, involve some amount of tradeoff between the goals of advertisers on one side and consumers on the other. In particular, consumers usually find it distracting or annoying to have their content consumption interrupted by advertising, but this is the price they pay for the consumption of their free content. In this paper, by varying the number of ads in a long-term randomized experiment, we measure the demand curve for free ad-supported music listening on Pandora, the world's largest Internet radio service.[1] To our knowledge, this is the first study to measure the extent to which customers are willing to reduce their consumption of a digital good as the amount of advertising (aka the price) is increased.

Pandora launched its Internet radio service in 2005, and as of the time of this study, it operates in the United States, Australia, and New Zealand. The service began with the Music Genome Project, which classifies each song on hundreds of musicological attributes in order to assess similarity between songs. A listener can specify a song, artist, album, or genre that she likes, and the Pandora service uses that information to create a station recommending similar music for that listener. The listener can further personalize her station by clicking thumbs-up or thumbs-down feedback on each song. Pandora algorithms use this thumb data, with over 50 billion thumbs recorded to date, to supplement the Music Genome, recommending music that has been liked by listeners with similar tastes. A team of scientists at Pandora has for years been developing and improving the recommendation algorithms. More recently, Pandora has also started hiring scientists to improve the advertising side of the business, and this paper is an example of the work of the ad science team.

Like many other online content services, Pandora earns its income primarily through advertising, which at first was digital display advertising displayed to listeners only when they looked at the Pandora website. In 2008, Pandora developed a smartphone app to deliver music on mobile devices. Since smartphones have very little screen space on which to display graphical ads, and since listeners typically spend most of their listening time not looking at the website or the mobile app, in December 2008 Pandora introduced audio ads. Audio ads command attention by interrupting the music, just as they do in traditional broadcast radio, though the digital world offers much more scope for the targeting of ads based on listener characteristics. When an audio ad runs on Pandora, if the listener is looking at the Pandora app or website during the ad, she will see a companion display ad tile on the screen. If she looks at the screen after the audio ad has ended, she will see a follow-on display ad for the most recent audio ad. Both of these companion graphics are clickable, just like standard online display advertisements. While Pandora separately sells both display and video advertising ad campaigns for both web and mobile devices, today audio advertising represents the majority of Pandora's advertising revenue.

As of the time of the study, Pandora had approximately 80 million monthly active listeners, each listening an average of about 20 hours per month. More than 80% of listening now takes place on mobile devices. The remainder takes place on the website as well as on other consumer-electronics devices such as Roku, AppleTV, Sonos speakers, and automobile stereo systems. Given engineering resource constraints, we chose to focus on mobile listening in our experiment. In other words, while our experiment generates variation in the quantity of advertising delivered to each listener, these deliberate differences take place only in the 80% of listening that takes place on mobile devices. Listeners in different treatment groups receive the same advertising treatment when listening on the website or on other consumer-electronics devices; their treatment differs only for the listening that takes place via iOS and Android apps.

This ambitious experiment generates a very precise estimate of Pandora listeners' sensitivity to advertising. By varying the number of ads each listener receives, we measure the long-run effects of advertising intensity on a variety of outcomes, most notably the total quantity of hours listened by Pandora users. Because we experiment with nine different levels of advertising, we are able to measure a continuous curve

[1] Because Pandora also pays royalties to musicians and promotes their music to listeners, we might actually consider its platform to be a three-sided market. But in this paper, we focus mainly on listeners and advertisers rather than the music industry.

that traces out the relationship between the ad load and the number of listening hours, and show that it is strikingly linear over a range that involves doubling the total quantity of audio advertising heard by listeners.

Using observational data relies on untestable assumption of unconfoundedness, which states that the set of observable variables properly controls for all confounding factors. Although we can always condition on observable variables, an endogenous treatment may well be correlated with unobservable variables, causing bias. By contrast, experiments allow us to avoid having to make such untestable assumptions in order to identify the causal estimates of ad load; the randomized experiment guarantees that the effects we measure are true causal effects rather than spurious correlation.[2]

Nevertheless, field-experimental measurements of how customers respond to prices are still extremely rare. Amazon experimented with randomized prices to customers in 2000-2001, but the experiment was not well-received by the general public. Customers compared notes online and felt it was unfair that they were receiving different prices from each other. CEO Jeff Bezos publicly stated that it was a "mistake because it created uncertainty for customers rather than simplifying their lives." (PSBJ (2000)) The company announced that if it ever again tested differential pricing, it would subsequently give all buyers the lowest price tested. We are not aware of any reports of Amazon ever resuming this type of randomized pricing across individuals, but we have (as individual consumers) noticed seemingly random changes in Amazon prices over time. Perhaps due in part to this bad public-relations outcome of the Amazon experiment, most firms have been reluctant to run price experiments to estimate demand. Pandora's decision to run this experiment is an important exception to this rule.

Field experiments in development economics have often performed binary pricing tests, usually finding large effects of changing the price from zero to some positive (but subsidized) price. Examples include Miguel et al. (2002) on free versus \$0.30 deworming medicine, Cohen & Dupas (2007) on free versus \$0.60 mosquito nets, and Ashraf et al. (2010) on free versus positive prices for water treatment. Berry et al. (2015) advocate the use of the Becker-DeGroot-Marschak (BDM) mechanism to estimate demand by eliciting each individual's willingness to pay, though in their water-filtration application they find some tendency for BDM to underestimate demand relative to an experiment that varies posted prices. Lusk et al. (2010) reviews the use of BDM and other lab-in-the-field auction mechanisms to estimate demand for food products in the United States; Lusk notes privately to us via email that his repeated attempts to run pricing experiments in grocery stores have generally failed due to implementation errors by grocery personnel. A rare published example of a posted-price experiment in the developed world is that of Gneezy et al. (2012), who varied the price (from \$5 to \$15) of a souvenir photo offered to thousands of tourists at the conclusion of a boat sightseeing tour.

To the best of our knowledge our study is the first field experiment to measure the extent to which consumers avoid ads by choosing to consume less media content as the ad load increases. Perhaps most closely related to our work is the research of Hohnhold et al. (2015), who vary the quality and quantity of advertising served in Google search results, finding that increased quantity and decreased quality increase the "ads blindness" of consumers, making them less likely in the long run to click on ads. Our exercise differs from theirs in that we estimate not just the effects on consumers' interactions with ads, but the extent to which consumers reduce their consumption of music on Pandora as a result of an increase in ad volume.

