Measuring Strategization in Recommendation: Users Adapt Their Behavior to Shape Future Content

SARAH H. CEN*, Massachusetts Institute of Technology, USA
ANDREW ILYAS*, Massachusetts Institute of Technology, USA
JENNIFER ALLEN, Massachusetts Institute of Technology, USA
HANNAH LI, Columbia University, USA
DAVID G. RAND, Massachusetts Institute of Technology, USA
ALEKSANDER MADRY, Massachusetts Institute of Technology, USA

Modern recommendation algorithms are data-driven: they generate personalized recommendations by observing users' past behaviors. A common assumption in recommendation is that how a user interacts with a piece of content (e.g., whether they choose to "like" it) is a reflection of the content, but *not* of the algorithm that generated it. Although this assumption is convenient, it fails to capture user strategization: that users may attempt to shape their future recommendations by adapting their behavior to their recommendation algorithm. In this work, we test for user strategization by conducting a lab experiment and survey. We begin with a model of user strategization that captures how strategic users select their current actions to improve their downstream recommendations. We use this model to formulate two testable hypotheses. Using a music platform that we built, we study how users respond to different *information* about their recommendation algorithm as well as different *incentives* about how their current actions affect downstream outcomes. We find strong evidence of strategization in both dwell time and engagement metrics. For example, participants who are told the algorithm will generate personalized recommendations primarily based on their "likes" and "dislikes" used "likes" and "dislikes" 1.9 times more than participants who are told the algorithm learns primarily from their dwell time. In the post-experiment survey, 60 percent of participants self-reported strategizing. We also document how and why users strategize "in the wild." Ultimately, our findings indicate that user strategization in recommendation is common, suggesting that platforms cannot assume away the effect of the recommendation algorithm on user behavior.

*Both authors contributed equally to the paper

Authors' addresses: Sarah H. Cen, Massachusetts Institute of Technology, Cambridge, USA, shcen@mit.edu; Andrew Ilyas, Massachusetts Institute of Technology, Cambridge, USA, ailyas@mit.edu; Jennifer Allen, Massachusetts Institute of Technology, Cambridge, USA, jnallen@mit.edu; Hannah Li, Columbia University, New York, USA, hannah.li@columbia.edu; David G. Rand, Massachusetts Institute of Technology, Cambridge, USA, drand@mit.edu; Aleksander Madry, Massachusetts Institute of Technology, Cambridge, USA, madry@mit.edu.

1 INTRODUCTION

Recommendation platforms—like TikTok, Netflix, and Amazon—attract and retain users by tailoring content (e.g., videos, shows, and products) to each user's interests. Although platforms employ a wide variety of algorithms, all of them are trained on *past user behavior*. For instance, Netflix generates recommendations based on each user's watch and rating history.

These data-driven algorithms typically assume that user behavior is *exogenous*: how a user reacts to a recommendation depends on that recommendation alone, and *not* on the algorithm that generates it [Adomavicius and Tuzhilin, 2005, Ricci et al., 2010]. This assumption implies, for example, that a user will "like" a video with the same probability irrespective of the recommendation algorithm that produces it. In other words, a user's revealed preferences (implied by their engagement behavior) remain consistent across recommendation algorithms as long as their true preferences (their unknown utility function) remain the same.

What this exogeneity assumption fails to capture is *strategic* behavior: that users may attempt to shape their future recommendations by adapting their revealed preferences to their recommendation algorithm, even if their true preferences do not change. For example, a TikTok user might "heart" a video not because they enjoy it, but because they like the creator and believe TikTok's algorithm will recommend more content from creators they "heart" in the future. Or a Spotify user might choose to ignore a "guilty pleasure" song that they actually like because they are worried Spotify's algorithm will recommend too many similar songs later on. In these example, the true, unknown utility that the user receives from each recommendation does not change across algorithms, but the user's behavior may. Aware that their actions serve as training data for future recommendations, the user may adjust their actions to improve their downstream outcomes.

User strategization would have important implications for recommendation algorithm design. Since recommendation algorithms are continually trained on user data, strategization can lead to unintended effects (such as feedback loops [Perdomo et al., 2020]). User data is also used for a variety of other purposes (e.g., to estimate off-platform behavior or to synthetically test new algorithms), and strategization would hurt a platform's ability to perform these tasks, as the data that a platform gathers would become *algorithm-dependent* [Cen et al., 2023].

Although strategization would have significant impacts on the data that platforms gather, to our knowledge, there has not been a lab experiment investigating whether strategization in recommendation does indeed occur. The goal of the current work is to fill this gap. To do so, we conduct a survey and a lab experiment that uncover user strategization and insights into why users strategize.

1.1 Our Contributions

Definition of user strategization. We begin with a formal definition of strategization in Section 2, which we adapt from [Cen et al., 2023]. Formally, each user is characterized by a utility function U, where $U(Z, B)$ denotes the payoff that the user internalizes if they take action B in response to recommendation Z (e.g., click on the recommendation). One can think of U as capturing the user's *true, unknown preferences*. Typically, it is assumed that users behave naively, e.g., play an action $B^{\text{naive}}(Z) \in \arg\max U(Z, B)$ that maximizes their payoff under recommendation Z. This assumption is convenient because it implies that a user's *revealed* preference is a function of the recommendation alone. On the other hand, a *strategic* user is aware that their current actions are used to generate future recommendations under some data-driven algorithm $\mathcal{A}$. A strategic user therefore anticipates how possible current actions affect future recommendations under $\mathcal{A}$ and chooses an action $B^{\text{strat}}(Z, \pi)$ that maximizes their long-term payoff, as formalized in Section 2.[1] In

[1]There is a distributional version of naive and strategic behavior such that a user's actions are not deterministic. The same reasoning applies under distributional actions, so we use the deterministic setting for ease of exposition.

other words, a strategic user's revealed preferences would be *algorithm-dependent*, which would complicate the platform's ability to estimate U.

Testable hypotheses for strategization. Testing for user strategization is challenging. For one, each user's true preferences are unknown, making it difficult to determine whether their revealed preferences (i.e., what is observable) matches their true preferences, which is at the core of determining whether users are strategic. For another, users have heterogeneous preferences (i.e., there is a different, hidden U for each user).

Despite these challenges, the definition of strategization in Section 2 suggests that there are two hypotheses that we can use to test for strategization.

Hypothesis 1 (Information Hypothesis, informal). *Different descriptions of how a participants's preferences will be learned prompt participants to behave differently.*

Hypothesis 2 (Incentive Hypothesis, informal). *Participants who are and are not told they will receive recommendations behave differently.*

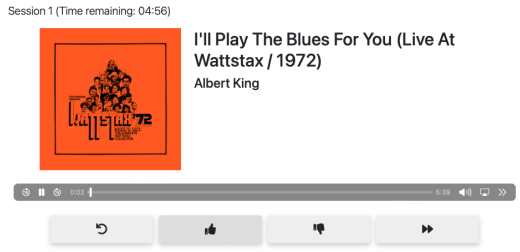
The first hypothesis implies that users are not only aware of their algorithm, but also adapt their behavior based on their understanding of the algorithm. The second hypothesis implies that users adapt in a way that is aware of the *data-driven* nature of algorithms, i.e., that their current actions influence downstream recommendations. Together, these hypotheses would indicate that users are strategic in that they *adapt* their current behavior in order to elicit good *future* payoffs.

Behavioral experiment on custom music streaming platform. We run a lab experiment with 750 participants. We build a basic music streaming platform that allows us to observe how participants interact with their songs (e.g., which songs participants "like" and how long they listen to each song), as shown in Figure 1a. Half of the participants are told that they will receive recommendations at the end of the study (see Figure 1b), and the other half are told that their behavior is used to learn people's music preferences, but not that they will receive recommendations (see Figure 1c). These "Incentive" conditions mirror those given in Hypothesis 2.

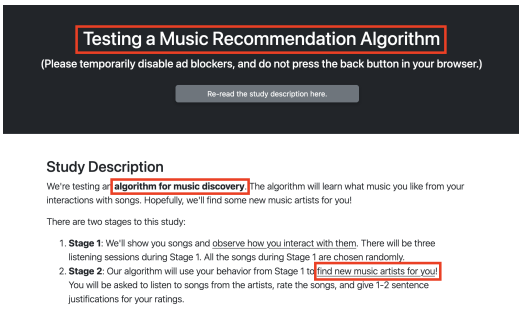
Each participant undergoes two listening sessions. During the first session (the warm-up), participants are asked to behave as they would on their typical music recommendation platform (e.g., Spotify). During the second, participants are randomly exposed to one of three descriptions of the platform's recommendation algorithm before interacting with the songs, as per Hypothesis 1 (see Figure 2). Our goal is to test for user strategizaton by study whether and how (i) Information and (ii) Incentives influence user behavior.

Evidence of user strategization. Our findings show strong evidence of user strategization. There are marked changes in user behavior not only across different Information conditions, but also across the two Incentive conditions, confirming both Hypothesis 1 and Hypothesis 2. These effects are not just concentrated among highly active participants, but are observed across our outcome distributions, suggesting strategization is not a rare behavior. Furthermore, while we find that the nature and degree of strategization differ by individual characteristics, we observe consistent evidence of strategization even among participants one might expect to be naive (e.g., older participants). We further survey participants at the end of the study to understand whether users strategize "in the wild" and whether they do so intentionally. Of those who received the Incentive treatment, 60 percent reported strategizing in our experiment. Moreover, many report definitive strategization "in the wild" (e.g., some say they do not like being "pigeonholed" by the algorithm and maintain multiple user accounts for different "moods").

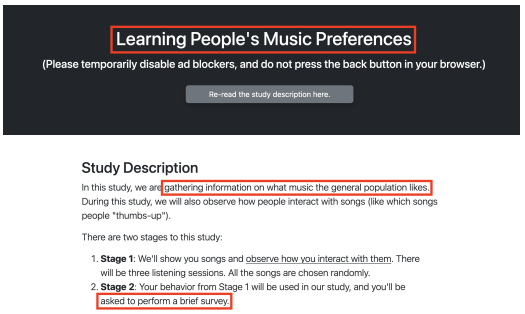
These results provide a first step in documenting user strategization in recommendation. Although we study the recommendation setting in this work, our findings suggest that people are



(a) Our interface



(b) "Treatment" Incentive study description



(c) "Control" Incentive study condition

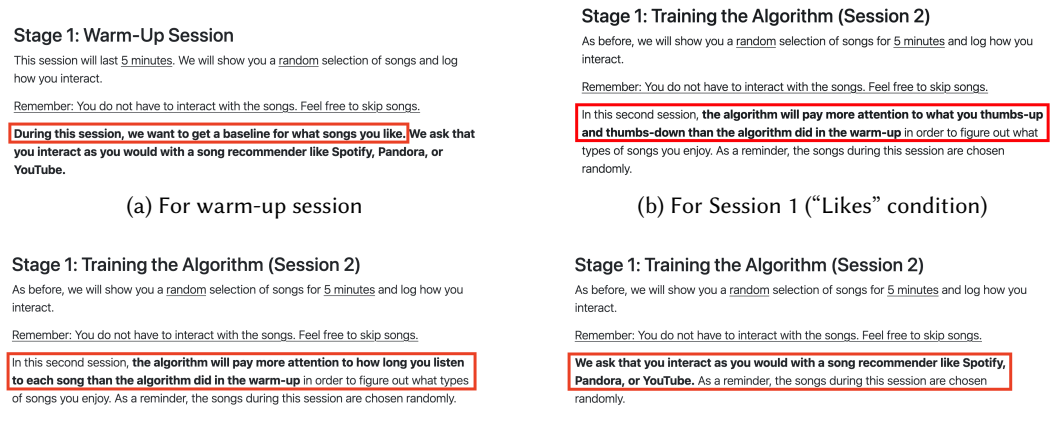
Fig. 1. (a) The music player interface with which participants interact. (b) The study description that participants in the "Treatment" Incentive condition see. These participants are told that their behaviors will be used to generate personalized music recommendations at the end of the study. (c) The study description that participants in the "Control" Incentive condition see. These participants are told that their behaviors are used to learn what music the general population likes. The participants are randomly divided into the "Treatment"and "Control" Incentive conditions.

increasingly aware of data-driven algorithms and have begun to develop strategies to improve their outcomes. Ultimately, this behavior can hurt platforms, which rely on users' *revealed* preferences to make predictions, train new algorithms, and even draw broader conclusions about their users. Strategic behavior can therefore hurt platforms because their revealed preferences do not necessarily reflect how they would behave under a different algorithm (or, more broadly, different circumstances). Thus, unless platforms take user strategization into account, they can be misled, to the detriment of both themselves and their users.

The rest of the paper is organized as follows. Section 1.2 discusses the related work. Section 2 presents a formal definition for strategization and, in accordance, the hypotheses that we wish to test. Section 3 describes the methodology and analysis we employ to test these hypotheses. Section 4 presents our results and evidence of strategization. Finally, we discuss the implications of our findings and future directions in Section 5.

1.2 Related Work

User awareness of recommendation algorithms. Existing work shows evidence that users are *aware* of recommendation algorithms and have beliefs about how they work [DeVito, 2021, DeVito et al., 2018, Eslami et al., 2016, Newman et al., 2018, Sirlin et al., 2021, Taylor and Choi,



(a) For warm-up session

(b) For Session 1 ("Likes" condition)

(c) For Session 1 ("Dwell" condition)

(d) For Session 1 ("Control" condition)

Fig. 2. Participants undergo two listening sessions. The first session for all participants is the warm-up session, as shown in (a). In the second session, participants are randomly assigned to one of three Information conditions, as shown in (b), (c), and (d). (The descriptions above are shown to participants in the "Treatment" Incentive condition. The "Control" Incentive descriptions are analogous.)

2022]. The existence of these beliefs is a precursor to the phenomenon that we study, that users behave strategically in response to these beliefs.

Qualitative evidence of user strategization on recommendation systems. More recently, there has been growing evidence of users attempting to *influence* what their recommendation systems show them, primarily relying on self-reported survey data [DeVito et al., 2017, Haupt et al., 2023, Lee et al., 2022, Shin, 2020, Simpson et al., 2022] There are tutorials aimed to teach the general population how to "train" their social media algorithms—that is, explicitly changing their behavior to induce a better feed [Narayanan, 2022, 2023, WSJ, 2021]. To the best of our knowledge, our work provides the first large-scale behavioral study that quantifies the existence of strategization by measuring observed behavior change, rather than relying on self-reported data.

Theoretical models of user strategization on recommender systems. Recent theoretical work shows how user strategization on recommender systems may result from "long-term planning" [Cen et al., 2023, Haupt et al., 2023]. Haupt et al. [2023] proposes a model where strategic users can modify their consumption patterns and Cen et al. [2023] proposes a model where strategic users can modify any action, including, but not limited to, consumption. We adapt the model from the latter to develop testable hypotheses in Section 2 for the existence of strategization and analyze consumption patterns in addition to user feedback ("likes" and "dislikes"). Our work provides empirical evidence for "long-term planning" as a mechanism leading to strategization.

Strategization in other contexts. We note that in the context of recommendation systems, there is an existing body of work showing that *content creators* strategize in terms of the type and frequency of the content they create [Arriagada and Ibáñez, 2020, Hron et al., 2023, Huang et al., 2022, Huttenlocher et al., 2023, Immorlica et al., 2024, Jagadeesan et al., 2022, Mummalaneni et al., 2023]. We consider our focus of user-side strategization as a distinct phenomenon from creator-side strategization, as well as documented strategization in online auctions [Edelman and Ostrovsky, 2007] and freelancing [Rahman, 2021] and ride-share platforms [Marshall, 2020]. This user-side strategization distinctly aims to assist the algorithm in learning the user's own preferences.

We further distinguish our work from strategic classification [Brückner and Scheffer, 2009, Hardt et al., 2016] and generalized strategic classification introduced in Levanon and Rosenfeld [2022]. In the original strategic classification formulation, agents strategize to induce a positive decision (e.g., loan approval) whereas agents in the generalized strategic classification model strategize to induce the *correct* decision. Our formulation is closed to the latter in the sense that recommender systems seek to personalize to user preferences. However, our work is distinct in that we examine settings with repeated interactions, where an individual is cognizant of the fact that their current actions are used as training data and may therefore influence future outcomes, rather than a one-off prediction setting.

Learning from revealed preferences. Our work contributes to the observation that users' revealed preferences (observed behavior) may not be indicative of their true interests [Beshears et al., 2008]. In the recommender system space, Kleinberg et al. [2022], Morewedge et al. [2023] suggest that relying revealed preferences can result in suboptimal recommendations, due to factors like users' inconsistent preferences and habitual behavior. Our work shows how revealed preferences depend not only on users' true interests, but also their beliefs about the algorithm and whether they are forward looking.

2 HYPOTHESES

Before describing our methodology, we begin with a model of user strategization. This model characterizes how users strategize relative to the *data-driven nature* of their algorithm by anticipating how actions that they take now affect downstream recommendations. Under this model, there are two components of user strategization in recommendation. First, a user strategizes based on their understanding of the algorithm (as they do in the setting of strategic classification). Second and of particular interest in this work, strategization is *forward-looking*: cognizant of the fact that their recommendation algorithm is data-driven, strategic users select actions that will benefit them in the long run.[2] Using this model, we formulate two testable hypotheses of strategization.

2.1 Model of Strategic Users

We use a model similar to that of Cen et al. [2023], distilled to the components that we test experimentally in this work and adapted to a finite-horizon recommender system. Formally, let $\mathcal{Z}$ be the set of content available on a platform, and let $\mathcal{B}$ be the set of ways (or behaviors) with which a user can respond to a piece of content $Z \in \mathcal{Z}$. The user and platform engage in $T > 0$ repeated interactions where at each time step $t \in \{1, \ldots, T\}$, the platform gives a *recommendation* $Z_t \in \mathcal{Z}$ and the user responds with a *behavior* $B_t \in \mathcal{B}$. Based on both the recommendation Z_t and their behavior B_t, the user collects a reward $U(Z_t, B_t)$.