Our research design has a number of advantages. First, we are able to run a field experiment, allowing us to avoid the untestable assumptions required to infer causal effects from observational data. Our online setting made it possible for us to implement the randomized experiment without administrative errors or subject noncompliance. Also, because our price is an experiential one (how many ads am I listening to?), and Pandora's audio ad volume naturally varies across individuals for a variety of institutional reasons described

[2] A nice example of an observational study is the recent paper of Cohen et al. (2016), who use a regression-discontinuity analysis to estimate the demand curve for Uber based on discrete changes in "surge pricing." Because the authors are able to observe a continuous scarcity state variable that leads to discrete changes in price, they are able to argue that comparisons above and below each price-change threshold are as good as an experiment. This argument would be incorrect if consumers and drivers were somehow gaming their behavior to act differently just above a threshold than they do just below a threshold, which seems unlikely. A stronger concern about their study is how representative their results may be, since high surge prices in their study apply mainly to those who travel late at night or during morning rush hour, and these demand elasticities might not be representative of all Uber riders. Finally, because they exploit variation at the level of an individual ride, they are measuring a short-run elasticity, which might well be smaller than the long-run elasticity that would result if a treatment consumer received consistently high prices and got used to Uber being relatively expensive versus other transit options. Our experiment scores well on both representativeness and on measuring the relevant long-run demand elasticity, because we apply experimental variation in prices to all Pandora listeners over a long period of time.

below, the risks of consumers feeling unhappy about perceiving unfair differences in treatment are lower. Second, we managed to expose listeners not just to a binary treatment, but to nine different intensities of advertising, with the maximum treatment approximately twice as large as the minimum treatment. Third, we conduct the experiment for 21 months, enabling us to estimate long-run effects on listening behavior. Fourth, our sample size in the millions allows us to estimate the treatment effects with considerable precision. The results add to our scientific understanding of ad-avoidance behavior by consumers, which contributes in turn to our understanding of two-sided markets for ad-supported free content.

The remainder of this paper proceeds as follows. In the next section, we describe the details of the experimental design and provide validation of the randomized treatment assignment. Next we present our main results on the sensitivity of listeners to advertising intensity, including outcomes of hours listened, days listened, and probability of listening at all. In section 4, we measure the increased demand for a substitute product (ad-free subscriptions) due to increased advertising load. Next we look at heterogeneous treatment effects, finding interesting differences between older and younger listeners. In section 6 we take advantage of the experimental design to look at the effect of the number of commercial interruptions. The final section concludes.

2 Experimental Design

Between June 2014 and April 2016, Pandora Internet Radio conducted a large-scale experiment to measure the sensitivity of listeners to the number of audio ads they hear interrupting their music stream. During this time period, we randomized 19% of listeners into one of ten different treatment groups: nine treatment groups each with 1% of listeners, plus a 10% control group receiving the Pandora status quo. Each treatment group received different levels of audio advertising when they listened to Pandora via their mobile devices, with an individual's treatment kept consistent for the entire period of the experiment. New listeners joining Pandora were assigned to treatment using the same randomization (a hashing function of user ID with a unique randomization seed not used in any other current or historical experiments at Pandora).

We implemented a 3x3 experimental design, separately varying the number of commercial interruptions per hour and the number of ads per commercial interruption. At the time the experiment began, the status quo was for listeners to receive four commercial interruptions per hour. We will often refer to each commercial interruption as a "pod" of ads. The status quo at the start of the experiment was for pods to alternate deterministically between one and two ads per interruption, for an average of 1.5 ads per pod. The experiment assigned listeners to receive either 3, 4, or 6 interruptions per hour, and to receive either 1, 1.5, or 2 ads per interruption, with each of the 9 possible combinations getting its own 1% group of listeners. We will use the following shorthand to refer to these treatments: 3x1 for three 1-ad pods per hour, 6x1.5 for six pods per hour alternating between one and two ads per pod, and so on. We note that the large 10% control group was redundant with the status-quo 4x1.5 treatment. By varying ad load in these two separate dimensions, we enable ourselves to measure whether listeners have a preference for more commercial interruptions versus longer commercial interruptions, conditional on the platform's desired number of ads per hour.

The actual number of ads received by an individual will usually differ from the number of ads just described in our experimental design. This happens for a number of institutional reasons. First, ad delivery depends somewhat on listener behavior. For example, ads are delivered according to a given schedule, and a listener who listens to only a single song might not receive any ad at all during that listening session. Also, due to resource constraints, Pandora chose to implement the experiment only in its mobile application (Android and iOS), not on the Pandora website or on other pieces of client software, such as those in automobiles. Listening on mobile devices constituted approximately 80% of Pandora listening during this time period, so the experiment does generate considerable variation in advertising exposure to individuals, but listeners who listened only on the website from their desktop computers would receive no effective differences in treatment. The total amount of realized experimental treatment, therefore, depends on how much listening an individual does on different devices. In our analysis, we consider the impact on total listening across all devices, rather than restricting attention to listening on mobile devices, since many consumers listen via multiple devices and we would expect listener perceptions of Pandora's value to depend on the total amount of advertising they receive during all listening sessions they do.

Second, Pandora's ad delivery also depends on advertiser demand. Unlike in many digital advertising markets, there is no well-developed auction market for online audio advertising. Instead, all audio ads on Pandora are sold via forward contracts with advertisers. These contracts generally specify targeting attributes (such as "males aged 25-44 who live in San Diego and who are listening to punk-rock music"), as well as frequency caps that prevent the same listener from hearing the same advertisement more than a specified number of times per day or week. Given these delivery constraints, Pandora sometimes has no appropriate ad to play for a given listener on a given scheduled ad-delivery opportunity. At such moments, the ad opportunity will go unfilled and the listener will get to listen to more music instead of the scheduled advertisement. The realized "fill rate" (fraction of Pandora ad-delivery opportunities that actually get filled with an ad) can vary with listener attributes, with the time of day, and with the time of year.

Thus, we observe a considerable difference between the intended ad load and the realized ad load for each listener, with the latter being what we expect to affect listener behavior. In our analysis, we therefore use instrumental-variables estimation, as treatment assignment causes plenty of exogenous variation in realized ad load, amidst other potentially endogenous variation.

We consider several different outcome measures in this experiment. Listeners may react to the increased "price" of additional audio advertising by reducing their listening in one of several ways. We consider total hours listened, number of days active in a given month, and probability of listening at all to Pandora during a given time period. Finally, we measure the impact on the probability of purchasing an ad-free subscription to Pandora, which represents a substitute (costing approximately $5 per month) for the ad-supported version of the music service. For confidentiality reasons, we have normalized the listening hours and days metrics by dividing all observations by the control-group mean and multiplying by 100. We remind the reader of the normalization with the abbreviation "norm." in our tables of results.

We next validate our randomized experiment by verifying that treatment assignment is uncorrelated with other variables of interest. In particular, we compute means of the outcome variables during the pretreatment period, which we define as the month of May 2014. Table 1 displays comparisons of the outcome variables for the two treatment groups with the highest (6x2) and lowest (3x1) ad loads, showing that the variables are fairly similar across treatment groups. Each observation corresponds to a listener who used the ad-supported version of Pandora at least once during the experiment period. We next perform a χ^2 test for the equality of these means over all ten of the treatment groups (nine plus control, where control receives the same treatment as the 4x1.5 treatment group). As can be seen in table 2, each of these tests fail to reject the null hypothesis of equality at the 5% level, giving us a successful randomization check.