The user believes that the platform generates its recommendations using an *algorithm* $\mathcal{A}$, which maps the interaction history $\mathcal{H}_t = \{(Z_1, B_1), \ldots, (Z_t, B_t)\}$ at time t to a new recommendation Z_{t+1} at time $t + 1$. (Note that for the purposes of this model, the way the platform *actually* generates recommendations is irrelevant, and in particular, it need not match the user's belief.) Based on the algorithm $\mathcal{A}$ and the system parameters (i.e., $\mathcal{B}, \mathcal{Z}$, and T), the user chooses a *behavior policy* $\pi(\cdot\,; \mathcal{A}, \mathcal{B}, \mathcal{Z}, T)$ that maps a recommendation Z_t to a distribution over behavior. [3] At time t, the user responds to the recommendation Z_t by sampling $B_t \sim \pi(Z_t; \mathcal{A}, \mathcal{B}, \mathcal{Z}, T)$.

[2]Note that this phenomenon is distinct from the inconsistent preferences (e.g., "junk" versus "healthy" content preferences) phenomenon highlighted by Kleinberg et al. [2022]. In this work, we consider strategization with respect to the algorithm, whereas Kleinberg et al. [2022] consider strategization with respect to one's own internal preference inconsistencies.

[3]We consider users that are *stationary* in that they do not use the history H_t. This is without loss of generality, given that we do not restrict $\mathcal{Z}$, which can be adjusted to "contain" the history.

The simplest behavior policy is the *naive* policy, which simply responds to recommendation Z_t with the behavior that maximizes the user's immediate utility, i.e.,

$$\pi^{\text{naive}}(Z_t; \mathcal{A}, \mathcal{B}, \mathcal{Z}, T) = \delta\left\{\arg\max_{B \in \mathcal{B}} U(B, Z_t)\right\}, \tag{1}$$

where $\delta\{\cdot\}$ is the Dirac delta distribution (i.e., a degenerate distribution which places probability one on its argument and probability zero on everything else). In the language of recommender system, the naive policy behaves according to users' "ground-truth" preferences about the content they see. The typical recommender system assumes that the user follows the naive policy and typically refers to it as the user's "ground-truth" behavior. In particular, note that π^{naive} is exogeneous in that does not depend on either the perceived algorithm $\mathcal{A}$ nor the time horizon T.

However, a user may indeed account for $\mathcal{A}$ and T when deciding how to behave. They may choose sub-optimal short-term behavior to receive better treatment by $\mathcal{A}$ in the long term. Such a policy depends on both $\mathcal{A}$ and T. Thus we define the following class of *strategic policies* where the user chooses a strategy that optimizes their utility over T time steps for $T > 1$:

$$\pi^{\text{strat}}(\cdot; \mathcal{A}, \mathcal{B}, \mathcal{Z}, T) = \arg\max_{\pi} \mathbb{E}_{\mathcal{H}_T(\pi, \mathcal{A})}\left[\sum_{t=1}^{T} U(B_t, Z_t)\right], \tag{2}$$

where $\mathcal{H}_T(\pi, \mathcal{A})$ is a function mapping from a user policy and a platform algorithm to the (endogenous) T-step rollout of recommendations and actions generated by the user and platform interacting according to π and $\mathcal{A}$ respectively. For simplicity, we assume that the user chooses their strategic policy once; given knowledge of $\mathcal{A}$ and T, the strategic user computes $\pi^{\text{strat}}(\cdot; \mathcal{A}, \mathcal{B}, \mathcal{Z}, T)$ before $t = 1$. Note that this policy class reduces to the naive policy (1) for $T = 1$.

2.2 Hypotheses

In this study, we present participants with a synthetic recommendation platform for which $\mathcal{B}$ and $\mathcal{Z}$ are universally fixed. For a given participant, let $\mathcal{A}$ denote their belief about the recommendation algorithm and $\pi^*(\cdot; \mathcal{A}, \mathcal{B}, \mathcal{Z}, T)$ denote their chosen behavior policy. We observe samples from π^* by tracking the participant's behavior. To construct our hypotheses, we make two observations:

(1) The naive policy does not depend on (the user's belief of) the platform's recommendation algorithm $\mathcal{A}$, while the strategic policy changes for different beliefs $\mathcal{A}$.
(2) The naive policy is independent of the time horizon T of the platform-user interaction, whereas the strategic policy is not.

Below, we translate these two observations into concrete hypotheses about user behavior, by varying the user's percieved algorithm $\mathcal{A}$ and the time horizon T exogenously, and estimating the effect on $\pi^*(\cdot; \mathcal{A}, \mathcal{B}, \mathcal{Z}, T)$ from samples.

Hypothesis 3 (Information Condition). *Holding all else constant, changing the participants' beliefs about the recommendation algorithm changes the way they behave. Formally, there exist $\mathcal{A}$ and $\mathcal{A}'$ such that $\pi^*(\cdot; \mathcal{A}, \mathcal{B}, \mathcal{Z}, T) \neq \pi^*(\cdot; \mathcal{A}', \mathcal{B}, \mathcal{Z}, T)$.*

Hypothesis 4 (Incentive Condition). *Holding all else constant, changing the time horizon of the platform-user interaction will change the participant's behavior. Formally, there exist time horizons T_1 and T_2 such that $\pi^*(\cdot; \mathcal{A}, \mathcal{B}, \mathcal{Z}, T_1) \neq \pi^*(\cdot; \mathcal{A}, \mathcal{B}, \mathcal{Z}, T_2)$.*

These two hypotheses formalize those presented in Section 1. Our study—a controlled lab experiment—enables us to test the two hypotheses above *directly*, as discussed next.

3 METHODOLOGY

In this section, we describe our experimental methodology and analysis. In short, we build a music recommendation interface that allows participants to listen to and interact with songs, as they similarly would on Spotify or Pandora. We conduct a behavioral experiment in which participants are randomly exposed to different Information and Incentive conditions, as discussed in Section 2. We use data about each participant's behavior (e.g., the number of likes, skips, and replays on the platform) to determine the treatment effects of different Information and Incentive conditions. Specifically, we sought to answer the questions: Do different Information and Incentive conditions affect participant behavior in a systematic way? If so, do the observations support the strategization hypotheses given in Section 2? We additionally asked participants to complete a post-experiment survey to determine whether they intentionally strategize on their chosen recommendation platforms. All analyses are pre-registered, except where they are designated "post-hoc." Our pre-registration and analysis plan is available at https://aspredicted.org/WVF_6SH.

3.1 Participants

We recruited 750 participants from CloudResearch Connect. Of the recruited participants, we exclude 28 participants who ran into technical issues. Of the remaining participants, 50 failed at least two audio-visual attention checks or written attention checks. Finally, another 15 participants had metrics (likes, dislikes, skips, and dwell time) that were more than four standard deviations away from the average. The final sample has 657 participants, of which 48% are male, 52% are female; 45% are 35 years-old and below, and 70% use Spotify or music recommendations platforms at least a few times a week. In accordance with the standards of [anonymized]'s Institutional Review Board (IRB), this study was granted an exemption from full IRB review on March 9, 2024.

3.2 Music Platform

We build a basic music recommendation platform on which participants can listen to and interact with songs. Each participant undergoes two 5-minute listening sessions. During each session, participants can thumbs-up, thumbs-down, skip, and restart each song as well as skip to any time within a song, as shown in Fig. 1a. We log all participants' actions that involve clicks (such as the actions listed above). Each song is chosen randomly from a song bank of 196 songs purchased from iTunes on March 5, 2023 from the "Top Songs" of 16 genres. The songs presented in the listening sessions are chosen uniformly at random (without replacement) for all participants and therefore do *not* depend on the participants' interaction behavior, as described next.

3.3 Experimental Conditions

Our experiment uses a 3-by-2 factorial design with the conditions described below. The first factor, which we refer to as the "Information condition," tests Hypothesis 3. The second factor tests, which we refer to as the "Incentive condition," tests Hypothesis 4. In total, we have six different pairs of conditions that determine each participant's knowledge about (i) the algorithm used to learn their preferences and (ii) whether the learned preferences are used to generate personalized recommendations for the participant. Since participants are first exposed to an Incentive condition, then an Information condition, we describe the conditions in that order.

Study description. At the start of the study, participants are randomly exposed to one of two descriptions of the study's purpose, or *Incentive conditions*:

- **"Control" Incentive** Participants in the Incentive control are told that we are learning what music the general population likes based on their interactions with songs. They are told that they will first undergo listening sessions, then be asked to answer a brief survey.
- **"Treatment" Incentive**: Participants are told that we are testing an algorithm for music discovery that will try to learn their preferences from their interactions with song. They are told that they will first undergo listening sessions, then be asked to give feedback on songs that our algorithm recommends.