Table 1: Pretreatment Period Summary Statistics (Means and Standard Errors)

	Control	Lowest Ad Load	Highest Ad Load
Total Hours (norm.)	100.000	100.042	100.026
	(0.070)	(0.217)	(0.220)
Days Active (norm.)	100.000	99.943	99.943
	(0.043)	(0.137)	(0.137)
Audio Ads (norm.)	100.000	100.029	99.928
	(0.072)	(0.228)	(0.228)
Audio Pods (norm.)	100.000	100.001	99.956
	(0.071)	(0.223)	(0.224)
Ad Capacity (norm.)	100.000	100.057	99.946
	(0.068)	(0.216)	(0.216)
Percent Paid Users	1.098	1.097	1.104
	(0.002)	(0.008)	(0.008)
Percent Male	46.681	46.618	46.641
	(0.012)	(0.037)	(0.037)
Sample Size	18,342,916	1,833,826	1,831,909

Table 3 shows the means of the amount of treatment received between June 2014 and March 2016[3] for the

[3]Technically, the experiment ended on April 7, 2016. For expositional convenience, our outcome period will be the "final

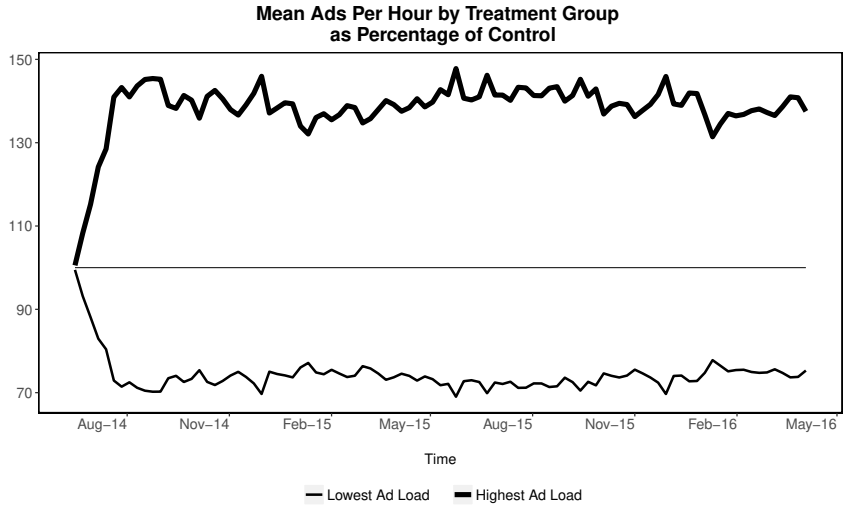

Figure 1: Realized Mean Ad Load per Hour by Treatment Group

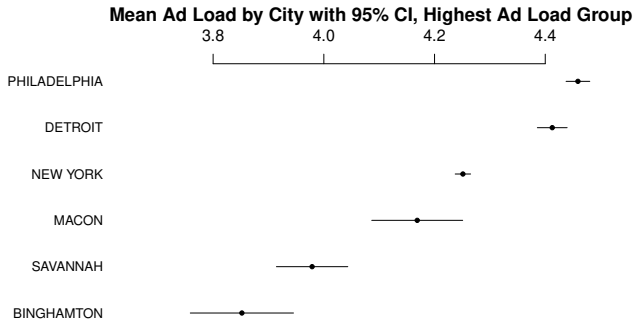

Figure 2: Mean Final Month Ad Load by Place of Residence, Highest Intended Treatment Group

Table 2: χ^2 Test for Equality of Means Across All Treatment Groups, Pretreatment Period

	Test Stat.	P Value
Total Hours	4.335	0.888
Days Active	3.562	0.938
Audio Ads	7.372	0.598
Ad Capacity	4.734	0.857
Paid Users	9.809	0.366
Male	7.362	0.599

Table 3: Treatment Period Summary Statistics

	Control	Lowest Ad Load	Highest Ad Load
Audio Ads per Hour	3.622	2.732	5.009
	(0.001)	(0.001)	(0.002)
Audio Ad Pods per Hour	2.999	2.645	3.521
	(0.000)	(0.001)	(0.002)
Ad Capacity per Hour	5.274	3.601	8.130
	(0.005)	(0.009)	(0.032)
Percent with Non-empty Pods	93.408	93.248	93.578
	(0.006)	(0.019)	(0.018)
Audio Ads per Pod	1.228	1.054	1.428
	(0.000)	(0.000)	(0.002)
Sample Size	18,342,916	1,833,826	1,831,909

highest and lowest treatment groups (6x2 and 3x1) compared with the control (4x1.5). We see that treatment assignment does, as intended, manipulate the realized ad load. The highest treatment group receives 37% more ads per pod and 33% more pods per hour than the lowest treatment group, for a total of 80% more ads per hour. For reference, the third row of the table shows the mean ad capacity per hour, or the number of audio ads that listeners would have received if every ad opportunity were filled by an advertiser. These numbers differ from the intended ad load numbers (6x2=12, 4x1.5=6, and 3x1=3) for several reasons having to do with the details of ad serving on Pandora. For example, listening sessions of different lengths can result in different numbers of ads per hour due to the details of the timing of ads. The most notable reason for differences is that the experiment manipulated advertising only for those listening via mobile apps, not for those listening via the Pandora website.

Figure 1 shows how the amount of treatment varies from one week to the next. This time series plot shows the amount of ad-load treatment received by the highest (6x2) and lowest (3x1) treatment groups, divided by the ad load of the control group. We point out two key features. First, the experiment was designed to ramp up slowly over a period of six weeks, visible at the left side of the graph, in case listeners would find a sudden increase in ads to be jarring. Second, the amount of treatment-control difference varies somewhat over time, since realized ad load differs from intended ad load in a way that depends on advertiser demand. In particular, we can see that the treatment effect is a bit higher in December than it is in January, because advertiser demand is very high in December and relatively low in January.

While we will use the experiment to identify differences in listening behavior across treatment groups, it is interesting to note that there also exists considerable variation in ad load within each treatment group. Advertisers' demand varies, for example, by different ages, genders, and cities of residence (DMAs). Figure 2 illustrates the variation in realized ad load across six randomly selected cities, within the highest treatment group. The overall distribution of realized treatment within this treatment group can be seen in Figure 3.

month" of the experiment, defined as March 8 to April 7, 2016.

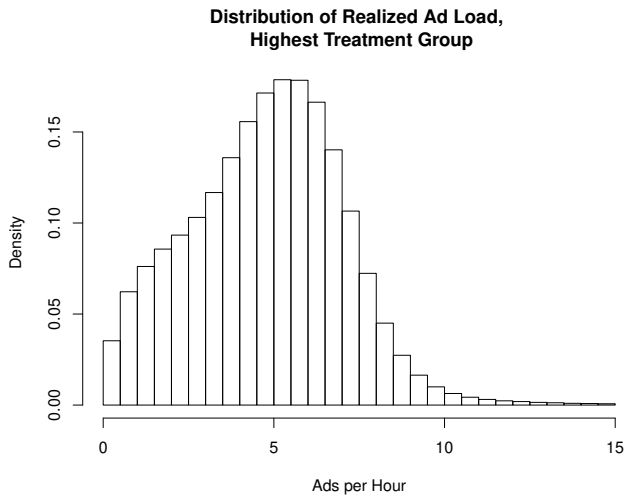

Figure 3: Distribution of Ad Load, Highest Intended Treatment Group

3 Measuring the Sensitivity of Listeners to Advertising

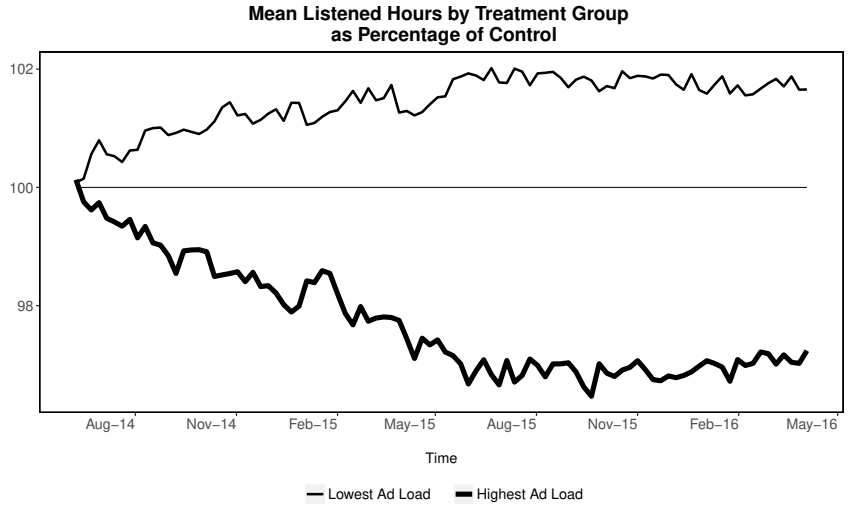

Figure 4: Mean Total Hours Listened by Treatment Group

Figure 4 plots the number of hours listened each week in the highest and lowest ad-load treatments, each relative to the control group. Total hours listened diverges relatively gradually from that of the control group, with the highest ad-load treatment group gradually listening fewer and fewer hours relative to control, while the lowest ad-load group gradually listens more and more hours. Figure 5 shows that this gradual change also holds true for the number of listeners actively listening to Pandora each week. By the end of the experiment, the highest treatment group has 2% fewer active listeners than control, while the lowest treatment group has 1% more listeners than control. Most importantly, we see in both graphs of the weekly treatment effect how important it is that we ran the experiment for over a year. In both graphs, we see that the treatment effect grows over the course of an entire year, stabilizing for the most part only after 12-15 months of treatment. Table 4 shows that the treatment assignment impacted the total hours and active days

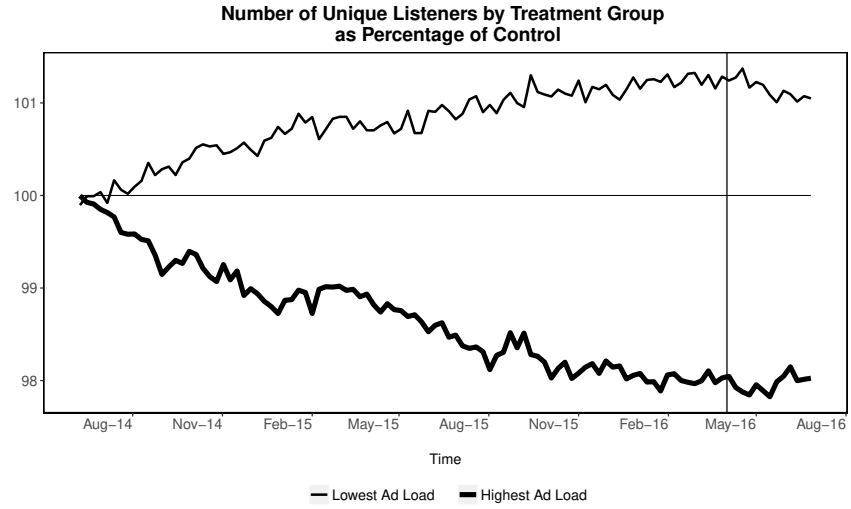

Figure 5: Mean Weekly Unique Listeners by Treatment Group

Table 4: Final Month Summary Statistics

	Control	Lowest Ad Load	Highest Ad Load
Total Hours (norm.)	100.000	101.740	97.169
	(0.083)	(0.243)	(0.225)
Days Active (norm.)	100.000	101.683	97.403
	(0.042)	(0.134)	(0.131)
Sample Size	18,342,916	1,833,826	1,831,909

in a fashion consistent with theory (i.e. users who were exposed to a higher ad load listened for fewer hours relative to the control, while those who were exposed to a lower ad load listened to more hours).

The experimental design allows us to measure the changes in listener behavior due to changes in the number of audio ads they receive. Figure 6 shows the estimated demand curve for Pandora listening in the final month of the experiment, as a function of the number of ads per hour received during the previous 21 months of treatment. Each of the nine treatment groups and the control group are plotted as a single point. We can see that this demand curve is strikingly linear, much like the simplified demand curves plotted in principles-of-economics textbooks. Since none of the points deviates very much from the best-fit line, we infer that the number of pods per hour and the number of ads per pod have only second-order impacts on listening behavior, with ads per hour being the first-order explanatory variable. We will return to this question below with an explicit model.

The best-fit line displayed in Figure 6 is the result of two-stage-least-squares (2SLS) estimation. We use 2SLS instead of ordinary least squares (OLS) because we don't have complete control over treatment, for the variety of reasons discussed above, so that the realized ad load (number of ads per hour, aka "price") differs from what the experiment set as the intended maximum ad load. Here, the first stage regresses the realized ad load on nine treatment dummy variables, which simply estimates the mean realized ad load in each treatment group, as illustrated by the horizontal coordinates in Figure 6. The main equation (the second stage) regresses the outcome variable (hours listened in the final month) on the realized ad load, where the first stage causes us to use the mean ad load for each group (rather than the ad load for each individual) as the regressor. The first stage guarantees that we exploit only the experimental differences between groups to identify the causal effects, removing all within-group variation that might yield spurious correlation - for example, urban listeners might both receive more ads and listen fewer hours, on average, than rural listeners, even though this correlation is not at all causal. In addition to hours listened, we also

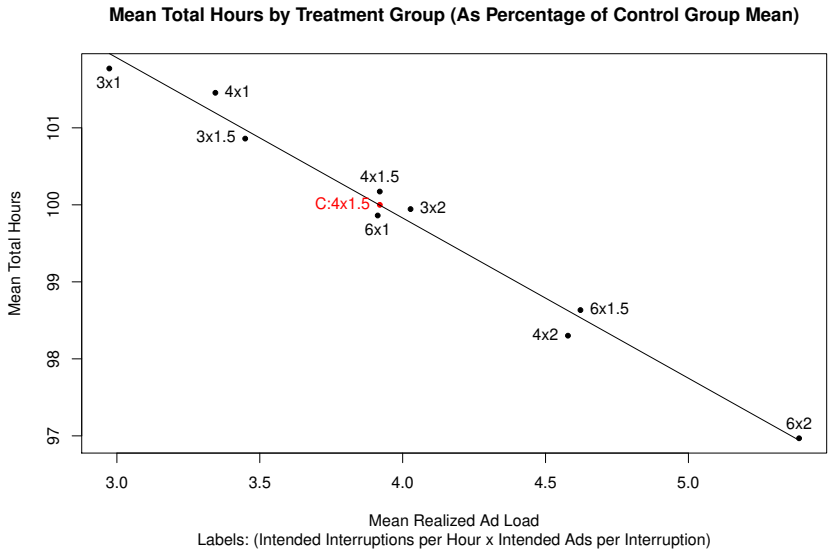

Figure 6: Impact of Ad Load on Final Month Hours per Listener, By Treatment Group

consider the effects of treatment on several other outcome measures from the final month of the experiment.