The Incentive conditions are intended to determine whether users plan ahead when they know that their current actions affect their future recommendations, as per Hypothesis 4.[4] Note that participants under the "Control" Incentive may plan ahead to some degree if they naturally strategize "in the wild" because planning ahead on recommendation platforms has become habitual for them. If this does occur, it would dampen our observed treatment effect. If, however, we still observe a treatment effect despite this spillover, it provides even stronger evidence in support of Hypothesis 4.

First session. The participants then undergo their first listening session, which we call the "warm-up session." For this session, *all* participants are told to interact as they would with a song recommender like Spotify or Pandora in order for us to get a baseline for the music they like.

Second session. We then randomly split participants into three groups for their second listening session. Before the second session, participants are exposed to one of three *Information conditions*:

- **"Control" Information** As in the warm-up, participants receive no information about how their preferences are learned. They are told to interact as they would with Spotify or Pandora.
- **"Likes" Information:** Participants are told that, in order to learn their music preferences, we pay more attention to how they "like" (thumbs-up) and "dislike" (thumbs-down) songs as compared to the warm-up session.
- **"Dwell" Information:** Participants are told that, in order to learn their music preferences, we pay more attention to their dwell time (how much time they spend on each song) as compared to the warm-up session.

As such, some participants undergo the Information control for both listening sessions, some undergo the Information control *then* the "Likes" condition, and the rest undergo the Information control *then* the "Dwell" condition. Note that the way we generate songs for participants does *not* change across participants (all songs during the listening sessions are generated randomly). We only change the information that participants receive.

3.4 Summary of Lab Experiment

In summary, we conduct our experiment on a custom-built platform. At the start of the experiment, each participant reads and agrees to the study instructions, which depend on the participant's Incentive condition (participants in the "Control" Incentive group are told they are participating in a general-interest survey, and participants in the "Treatment" Incentive group are told they will be given recommendations at the end of the study). After accepting the study instructions and passing an A/V check, every participant undergoes a five-minute warm-up listening session. For this session, they receive no information on how their preferences are learned and are told to behave as they would on Spotify or Pandora. Next, participants are randomly assigned to one of three

[4]In our experiment, songs are generated randomly (not personalized) until the very end (and only if the participant is in the Incentive treatment group. Therefore, under our formalization, one can view $\mathcal{A}$ as random until $t = T$.

Information conditions, which affects the instructions they receive for their second five-minute listening session, which we refer to as the test session.[5]

3.5 Post-Experiment Survey

At the end of the study, all participants are asked to complete a survey. The full list of questions is given in Appendix B.1. In addition to demographic information, we ask participants several multiple-choice/checkbox questions to query: (1) whether they changed the way they interacted across sessions and, if so, how; (2) how they believe their recommendation algorithms work on Spotify, Facebook, etc.; and (3) how much time they spend online. In addition, we ask one open-ended text question: *Do you ever try to "talk" to your algorithm or "hide" things from it? For example, do you ever give a song a "thumbs-up" just to Spotify that you want to see similar songs? Or do you sometimes avoid clicking on an advertisement just because you're worried about getting many similar advertisements in the future? If you do, tell us how and why.*

3.6 Analysis

We examine the data collected from our experimental procedure for signs of strategization. To test Hypotheses 3 and 4 from Section 2, we look at *average treatment effects* of the Information conditions and Incentive conditions on participant behavior. We then look at how these effects manifest across our outcome distributions, and whether there are heterogeneous effects by individual-level characteristics. Finally, we analyze a post-study survey that participants took in order to get a qualitative picture of strategization.

3.6.1 Outcome Variables. We pre-registered the following outcome variables.

(1) **Likes + Dislikes**. The number of songs that the participant has either liked (thumbs-up) or disliked (thumbs-down) during the session.
(2) **Fast Skips**. The number of times that the participant skips a song during the first 5 seconds of song during the session.
(3) **Dashboard Clicks**. The number of times that the participant clicks on the song player dashboard during the session. Any click on the like button (thumbs-up), dislike button (thumbs-down), or skipahead button (which allows participants to scroll through the song) counts as a click.
(4) **Average Song Dwell Time, Logged**. Average length of time participant listens to each recommended song in milliseconds, logged.
(5) **Standard Deviation Song Dwell Time, Logged**. Standard deviation of the time participant listens to each recommended song in milliseconds, logged.

Additionally, we pre-registered an analysis examining the proportion of (i) "likes" and "dislikes" and (ii) fast skips per song listened. Due to space constraints, we report these results in Appendix A; they do not change the interpretation of our findings.

3.6.2 Group Means and Average Treatment Effects. To test for the presence of strategization, we examine how the outcome variables (averaged across participants) in the test listening session differ across our (i) Information and (ii) Incentive conditions.

For each of our outcome variables, we fit a model with the respective outcome variable of interest as our dependent variable and treatment dummies for (i) the Incentive condition $D_{\text{Incentive}}$, (ii) the Information condition $D_{\text{Information}}$, and (iii) their interaction. In addition, we fit an additional

[5]After the two listening sessions, participants in the Incentive treatment are presented with three recommendations and asked to provide feedback on them. The data from this step is not used in our analysis; we undergo this step in order to fulfill our promise to these participants that they will receive recommendations.

specification that includes participants' pre-treatment behavior X_{pre} in the Warmup session as a control variable (e.g., we include the number of "likes" and "dislikes" in the Warmup Session as a control when Likes + Dislikes is our outcome of interest). In other words, we fit the outcome variable (with the appropriate model, as specified next) to the following:

$$\beta_0 + \beta_1 D_{\text{Incentive}} + \beta_2 D_{\text{Information}} + \beta_3 (D_{\text{Incentive}} \times D_{\text{Information}}) + \beta_4 X_{\text{pre}} + \varepsilon.$$

We now specify the models used for each of the outcome variables in Section 3.6.1. For our three count variables, (i) number of "likes" and "dislikes", (ii) number of fast skips, and (iii) number of dashboard clicks, we use a Poisson quasi-maximum-likelihood (quasi-Poisson) regression. We use a quasi-maximum-likelihood model in order to account for potential overdispersion in the engagement data [Wooldridge, 1999]. For our continuous dwell time variables, (i) log average song dwell time and (ii) log standard deviation of song dwell time, we use an OLS regression.

To report interpretable versions of the main effects of each condition, we calculate the average marginal effect (AME) of each treatment condition compared to the respective control group in that condition. For example, we report the difference in the average number of Likes + Dislikes predicted by our model between the Incentive "Treatment" and the Incentive "Control" condition, pooling across all levels of the Information condition.

3.6.3 Subgroup Means and Treatment Effects. In our pre-registration, we also stated we would examine potential heterogeneous effects among (i) participants younger than 25 years old and (ii) those who use TikTok, since we hypothesized that these subgroups might be more prone to strategic behavior. However, because we only had 72 participants below 25 (as our participants must also be at least 18 years old), we instead chose to look at participants who are either below and above 35 years old for greater statistical power. Due to a mistake, our final survey did not include a question specifically about TikTok (although we do ask about online platform use). We therefore divide our participants based on a different question where we ask participants about their use of music recommendation platforms, as this question closely aligns with our experimental setup. We refer to this question as Spotify Use for brevity, and we code Spotify Use greater than once per week as "Often", and less than or equal to once per week as "Rare."

We then calculate the conditional Average Treatment Effect (CATE) for each subgroup of interest, using the same methodology described in Section 3.6.2. We test for heterogeneity in our treatment effects across subgroups by examining the difference-in-CATEs (DICs) using a Wald Test.

4 RESULTS

In this section, we present and discuss our main results, as per Section 3. We provide further results in Appendices A and B. We find strong evidence supporting both Hypothesis 3 and Hypothesis 3. We also find that Age and Spotify Use do not moderate the effects of the Information condition and mildly moderate the effects of the Incentive condition, suggesting that strategization occurs across subgroups though is more prominent among users who are expected to gain more from strategizing "in the wild." We analyze why users strategize in Section 4.3.