Model 1 (Simple IV) *Let y_i denote total listening hours or days for listener i in the final month. The regression model we run is given by:*

$$y_i = \beta_0 + \widehat{\frac{ads}{hours}}_i \beta_1 + \varepsilon_i \tag{1}$$
$$\frac{ads}{hours}_i = TG_i'\gamma + \eta_i$$

Table 5: IV, Effect of Number of Audio Ads per Hour on Outcomes

	Total Hours	Active Days
Ad Per Hour	-2.0751^{***}	-1.8965^{***}
	(0.1244)	(0.0663)
(Intercept)	107.4540^{***}	106.8568^{***}
	(0.4601)	(0.2451)
Observations	34,858,249	34,858,249
Note:	*p<0.1; **p<0.05; ***p<0.01	

The 2SLS parameter estimates and standard errors are given in Table 5; we use heteroskedasticity-robust (White) standard errors when reporting all regression results in this paper. As shown in Figure 6, the number of hours listened is a linearly decreasing function of the number of ads per hour during the treatment period.[4] The second column of the Table 5 shows that the number of days listened is similarly decreasing in the ad load (or price of listening). The coefficients show us that one additional ad per hour results in mean listening time decreasing by $2.075\% \pm 0.226\%$, and the number of active listening days decreasing by $1.897\% \pm 0.129\%$. We

[4]The linearity was quite clear in Figure 6, but we also checked the specification by adding a quadratic ad-load term or a logarithmic ad-load term to the linear term in the regression. Unsurprisingly, both specifications produced small and statistically insignificant coefficients on the nonlinear terms.

Table 6: IV, Effect of Audio Ads per Hour on Overall Listening and Listening Intensity in the Final Month

	Hours	Hours/Active Day	Days/Active Listener	Active
Ad Per Hour	-2.0751^{***}	-0.4073^{***}	-0.9420^{***}	-0.9150^{***}
	(0.1244)	(0.0639)	(0.0474)	(0.0442)
(Intercept)	107.4540^{***}	101.4002^{***}	103.2733^{***}	103.3241^{***}
	(0.4601)	(0.2288)	(0.1696)	(0.1634)
Observations	34,858,249	14,312,482	14,312,482	34,858,249
Note:				*p<0.1; **p<0.05; ***p<0.01

note that the large size of our experiment achieves considerable precision: the width of each 95% confidence interval is less than one-fifth the size of the point estimate.

How much does it matter that we conducted a long-run rather than a short-run experiment? To see how our estimates change with longer exposure to the treatment, we run a 2SLS regression for each month of the experiment as if that month were the final one. This time, instead of just a slope coefficient, we present the elasticity of demand with respect to changes in ad load. Figure 7 shows the estimated elasticity across time, with the solid line tracing the point estimates and the dotted lines tracing pointwise 95% confidence intervals. For each month, we calculate the elasticity as the slope divided by the mean ad load of the control group from the beginning of the experiment to the end of that month. The estimated effects of a 1% increase in ad load, on hours and days active, respectively, start out at around -0.02% and -0.025%, slowly increasing to effects of -0.070% and -0.076%. Had we run an experiment for just a month or two, we could have underestimated the true long-run effects by a factor of 3.

Does this decrease in total listening come from shorter sessions of listening, or from a lower probability of listening at all? To answer this question, Table 6 breaks the decrease in total hours down into three components: the number of hours listened per active day, the number of active days listened per active listener, and the probability of being an active listener at all in the final month of the experiment. We have normalized each of these three variables so that the control group mean equals 100, so each of these treatment effects can be interpreted as a percentage difference from control. We find the percentage decrease in hours per active day to be approximately 0.41%, the percentage decrease in days per active listener to be 0.94%, and the percentage decrease in the probability of being an active listener in the final month to be 0.92%. These three numbers sum to 2.27%, which is approximately equal to the 2.08% percentage decline we already calculated for total hours listened.[5] This tells us that approximately 18% of the decline in the hours in the final month is due to a decline in the hours per active day, 41% is due to a decline in the days per active listener, and 41% is due to a decline in the number of listeners active at all on Pandora in the final month. We find it interesting that all three of these margins see statistically significant reductions, though the vast majority of the effect involves fewer listening sessions rather than a reduction in the number of hours per session.

3.1 Causality from Experimental Versus Observational Data

How valuable is the variation generated by this experiment? Since it can be difficult to convince decision makers to run experiments on key economic decisions, and it consumes engineering resources to implement such an experiment properly, could we have done just as well by using observational data? To investigate

[5]These three outcomes - hours per active day, days per active listener, and the probability of a listener being active - can be multiplied together to produce the total hours listened. Therefore, the sum of their percentage changes should equal the percentage change for the total hours. This arises from the fact that, for small changes, the percent change in a product is approximately the sum of the percentage changes:

$$\frac{(X+\partial X)(Y+\partial Y)-XY}{XY} = \frac{\partial X}{X} + \frac{\partial Y}{Y} + \frac{\partial XY}{XY} \approx \frac{\partial X}{X} + \frac{\partial Y}{Y}$$

The small disagreement between our two reported numbers likely results from the fact that the changes in ad load are not small ones.

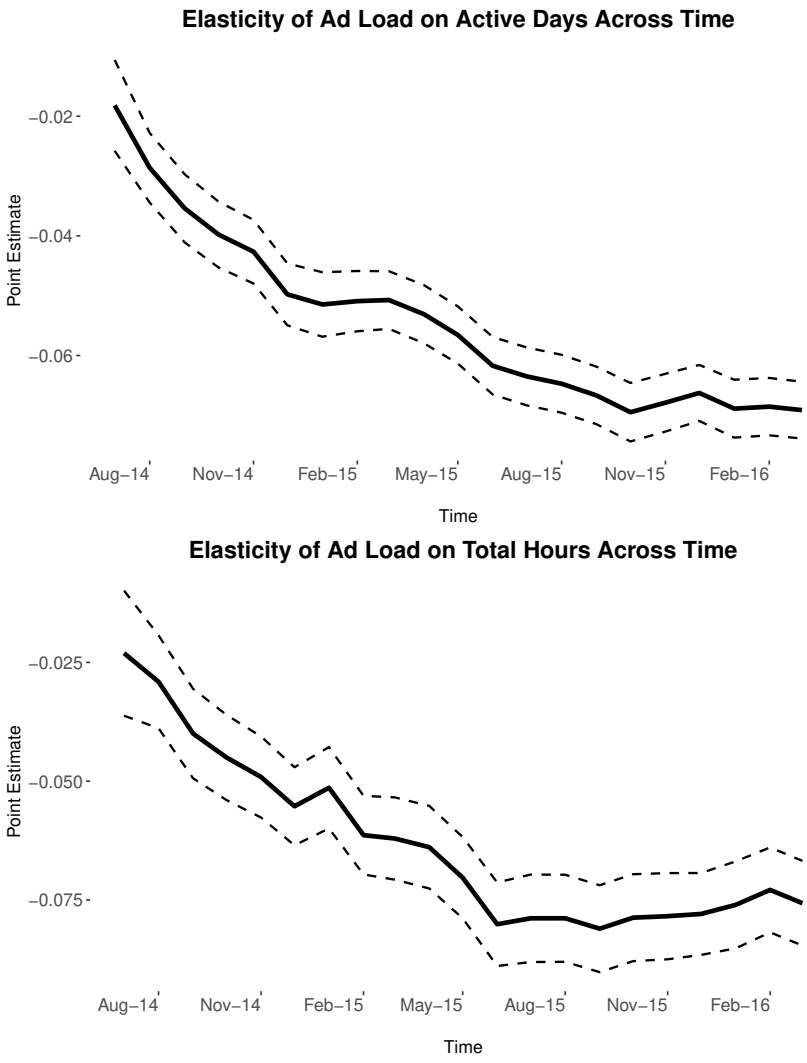

Figure 7: Impact of Ad Load Across Time

this question, we re-run our analysis using only the 17 million listeners in the control group, since they were untouched by the experiment in the sense that they received the default ad load. In the absence of experimental instrumental variables, we run the regression based on naturally occurring variation in ad load, such as that caused by higher advertiser demand for some listeners than others, excluding listeners who got no ads during the experimental period.[6] The results are found in Table 7. We find that the endogeneity of the realized ad load (some people get more ad load due to advertiser demand, and these people happen to listen less than people with lower advertiser demand) causes us to overestimate the true causal impact of ad load by a factor of approximately four. The coefficients for the impact on the number of listening hours and the number of active days both show this endogeneity bias, if we compare Table 7 to Table 5.

Table 7: Endogenous Regression: Effect on Ad Load

	Total Hours	Active Days
Ad Per Hour	-9.800^{***}	-7.224^{***}
	(0.042)	(0.022)
(Intercept)	143.990^{***}	134.373^{***}
	(0.184)	(0.094)
Observations	17,147,544	17,147,544
Note:	*p<0.1; **p<0.05; ***p<0.01	

One obvious criticism of a cross-sectional regression is that it suffers from omitted variable bias: unobservable listener-level heterogeneity may produce non-causal correlation between the ad load and the dependent variable. A valuable technique to address this issue when performing causal inference with observational data is to use a panel regression with listener and time fixed effects. Unfortunately, without a long-run experiment we would not have known the relevant time period to use in our panel estimation; we might have chosen to look at variation from a single month, which clearly would not have produced the correct long-run causal effects. To give the panel estimator its best shot, we use what we have learned from the experiment, and allow the panel estimator a 20-month time period between observations. We run the following model:

Model 2 (Panel Regression) *Let y_{it} denote the outcome of interest for listener i in month t, where t is either the month prior to the start of the experiment or the final month. The regression model we run is given by:*

$$y_{it} = \delta_i + \tau_t + \frac{ads}{hours}_{it} \beta_1 + \varepsilon_i, \tag{2}$$

where δ_i and τ_t are listener and month fixed effects, respectively. For the final month, $\frac{ads}{hours}_{it}$ is calculated based on the experimentation period while for the month prior to the experiment, that ratio is calculated based on May of 2014. To ensure a balanced panel, we restrict the sample to listeners active in May of 2014.

We see from Table 8 that the point estimate for active days is much closer to that found in Table 5, but it still overestimates the impact of ad load by more than the width of our 95% confidence intervals. The panel point estimate for total hours, while an improvement over the cross-sectional regression results, still overestimates the effect by a factor of 3. Our result suggests that, even after controlling for time-invariant listener heterogeneity, observational techniques still suffer from omitted-variable bias caused by unobservable terms that vary across individuals *and time* that correlate with ad load and listening behaviors. And without a long-run experiment, we would not have known the relevant timescale to consider to measure the long-run sensitivity to advertising (which is what matters for the platform's policy decisions).

[6] In the IV specifications, we include listeners with zero ad load. This is because whether a listener gets no ad may depend on the assigned treatment group, and excluding those listeners might cause selection bias. However, we also ran the regressions excluding listeners with zero ads, and we found the results to be very similar.

Table 8: Panel Regression

	Total Hours	Active Days
Ad Per Hour	-6.6263^{***}	-2.3571^{***}
	(0.0613)	(0.0303)
Observations	13,928,268	13,928,268
Note:	*p<0.1; **p<0.05; ***p<0.01	

4 Increased Demand for a Substitute: Ad-Free Subscriptions

Table 9: IV, Effect of Ad Load per Hour on Listener Churn and Pandora One Subscription

	Subscriber At End	Listener Churn
Subscriber At Start	0.3487^{***}	-0.0063^{***}
	(0.0002)	(0.0008)
Ad Per Hour	0.0014^{***}	0.0034^{***}
	(0.00004)	(0.0002)
(Intercept)	0.0046^{***}	0.5172^{***}
	(0.0002)	(0.0007)
Observations	34,858,249	34,858,249
Note:	*p<0.1; **p<0.05; ***p<0.01	

Another important effect of an increase in ad load is to cause some listeners to switch to the ad-free version of Pandora, which was called Pandora One and priced at $4.99 per month during the period of the experiment. We measure this effect in Table 9, where the binary outcome is the listener's subscription status on the last day of the experiment. This 2SLS regression measures how this outcome depends on ads per hour, with the additional covariate of subscription status at the beginning of the experiment. Recall that we study only the set of individuals who did at least some ad-supported listening during the experiment, so any listeners in this dataset who were subscribers at the beginning of the experiment must have allowed their subscriptions to lapse at some point during the experiment. Even for this set of individuals, we see that those who had paid subscriptions on the first day of the experiment were 35 percentage points more likely to be a subscriber at the end than those who didn't subscribe at the beginning of the experiment. For each additional one ad per hour during the experiment, we see a 0.14 percentage-point increase in the probability of being a paid subscriber at the end of the experiment. Multiplying by the subscription fee of $5 per month, we see that Pandora picks up additional monthly subscription revenue of approximately 0.75 cents per listener for each increase of one ad per hour (minus payment-processing costs). This turns out to be considerably smaller than the effects on advertising revenue implied by the demand-curve estimates above.

It is also instructive to compare the magnitude of this result to the similar coefficient for the probability of listening at all in the final month of the experiment, in the last column of Table 6. There we see an estimated decrease of 0.34%, for each one ad per hour, in the probability of listening at all in the final month, which is nearly three times as large as the estimate for the ad-free subscription probability. In other words, for each listener converted to a subscription by increased ad load, three more listeners leave Pandora entirely.