4.1 Do people strategize?

Our results provide strong evidence that users strategize when interacting with recommender systems. We find that participants change their behavior (i) when they receive different information about how the recommender algorithm learns preferences (Hypothesis 3) and (ii) when the time horizon for the user-platform interaction is changed (Hypothesis 4), supporting our strong evidence that users are in fact strategic rather than naive. Figure 3 summarizes our main results, showing the means of our outcome variables across our Information and Incentive conditions, as described

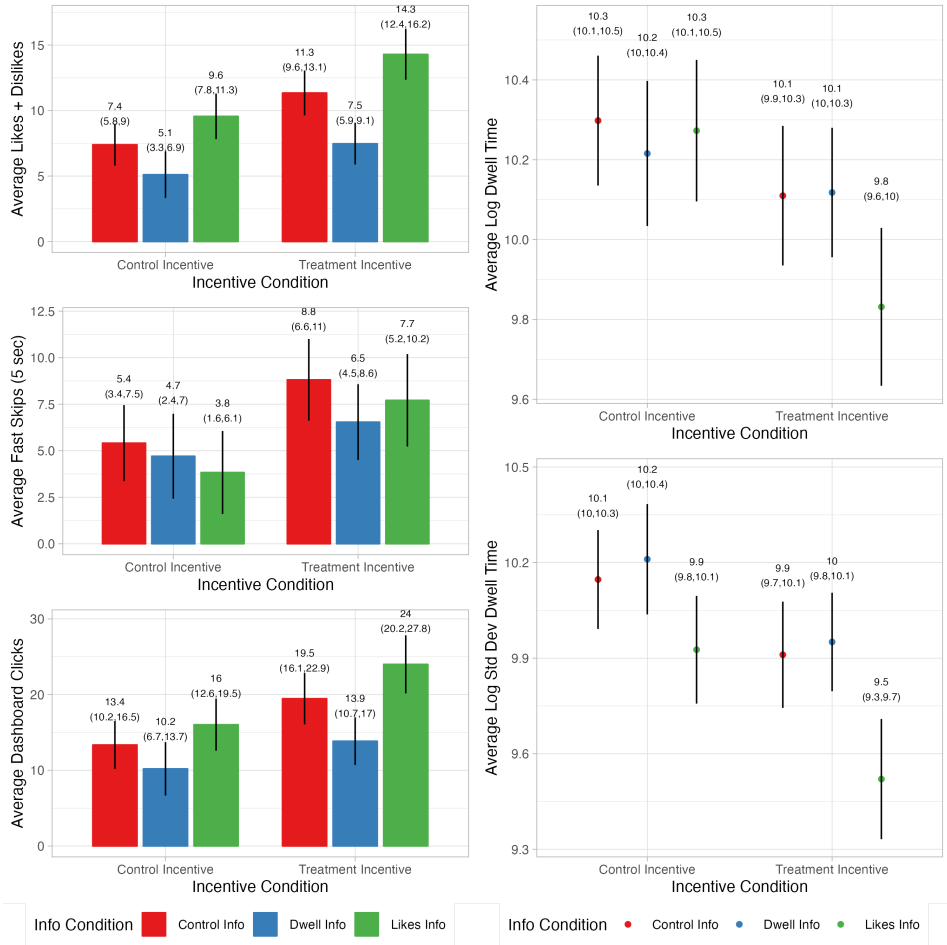


Fig. 3. Means and 95% CIs across our conditions for our 5 outcome variables of interest, as described in Section 3.6.1 and using the models in Section 3.6.2.

in Sections 3.6.1 and 3.6.2. We now discuss these results in light of our two Hypotheses: **H1** (which corresponds to Hypothesis 3) and **H2** (which corresponds to Hypothesis 4).

4.1.1 Strategization Under Different Information conditions. We first consider **H1**, which states that all else constant, providing participants with different descriptions of how their preferences are learned will lead to different participant behaviors. We test this hypothesis by examining how participant behavior changes across Information conditions, in which participants are either told that the algorithm is tracking (i) their "likes" and "dislikes," which we refer to as the "Likes" Information group or (ii) the time they spent listening to each song, which we refer to as the "Dwell" Information group, or else are told (iii) no explicit information about what data the algorithm is tracking, which we refer to as the "Control" Information group. (Note that although we pre-registered models that include interactions, which are shown in Figure 3 and Appendix Table A.3, we do not find any significant interactions between the Information and Incentive conditions and thus consider the main effects of these conditions separately in the discussion below.) Below, we

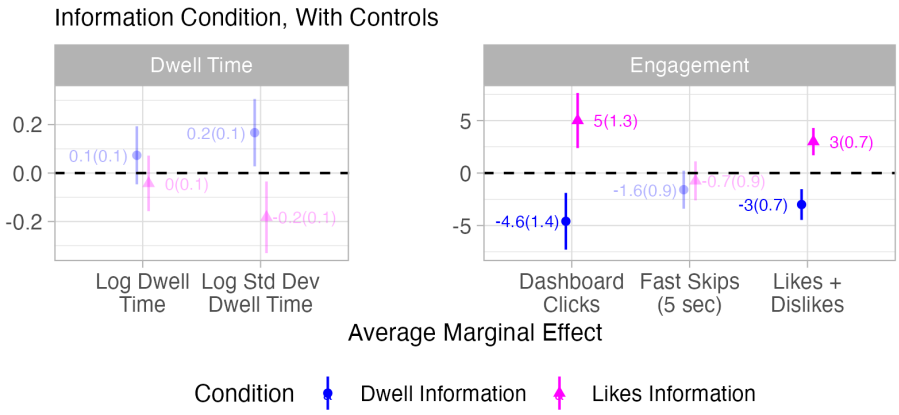


Fig. 4. Effect of the Likes Information condition and the Dwell-Time Information condition, compared to the Control Information condition, on participant behavior. *Left:* Average marginal effects of Information conditions on dwell time outcomes. Models are estimated using an OLS regression with controls for behavior in the Warmup session. *Right:* Average marginal effects of Information conditions on engagement outcomes. Models are estimated using a quasi-poisson regression with controls for behavior in the warmup session.

report the effects of the "Information" condition on our various outcomes (pooling across different levels of the Incentive condition) and controlling for participants behavior in the Warmup Session.

We find strong support for **H1**, as summarized in Figure 4, which visualizes average marginal effects, pooled across the Incentive conditions. Overall, participants' engagement patterns differ substantially across Information conditions, as we unpack below.

Engagement metrics. As expected, the number of "likes" and "dislikes" increases when participants believe that the algorithm is tracking that behavior. Participants in the "Likes" Information condition generate 3.0 (SE: .7, $p < .001$) more "likes" and "dislikes," and 5.0 more dashboard clicks (SE: 1.3, $p < .001$) on average than participants in the "Control" Information condition.

Interestingly, the number of "likes" and "dislikes" *decreases* when participants are in the "Dwell" Information condition. Participants in the "Dwell" Information condition, who are told that the algorithm is paying attention to how long they spend listening to each song, submit 3.0 (SE: .7, $p < .001$) *fewer* "likes" and "dislikes," and 4.6 (SE: 1.4, $p < .001$) fewer dashboard clicks on average than those in the "Control" Information condition. Since participants in the "Control" Information condition are told to behave as they would on their music platform of choice (e.g., Spotify), this result suggests that participants use "likes" and "dislikes" to strategize in the wild (e.g., based on their understanding of Spotify's algorithm). These results suggest that participants develop an understanding of how algorithms learn preferences and adjust their behavior based on this understanding, as we would expect based on our model of strategic behavior.

Overall, the effects of the Information treatment on dashboard clicks as well as "likes" and "dislikes" are substantial in magnitude. For example, participants in the "Likes" condition (M=11.7) like or dislike 80% more songs than participants in the "Dwell" condition (M=6.4). That is, with only minor changes in information about which data the platform is tracking, and *no actual changes to the algorithm itself*, our treatment induces large changes in engagement behavior. We observe the same pattern in the number of dashboard clicks but not in fast skips, as shown in Figure 3. Since skips are related to dwell time (a greater number of skips implies a shorter dwell time for a fixed, five-minute listening session), we discuss skips next.

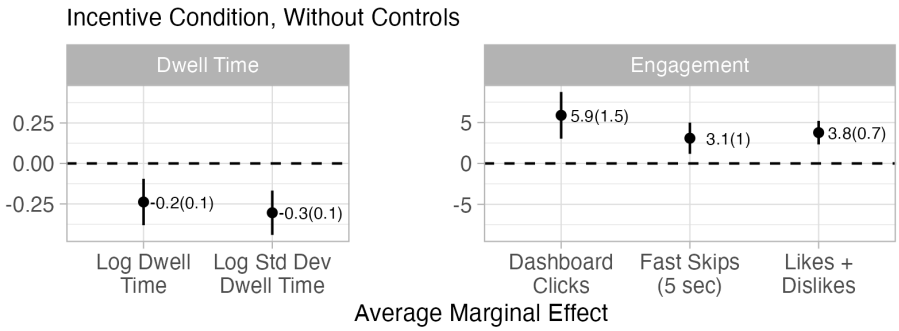


Fig. 5. Effect of the Treatment Incentive condition compared to the Control Incentive condition on participant behavior. *Left:* Average marginal effects of the Incentive condition on dwell time outcomes. Models are estimated using an OLS regression. *Right:* Average marginal effects of Incentive condition on engagement outcomes. Models are estimated using a quasi-poisson regression. Neither model is estimated with warmup session controls because participants are randomized into an Incentive condition before the Warmup session

Dwell time metrics. In contrast, we do not observe significant differences of the Information condition on the (i) number of "fast skips" or (ii) the average log dwell time of the participants. However, we do see weak evidence that the Information condition affects the variance of participants' time spent listening to songs. Participants in the "Likes" condition have a slightly smaller log standard deviation of dwell time than participants in the "Control" condition (δ = -.2, p = .06), and participants in the "Dwell" condition have a slightly larger log standard deviation of dwell time than participants in the "Control" condition (δ = .2, p = .06).