5 Heterogeneous Treatment Effects By Age and Pre-treatment Listening Behavior

Next we look at heterogeneous treatment effects. Pandora most commonly uses four different age categories (13-17, 18-24, 25-54, 55+) when analyzing listeners, corresponding largely to different demographic segments of interest to advertisers. We look for differences in treatment effects across age groups by allowing the coefficients in each regression to vary by age.[7]

Model 3 (Demographic IV) *Let y_i denote total listening hours or days in the final month. Let $Demo_i$ denote a vector of dummy variables corresponding to the demographic groups a given individual belongs to. The Demographic-specific regression model we run is given by:*

$$y_i = \boldsymbol{Demo}_i^T \boldsymbol{\beta}_0 + \widehat{\frac{ads}{hours}}_i \cdot \boldsymbol{Demo}_i^T \boldsymbol{\beta}_1 + \varepsilon_i \tag{3}$$

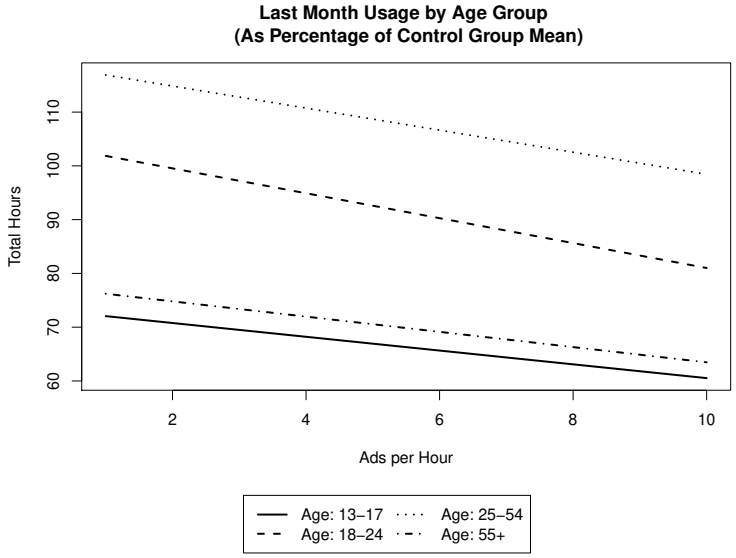

Figure 8: Impact of Ad Load on Final Month Hours per Listener, by Age Groups

Figures 8 and 9 summarize the heterogeneous treatment effects of increasing the ad load on the number of hours listened and the number of days listened in the final month. We see that the middle age groups listen more baseline hours than the oldest and youngest age groups, but the slope (the reduction in number of listening hours due to increased ad load) is quite similar across age groups.

More interesting results come from analyzing the heterogeneous treatment effects on subscription behavior, which we can see in Table 10. The first column shows that both the baseline probability and the marginal change in probability of holding an ad-free subscription at the end of the experiment are monotonically increasing in age. As before, we include subscription status at the beginning of the experiment as a covariate. We see that listeners over 55 years old are twice as likely as listeners between the ages of 13 and 24 (marginal impact of 0.21% versus 0.09% or less) to react to an increase in ad load by paying for the ad-free service. To the extent that older people have more disposable income, these results are consistent with a positive income elasticity of demand for the ad-free subscription.

[7] We actually allowed the coefficients to vary both by age and gender, but found that gender differences were negligible. We also allowed treatment effects to vary by city of residence, looking at the twenty largest DMAs, but found no statistically significant differences, as within-city variation appears to be large compared to between-city variation. We have chosen to report only the most interesting heterogeneous treatment effects we found, which are those by age

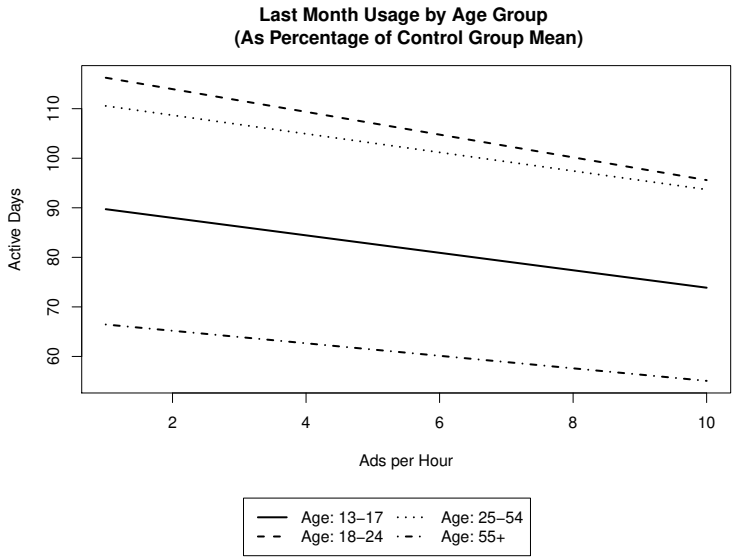

Figure 9: Impact of Ad Load on Final Month Active Days per Listener, by Age Groups

Table 10: IV Effect of Ad Load per Hour on Listener Churn and Pandora One Subscription by Demographics

	Subscriber At End	Listener Churn
Age: 13-17	0.0009	0.6082***
	(0.0007)	(0.0033)
Age: 18-24	0.0018***	0.5407***
	(0.0003)	(0.0014)
Age: 25-54	0.0058***	0.4841***
	(0.0002)	(0.0009)
Age: 55+	0.0095***	0.5453***
	(0.0007)	(0.0031)
Subscriber At Start	0.3472***	0.0106***
	(0.0002)	(0.0009)
Ad Per Hour Age: 13-17	0.0005	0.0017
	(0.0004)	(0.0019)
Ad Per Hour Age: 18-24	0.0009***	0.0032***
	(0.0001)	(0.0004)
Ad Per Hour Age: 25-54	0.0016***	0.0036***
	(0.00004)	(0.0002)
Ad Per Hour Age: 55+	0.0021***	0.0025**
	(0.0003)	(0.0012)
Observations	34,858,249	34,858,249
Note:		*p<0.1; **p<0.05; ***p<0.01

For comparison, in the second column of Table 10 we provide a related set of heterogeneous treatment effects, this time for the probability of leaving Pandora altogether. Pandora generally considers itself to have lost a listener (or the listener has "churned" away from the service) if that person has not listened to the service for two months. We therefore define a binary outcome variable equal to 1 if the listener did not listen at all during the last two months of the experiment. The treatment effects display less heterogeneity than in the subscription results. Comparing the two columns of Table 10 shows us that listeners under 25 years old are much more likely to react to increased ad load by leaving Pandora than by paying for an ad-free subscription (0.32% versus 0.09% for each additional ad per hour). By contrast, listeners over 55 years old are nearly as likely to react by switching to an ad-free subscription as by leaving Pandora altogether (0.21% versus 0.25% for each additional ad per hour).

6 The Effect of the Number of Commercial Interruptions

How do consumers react to the number of commercial interruptions, as opposed to merely the number of ads? Recall that the experimental design separately varied two different components of the ad load: the number of pods per hour and the number of ads per pod. Earlier, we noted that the main impact on listening comes through their product, the total number of ads per hour. To see how the number of commercial interruptions matters independently, we add pods per hour as a covariate in our main outcome regressions, with results shown in Table 11.[8]

Model 4 (Ads per Pod vs. Pods per Hour Model) *Let y_i denote total listening hours and days in the final month for listener i. The regression model we run is given by:*

$$
\begin{aligned}
y_i &= \beta_0 + \widehat{\frac{ads}{hours}}_i \beta_1 + \widehat{\frac{pods}{hours}}_i \beta_2 + \varepsilon_i \qquad (4)\\
\frac{ads}{hours}_i &= TG_i'\gamma_1 + \eta_{1,i}\\
\frac{pods}{hours}_i &= TG_i'\gamma_2 + \eta_{2,i}
\end{aligned}
$$

We find weak evidence to suggest that redistributing advertising across more interruptions, conditional on the number of ads per hour being held constant, may have a small impact on customers' propensity to listen to Pandora. The regression for total hours shows a small and statistically insignificant impact of pods per hour, with a coefficient of 0.204 ± 0.595, ten times smaller than the coefficient on ads per hour. The regression for active days shows a statistically significant impact of pods per hour, with coefficient 0.406 ± 0.332, five times smaller than the coefficient on ads per hour. The positive coefficients on pods per hour suggests that, holding fixed the overall ads per hour, listeners like the ads to be spread out across more pods and have fewer ads per pod. The main determinant of the number of days or hours of listening is the number of ads per hour. For the range we tested (3 to 6 pods per hour), the effect of the number of pods is relatively small once total ads per hour are taken into account.

7 Conclusion

We have reported on the results of an extensive experiment undertaken by Pandora to trace out its consumers' demand curve for Pandora listening as a function of the number of audio ads served. With more than 30 million listeners experiencing one of nine different randomized treatments consistently for a period of 21

[8]The data source for ad impressions differs from that for commercial interruptions. The ad-impression table logs all audio ads actually played to listeners, but it does not record groupings of ads into pods. We obtain pod data from an event table that includes all the ads that Pandora requests from Doubleclick for Publishers (DFP). Pandora pre-fetches ads to improve the listener experience, and some of these pre-fetched ads might never actually be played if the listener terminates her session early, but our source of pod data includes all such requests. To the best of our knowledge, the event table we use as a source of pod data includes approximately 20% unplayed, pre-fetched ad pods. We know we have some measurement error in our pod data, and this could lead to some attenuation bias, which is relevant since we find relatively small causal effects of pod quantity. We aim to fix the measurement problem in future research with better data logging.

Table 11: IV, Effect of the Number of Commercial Interruptions, Controlling for the Number of Ads Per Hour

	Total Hours	Active Days
Ads per Hour	-2.1509^{***}	-2.0472^{***}
	(0.1716)	(0.0914)
Pod per Hour	0.2043	0.4062^{**}
	(0.3185)	(0.1697)
(Intercept)	107.1143^{***}	106.1815^{***}
	(0.7015)	(0.3738)
Observations	34,858,249	34,858,249
Note:	*p<0.1; **p<0.05; ***p<0.01	

months, we have obtained what may be the best estimate of a demand curve ever performed. We found that the quantity of hours listened is strikingly linear in the price (number of ads per hour). By contrast, the number of commercial interruptions per hour (which we varied separately from the number of ads per interruption) has little effect on listening behavior once we have controlled for the number of ads per hour. We decomposed this clear reduction in listening due to increased ad load into three components: 41% due to Pandora listeners stopping their listening entirely, 41% due to active Pandora listeners listening fewer days during the month, and 18% due to listeners listening fewer hours on each day they choose to listen.

Our experiment measured the impact of the experienced quantity of advertising on listening behavior. While this gives us a very relevant demand curve, it is different from the exercise of announcing a new "price" (or advertising quantity) publicly and having it be consistent across consumers. To some people, the latter might seem a better example of what one usually means by a price change. However, giving everyone the same treatment would destroy our ability to measure the effects with a controlled experiment. We demonstrate the importance of using an experiment rather than relying on endogenous observational data, as a regression relying only on naturally occurring variation in ad load within the control group produces an estimate three times as large as the truth we obtain with our experiment.[9]

In addition to measuring the downward-sloping demand curve for Pandora ad-supported listening, we also measured the related increase in demand for a substitute: Pandora ad-free subscriptions. The impact on incremental ad-free subscriptions was highly significant, and monotonically increasing in the age of the consumer. The incremental probability of leaving Pandora entirely due to increased ad load is monotonically decreasing in age. These two effects combine so that for the oldest consumers, the incremental probability of switching to an ad-free paid subscription is almost as large as the incremental probability of leaving Pandora altogether.

Our estimates are quite precise, with the slope coefficients in our main outcome regressions having t-statistics of approximately 20 (i.e., point estimates five times larger than the width of their 95% confidence intervals). We also saw that the effects of a change in ad load take at least a year to be fully realized, demonstrating the importance of our having run a long-term experiment. These results go a long way towards helping us understand the science of two-sided markets, as we have managed to describe the behavior of one side of Pandora's advertising business in great detail. To more fully understand the demand for the platform from the advertiser's perspective, we would need to conduct a similar pricing experiment with advertisers. We have begun to contemplate how to run such an experiment, but the experimental design will be much harder, given that the number of advertisers is several orders of magnitude smaller than the number of listeners, and that advertising sales generally take place via bilateral negotiations.

The results are also quite relevant to Pandora's business, as the firm has gained a much better understanding of its listeners through this experiment. Prior to the experiment, Pandora had almost no data

[9]Note that one potential source of bias in our experiment is word-of-mouth effects. If some listeners stop listening to Pandora because they talk to friends who complain about how many ads they hear on Pandora, then this would cause spillovers between treatment and control, and thereby cause us to underestimate the true causal effects of ad load. We believe such a spillover effect to be negligible on the existing Pandora listeners we study, but we acknowledge the possibility that spillover effects could cause us to estimate consumers to be less sensitive to ad load than they really are.

on how sensitive its ad-supported listeners were to the level of advertising. Armed with the demand-curve estimates presented here, the company has much better information for deciding appropriate levels of advertising for its audience. The results on heterogeneous treatment effects by age are also quite valuable for understanding listeners' decisions to pay for ad-free subscriptions. Though such experiments have not been common in the past, we hope that firms will conduct more such important experiments in future.

References

Ashraf, N., Berry, J., & Shapiro, J. M. (2010). Can higher prices stimulate product use? evidence from a field experiment in zambia. *The American Economic Review*, 100(5), 2383–2413.

Berry, J., Fischer, G., & Guiteras, R. P. (2015). Eliciting and utilizing willingness to pay: evidence from field trials in northern ghana. *CEPR Discussion Paper No. DP10703*.

Cohen, J. & Dupas, P. (2007). Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment. *Brookings Global Economy and Development Working Paper*, (11).

Cohen, P., Hahn, R., Hall, J., Levitt, S., & Metcalfe, R. (2016). *Using Big Data to Estimate Consumer Surplus: The Case of Uber*. Working Paper 22627, National Bureau of Economic Research.

Gneezy, A., Gneezy, U., Riener, G., & Nelson, L. D. (2012). Pay-what-you-want, identity, and self-signaling in markets. *Proceedings of the National Academy of Sciences of the United States of America*, 109(19), 7236–7240.

Hohnhold, H., O'Brien, D., & Tang, D. (2015). Focus on the long-term: It's better for users and business. In *Proceedings 21st Conference on Knowledge Discovery and Data Mining* Sydney, Australia.

Lusk, J., Jaeger, S., MacFie, H., et al. (2010). Experimental auction markets for studying consumer preferences. *Consumer-Driven Innovation in Food and Personal Care Products*, (195), 332–357.

Miguel, E., Kremer, M., Beasley, E., Benaya, L., Brooker, S., Dupas, P., Luoba, A., Moulin, S., Namunyu, R., Waswa, P., Wafula, P., Akerlof, G., Imbens, G., Koszegi, B., Laibson, D., Munshi, K., & Rosenzweig, M. (2002). Why don't people take their medicine? experimental evidence from kenya. *mimeo, University of California, Berkeley and Harvard University*.

PSBJ (2000). Bezos calls amazon experiment 'a mistake'. Accessed: 2016-08-12.

Rysman, M. (2009). The economics of two-sided markets. *The Journal of Economic Perspectives*, 23(3), 125–143.