While the effects for the log standard deviation of dwell time are suggestive, we note that they are not significant at the α = .05 level after adjusting for multiple comparisons. Nonetheless, these results are consistent with an account of participants varying the time they spend listening to songs more when they receive information that the algorithm is tracking dwell time, even if they are not necessarily consistently increasing or decreasing the time spent per song, on average. That is, it suggests that when users are aware that their algorithm is tracking dwell time, they spend more time on songs they like and less time on songs they do not like, thus increasing the variance of dwell time across songs rather than the average dwell time itself.

4.1.2 Strategization Under Different Incentive conditions. We now turn to consider **H2**, which states that participants behave differently if they believe that the time horizon under which their interactions affect their later recommendations changes. We test this hypothesis by examining how participant behavior changes under the Incentive condition, in which participants are either told that the algorithm is learning their preferences in order to provide them with personalized recommendations that they must evaluate after their listening sessions ("Treatment" Incentive) or told that the algorithm is learning general music preferences ("Control" Incentive). Consistent with the previous analysis, the reported results are estimated by pooling across different levels of the Information condition. Because participants are randomized to the Incentive condition before the Warmup session, we do not include controls for participants' Warmup session behavior (since doing so would be controlling for a post-treatment variable).

Overall, we find strong support for **H2**: Participants' engagement patterns differ substantially across Incentive conditions, as summarized in Figure 5.

Table 1. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Incentive Condition, by Age, Without Controls, pooled across Information Conditions

Outcome	ATE		DIC
	Below-35	Above-35	
Likes + Dislikes	2.7*(1.06)	4.37***(1.02)	-1.66 (1.47)
Fast Skips (5 sec)	5.61***(1.58)	1.29(1.3)	4.32* (2.05)
Dashboard Clicks	3.24(2.42)	8.42***(1.85)	-5.18† (3.04)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

Engagement metrics. Participants in "Treatment" Incentive condition engage at higher rates than participants in the "Control" Incentive condition. The results are summarized in Figure 5. Participants in the "Treatment" Incentive condition, who are told that the algorithm is tracking their behavior in order to provide them with personalized recommendations, submit 3.8 (SE: .7, $p < .001$) more "likes" and "dislikes," 3.1 more "fast skips" (SE: 1.0, $p = .008$), and 5.9 more dashboard clicks (SE: 1.5, $p < .001$) than participants in the "Control" Incentive condition. These results show that the participants in the "Treatment" Incentive engage more with the music player than participants in the "Control" Incentive condition, presumably to shape their future recommendations.

Dwell time metrics. Additionally, we find that participants in the "Treatment" Incentive condition listen to songs for a shorter amount of time (δ = -.2, SE:.1, p =.006), and have a smaller standard deviation of dwell time (δ = -.3, SE: .1, $p < .001$) than participants in the "Control" Incentive condition. These results, along with the increase in the number of "fast skips", show that participants in the "Treatment" Incentive condition are skipping through songs relatively quickly on average, rather than exploring songs for longer amounts of time. These results are consistent with an account of participants who believe they are interacting with a personalization algorithm seeking to "train" the algorithm by sifting through content quickly to provide more feedback, rather than maximizing short-term utility by listening to songs they enjoy for longer periods of time.

4.2 Subgroup Analysis

We have thus far found evidence that participants, on average, engage in strategic behavior. In this section, we explore whether certain types of participants more likely to be strategic than others. In particular, we highlight two theoretically relevant individual characteristics: Age and Spotify Use. (For brevity, we only examine a subset of our outcome variables and potential moderating characteristics; for the full set of moderators and outcomes, see Appendix A.4). All the results below (and Tables 1 to 4) give the CATEs and DICs by subgroup across Incentive and Information conditions. Note that, when we refer to average treatment effects, we are computing the average marginal effect.

Age. First, we consider the participants' age as a potential moderator of our treatment effects. Research has shown that younger people are more familiar with technology and have higher digital literacy than older adults, with people born after 1980 sometimes referred to as "digital natives" [Broady et al., 2010, Palfrey and Gasser, 2011, Vercruyssen et al., 2023]. Moreover, TikTok, a platform heavily dependent on recommendation algorithms for content curation and known from previous qualitative studies to have users who show awareness of the algorithm's behavior, is exceptionally popular among the younger demographic [Anderson, 2021, Klug et al., 2021]. Therefore, we hypothesized that younger participants would show more evidence of strategization.

Table 2. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Information Condition, by Age, With Controls, pooled across Incentive Conditions

Outcome	Info Condition	ATE		DIC
		Below-35	Above-35	
Likes + Dislikes	Likes	3.54**(1.09)	2.15**(0.78)	1.38 (1.35)
Likes + Dislikes	Dwell	-2.27*(1.09)	-3.62***(0.95)	1.35 (1.44)
Fast Skips (5 sec)	Likes	-2.11(1.51)	-0.48(1.17)	-1.63 (1.92)
Fast Skips (5 sec)	Dwell	-1.78(1.34)	-1.52(1.11)	-0.26 (1.75)
Dashboard Clicks	Likes	6.81**(2.36)	2.88†(1.56)	3.94 (2.82)
Dashboard Clicks	Dwell	-3.52(2.42)	-5.26***(1.44)	1.74 (2.81)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

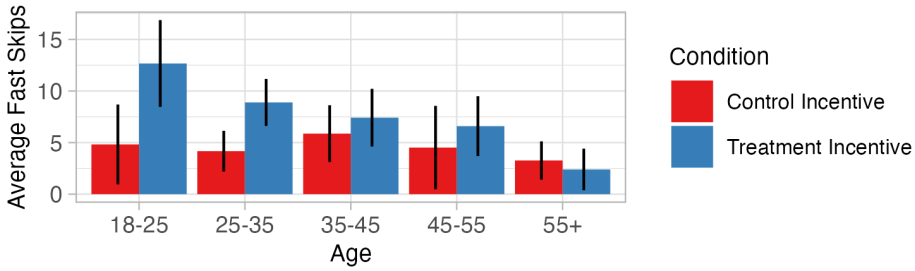


Fig. 6. Average number of “fast skips” (with 95% CIs) for “Treatment" vs. “Control" Incentive groups, for different Age groups

Specifically, we examine whether Age (split at 35 years old) moderates the effects of our Incentive and Information conditions.

Against our expectations, we do not find consistent evidence that younger participants are significantly more affected by our treatments. The results are summarized in Tables 1 and 2. As can be seen, participants below age 35 and above age 35 *both* exhibit evidence of strategic behavior. For example, both older and younger participants show a significantly increased number of “likes” and “dislikes” in the “Treatment" Incentive and "Likes" Information conditions, and significantly decreased number of “likes” and “dislikes” in the “Dwell" Information condition. Furthermore, the difference in conditional average treatment effects (DICs) for older and younger participants is not significant across any measures (with one exception, which we consider in the next paragraph). These findings demonstrate that even older participants—who one might expect would show less sophistication when interacting with algorithms—on average exhibit evidence of strategic behavior. That is, we find evidence of strategic behavior even in subgroups that one might expect to be naive.

Upon closer examination of the results, we do find some evidence that the *mechanism* of that strategization—i.e., the types of behaviors changed—differs for older vs. younger participants. In particular, we consider case of “fast skips"—when the participant skips past a song within 5 seconds of it starting. Figure 6 shows a substantial trend between participants’ age group and the average number of “fast skips" in the “Incentive" group where younger participants produce substantially more “fast skips."

Table 3. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Incentive Condition, by Spotify Use, Without Controls, pooled across Information Conditions

Outcome	ATE		DIC
	Spotify Use=Rare	Spotify Use=Often	
Likes + Dislikes	1.54(1.13)	4.68***(0.95)	-3.13* (1.47)
Fast Skips (5 sec)	0.87(1.46)	4.17**(1.28)	-3.3† (1.94)
Dashboard Clicks	1.91(2.28)	7.53***(1.88)	-5.62† (2.95)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

Spotify Use. We next consider the participants' prior Spotify use as a moderator. We hypothesized that participants with greater Spotify use would exhibit greater levels of strategization for two reasons. First, frequent participants are likely more familiar with music recommender platforms, and therefore, might be more comfortable in engaging with the platform. Second, frequent users of Spotify likely enjoy music, and therefore might be more incentivized to strategize in order to receive higher payoff in the form of better song recommendations.

As shown in Table 4, we find evidence that both frequent and infrequent Spotify users respond to our Information conditions, and no strong evidence of moderation of this effect across subgroups. However, as shown in Table 3, we do find some evidence that past Spotify use moderates treatment effects in the Incentive condition. Frequent Spotify users (i.e. those who use the app more than once a week) submitted 4.68 more "likes" and "dislikes," 4.16 more "fast skips", and 7.38 more Dashboard Clicks in the Treatment vs. Control Incentive conditions. In contrast, less frequent users submitted 1.54 more "likes" and "dislikes," .87 more "fast skips", and 1.91 more Dashboard Clicks in the Treatment vs. Control condition. These give us difference-in-CATEs of -3.13 (p = .03), -3.31 (p = .09), -5.67 (p = .05) respectively, showing that users who use Spotify more frequently have larger treatment effects in the "Incentive" condition. We note that while these differences are large in magnitude, they are imprecisely measured. This is because our analysis is underpowered to detect small differences across groups, largely because the relatively few number of infrequent Spotify users in our experiment (200 infrequent vs. 452 frequent users). Nonetheless, these findings are suggestive: Participants who we would expect to derive more long-term payoff from better song recommendations exhibit greater evidence of strategization in our "Incentive" condition.

In summary, while we find evidence that the degree and mechanism of strategization might vary across subgroup, we do not find evidence that strategic behavior is wholly concentrated within a particular type of user. Rather, some degree of strategization is common across types of users.

4.3 Post-Experiment Survey

We now add qualitative insights into users strategization behavior by analyzing the results of our post-experiment survey. Of the "Treatment" Incentive participants, 60 percent reported that they changed their behavior between sessions, 39 reported that they did not, and 1 percent reported "I don't know." Of the "Control" Incentive participants, 41 percent reported that they changed their behavior between sessions, 58 reported that they did not, and 1 percent reported "I don't know." We provide further details on participant responses in Appendix B.1.

We manually analyze the open-ended responses, in which we asked participants whether they strategize on their own recommendation platforms. We find that around 20 percent of participants report definitive strategization, 42 percent report not strategizing, and 38 percent provide information for which it is unclear (e.g., some participants indicate clear awareness of their algorithm, but

Table 4. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Information Condition, by Spotify Use, With Controls, pooled across Incentive Conditions

Outcome	Info Condition	ATE		DIC
		Spotify Use=Rare	Spotify Use=Often	
Likes + Dislikes	Likes	2.28*(1.09)	3.05***(0.82)	-0.77 (1.39)
Likes + Dislikes	Dwell	-4.88***(1.31)	-2.25*(0.89)	-2.63† (1.57)
Fast Skips (5 sec)	Likes	-2.47†(1.49)	-0.37(1.23)	-2.1 (1.94)
Fast Skips (5 sec)	Dwell	-1.47(1.3)	-1.43(1.18)	-0.04 (1.75)
Dashboard Clicks	Likes	3.72*(1.82)	5.25**(1.81)	-1.54 (2.56)
Dashboard Clicks	Dwell	-6.79***(1.9)	-3.94*(1.78)	-2.85 (2.6)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

it is unclear whether "liking" to express approval to the algorithm is naive or strategic). Finally, we analyze why and how users strategize. We identified several trends that persist across participants and include example responses below.

Being pigeonholed by algorithms: Some participants express that they do not like to be pigeonholed by their algorithm, with one stating "what I like today might not be what I will like tomorrow," another saying "Yes sometimes I may like a song but not thumbs-up the song because I don't want my feed filled with similar artists/videos. This is because I might like only one type of song by an artist," and a third sharing that "On YouTube I will like things I don't and dislike things I do and subscribe to dozens and dozens of channels, even walk out of the room with something I like or dislike playing just so I get lots of new stuff and they don't pigeonhole me too much and show me crap I don't want to see over and over. Basically, I try to be purposefully unpredictable and then go into my subscriptions and play from there the stuff I really want to see. My hope is the two are playing against each other and the algorithm doesn't know exactly what I want." Similarly, several say that they like to reset their algorithm, stating: "I have played some music I would not normally listen to or even like to throw off an algorithm" and "I might give thumbs-up to specific songs if I am trying to reset the algorithm and get it to forget what I have been listening to."

Helping the algorithm. Several participants suggest that they strategize to help their algorithm identify their preferences. One said: "[D]uring the sessions a Blink-182 song came on, and I'm not really crazy about them, but I was hoping to force the algorithm to swing more towards a 'rock' vibe" while another responded: "Yes. Thumbs upped songs in this survey that I didn't like because I wanted to hear similar bands. I hated that Blink 182 song, but I love Blink and I love punk music so I thumbs upped it anyway. Sometimes you gotta play along with the algorithm if you want it to work best for you." Others said "If I'm looking for more recommendations that are similar to a certain genre I will leave a playlist based on that genre playing for a day or two to try and get different recommendations matching those songs" and "I have frequently given thumbs up or not skipped a mediocre song by an artist that I otherwise love because I want their songs to continue to show up." Some even indicate awareness of more subtle recommendation tactics, like dwell-time tracking: "If I see something that I know I am not interested in, I quickly click away from it, I do not want to linger too long or the algorithm may think I am interested and show me more like it."

Preserving accounts. We find that several participants did not want to "ruin" their algorithm with unintended interactions: "I try not to link my account to others to avoid them "poisoning"

my algorithm with their preferences since algorithms assume there must be some kind of overlap between you and those you associate with" and "Yes, I often do like songs or avoid clicking links or ads that would impact my user profile on various platforms. I am aware that my activity often gets tracked and that the algorithms on social media or music sites detect the changes and cater to my new preferences. Sometimes, I do not want that to happen so I avoid clicking links. If I am with a friend who has a different music taste and wants to search something on my phone, I am often scared that it will impact my own music recommendations and so I try to limit that." Several even confess to creating multiple accounts: ""I have many YouTube accounts so my algorithm does not pick up a YouTube link a friend sends me to watch."

Private browsing. Many of our participants admit to using Incognito or private browsing mode to interact with interesting content: "If I want to just check something but not mess up my preferences, I will use incognito mode in Chrome so I'm not signed in" and "I avoid searching something embarrassing unless it is in incognito mode, because I expect I would get ads related to it after."

Using tracking to their advantage. Some participants strategize off-platform, as epitomized by the response: "If there's something I am interested in and haven't seen an ad for it, I will google it because I know within a very short amount of time, ads will start appearing in my feeds."

No strategization. Many participants reported not strategizing, responding: "I believe that algorithms are a useful tool that can help us make better decisions and find new insights" and "I'm pretty (and blatantly) honest about my feelings. And yes, this sometimes gets me into trouble, but it's easier to be honest about something than not."

5 DISCUSSION

Our results demonstrate that users are not only cognizant of recommendation algorithms, but also use knowledge of their algorithm to elicit better future recommendations. While much of the discussion around recommender systems focuses on the platform's role in shaping what users see or (in the case of social media) the content creators' role in choosing the type of content to create, we show that users also play an active role in shaping what they see.

We provide a first step into documenting and measuring strategization through a online lab experiment. We find strong evidence of strategization: participants change their behaviors in response to their beliefs about how the recommendation algorithm works. In particular, we find that participants change their behavior in response to (a) information about what types of user behaviors the algorithm pays attention to and (b) whether they will receive personalized recommendations based on their behaviors. The magnitude of this strategization is substantial and apparent across multiple outcome metrics, not just concentrated among particularly active users. Furthermore, even users thought to be naive (e.g., older participants) exhibit evidence of strategization.

Strategization implies that user behavior is not exogenous as commonly assumed, i.e., how a user interacts with their content does not simply depend on the content itself, but also on the algorithm that generated it. Such strategization can hurt the platform, e.g., its ability to repurpose data gathered under one algorithm for inference about a different algorithm [Cen et al., 2023].

There are several directions for future work. Two natural directions involve studying what *causes* users to strategize and *how* they strategize. For instance, do users want a more heterogeneous set of recommendations? Does strategization arise because users have changing (e.g., inconsistent [Kleinberg et al., 2022]) preferences? How do users encourage the algorithm to behave accordingly? Another direction is to measure and quantify the extent to which strategization happens on a real platform, as well as compare the magnitude of strategization across different platforms. This endeavor is difficult because it relies on users' perceptions of their recommendation algorithms,

which is difficult to account for because it may vary across users. Finally, in the face of strategization, platforms must find more robust preference elicitation methods. From an understanding of why and how users strategize in practice, the next step would be to develop such methods (e.g., [Cen et al., 2023] suggest pursuing trustworthy algorithm design).

ACKNOWLEDGMENTS

The authors would like to thank Professor Dean Eckles for fruitful conversations and the Mądry group for testing our system and providing feedback.

REFERENCES

Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.

Brooke Auxier and Monica Anderson. 2021. Social Media Use in 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Arturo Arriagada and Francisco Ibáñez. 2020. "You need at least one picture daily, if not, you're dead"': content creators and platform evolution in the social media ecology. *Social Media+ Society* 6, 3 (2020), 2056305120944624.

John Beshears, James J Choi, David Laibson, and Brigitte C Madrian. 2008. How are preferences revealed? *Journal of public economics* 92, 8-9 (2008), 1787–1794.

Tim Broady, Amy Chan, and Peter Caputi. 2010. Comparison of older and younger adults' attitudes towards and abilities with computers: Implications for training and learning. *British Journal of Educational Technology* 41, 3 (May 2010), 473–485. https://doi.org/10.1111/j.1467-8535.2008.00914.x

Michael Brückner and Tobias Scheffer. 2009. Nash equilibria of static prediction games. *Advances in neural information processing systems* 22 (2009).

Sarah Cen, Andrew Ilyas, and AleksandeR Madry. 2023. User Strategization and Trust on Data-Driven Platforms. In *arXiv preprint*.

Michael Ann DeVito. 2021. Adaptive Folk Theorization as a Path to Algorithmic Literacy on Changing Platforms. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 1–38.

Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms ruin everything": #RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3163–3174. https://doi.org/10.1145/3025453.3025659

Michael A DeVito, Jeffrey T Hancock, Megan French, Jeremy Birnholtz, Judd Antin, Karrie Karahalios, Stephanie Tong, and Irina Shklovski. 2018. The Algorithm and the User: How Can HCI Use Lay Understandings of Algorithmic Systems?. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18, Paper panel04)*. Association for Computing Machinery, New York, NY, USA, 1–6.

Benjamin Edelman and Michael Ostrovsky. 2007. Strategic bidder behavior in sponsored search auctions. *Decision support systems* 43, 1 (Feb. 2007), 192–198.

Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*. 2371–2382.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science* (Cambridge, Massachusetts, USA) *(ITCS '16)*. Association for Computing Machinery, New York, NY, USA, 111–122.

Andreas Haupt, Dylan Hadfield-Menell, and Chara Podimata. 2023. Recommending to Strategic Users. (Feb. 2023). arXiv:2302.06559 [cs.CY]

Jiri Hron, Karl Krauth, Michael I. Jordan, Niki Kilbertus, and Sarah Dean. 2023. Modeling Content Creator Incentives on Algorithm-Curated Platforms. arXiv:2206.13102 [cs.GT]

Justin T Huang, Rupali Kaul, and Sridhar Narayanan. 2022. The Causal Effect of Attention and Recognition on the Nature of User-Generated Content: Experimental Results from an Image-Sharing Social Network. (2022).

Daniel Huttenlocher, Hannah Li, Liang Lyu, Asuman Ozdaglar, and James Siderius. 2023. Matching of users and creators in two-sided markets with departures. *arXiv preprint arXiv:2401.00313* (2023).

Nicole Immorlica, Meena Jagadeesan, and Brendan Lucier. 2024. Clickbait vs. Quality: How Engagement-Based Optimization Shapes the Content Landscape in Online Platforms. *arXiv preprint arXiv:2401.09804* (2024).

Meena Jagadeesan, Nikhil Garg, and Jacob Steinhardt. 2022. Supply-side equilibria in recommender systems. *arXiv preprint arXiv:2206.13489* (2022).

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2022. The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization. In *Proceedings of the 23rd ACM Conference on Economics and Computation* (Boulder, CO, USA) *(EC '22)*. Association for Computing Machinery, New York, NY, USA, 29.

Daniel Klug, Yiluo Qin, Morgan Evans, and Geoff Kaufman. 2021. Trick and Please. A Mixed-Method Study On User Assumptions About the TikTok Algorithm. In *Proceedings of the 13th ACM Web Science Conference 2021 (WebSci '21)*. Association for Computing Machinery, New York, NY, USA, 84–92. https://doi.org/10.1145/3447535.3462512

Angela Y Lee, Hannah Mieczkowski, Nicole B Ellison, and Jeffrey T Hancock. 2022. The Algorithmic Crystal: Conceptualizing the Self through Algorithmic Personalization on TikTok. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022), 1–22.

Sagi Levanon and Nir Rosenfeld. 2022. Generalized strategic classification and the case of aligned incentives. In *International Conference on Machine Learning*. PMLR, 12593–12618.

Aarian Marshall. 2020. Uber Changes Its Rules, and Drivers Adjust Their Strategies. *Wired* (Feb. 2020).

Carey K Morewedge, Sendhil Mullainathan, Haaya F Naushan, Cass R Sunstein, Jon Kleinberg, Manish Raghavan, and Jens O Ludwig. 2023. Human bias in algorithm design. *Nature Human Behaviour* (2023), 1–3.

Simha Mummalaneni, Hema Yoganarasimhan, and Varad Pathak. 2023. How Do Content Producers Respond to Engagement on Social Media Platforms? (2023).

Arvind Narayanan. 2022. How To Train Your TikTok. https://knightcolumbia.org/blog/how-to-train-your-tiktok. Accessed: 2023-11-10.

Arvind Narayanan. 2023. Understanding Social Media Recommendation Algorithms. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms. Accessed: 2023-11-10.

Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David Levy, and Rasmus Kleis Nielsen. 2018. Reuters Institute Digital News Report 2018. (June 2018).

John Palfrey and Urs Gasser. 2011. *Born digital: Understanding the first generation of digital natives*. ReadHowYouWant. com. https://books.google.com/books?hl=en&lr=&id=wWTI-DbeA7gC&oi=fnd&pg=PR2&dq=Born+digital:+Understanding+the+first+generation+of+digital+natives&ots=wJiB7Fw2_C&sig=PTpXObKaqr576ANATxCY9kOMqW8

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.

Hatim A Rahman. 2021. The Invisible Cage: Workers' Reactivity to Opaque Algorithmic Evaluations. *Administrative science quarterly* 66, 4 (Dec. 2021), 945–988.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2010. Introduction to recommender systems handbook. In *Recommender systems handbook*. Springer, 1–35.

Donghee Shin. 2020. How do users interact with algorithm recommender systems? The interaction of users, algorithms, and performance. *Computers in human behavior* 109 (Aug. 2020), 106344.

Ellen Simpson, Andrew Hamann, and Bryan Semaan. 2022. How to Tame "Your" Algorithm: LGBTQ+ Users' Domestication of TikTok. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP (Jan. 2022), 1–27.

Nathaniel Sirlin, Ziv Epstein, Antonio A Arechar, and David G Rand. 2021. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. (2021).

WSJ Staff. 2021. Inside TikTok's Algorithm: A WSJ Video Investigation. *Wall Street Journal* (July 2021). https://www.wsj.com/articles/tiktok-algorithm-video-investigation-11626877477

Samuel Hardman Taylor and Mina Choi. 2022. An Initial Conceptualization of Algorithm Responsiveness: Comparing Perceptions of Algorithms Across Social Media Platforms. *Social Media + Society* 8, 4 (Oct. 2022), 20563051221144322.

Anina Vercruyssen, Werner Schirmer, and Dimitri Mortelmans. 2023. How" basic" is basic digital literacy for older adults? Insights from digital skills instructors. In *Frontiers in Education*, Vol. 8. 1–11.

Jeffrey M. Wooldridge. 1999. Quasi-Likelihood Methods for Count Data. In *Handbook of Applied Econometrics Volume 2: Microeconomics*. John Wiley & Sons, Ltd, 321–368. https://doi.org/10.1111/b.9780631216339.1999.00009.x Section: 8 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/b.9780631216339.1999.00009.x.

WSJ. 2021. https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/6C0C2040-FF25-4827-8528-2BD6612E3796

A ADDITIONAL EXPERIMENTAL DETAILS

A.1 Covariate Balance

In Table 5, we report covariates across condition. We also report Chi-Squared tests to assess covariate balance and find that our demographics are balanced across conditions.

Table 5. Percent of Condition that (i) Is Female, (ii) Is Under 35, (iii) Is College Educated, (iv) Is Privacy Concerned, (v) Frequently uses social media, and (vi) Frequently uses Spotify Balance Checks for Demographics, by Conditions. Balance check is done via chi-square test, with p-values adjusted with a Benjamini-Hochbert correction.

Variable	Control Incentive Control Info	Control Incentive Dwell Info	Control Incentive Likes Info	Treatment Incentive Control Info	Treatment Incentive Dwell Info	Treatment Incentive Likes Info	statistic	p.adj
Is Female	50%	43%	53%	48%	55%	59%	6.44	0.53
Is Under 35	50%	43%	52%	36%	41%	40%	8.80	0.47
Has College Edu	59%	57%	56%	69%	62%	51%	7.99	0.47
Is Privacy Concerned	29%	32%	24%	30%	31%	31%	1.85	0.87
Is Frequent Social Media User	75%	75%	77%	73%	68%	74%	3.13	0.87
Is Frequent Spotify User	42%	36%	40%	41%	36%	35%	2.30	0.87

A.2 Model, Proportion Regressions

In addition to the outcomes reported in the main text, we also pre-registered two outcome variables: Proportion of Likes + Dislikes (per song listened) and Proportion of Fast Skips (per song listened). To estimate this, we estimate the following models using a quasi-binomial rate regression, where Y_i is the count of our outcome of interest (either (i) Likes + Dislikes or (ii) Fast Skips, respectively) and S_i is the total number of songs for user i. The results are shown in Table 8.

Without Controls

$$\log \frac{Y_i}{S_i} \sim \beta_0 + \beta_1 \text{incent}_i + \beta_2 \text{likes}_i + \beta_3 \text{dwell}_i + \beta_4 \text{incent}_i \times \text{likes}_i + \beta_5 \text{incent}_i \times \text{dwell}_i + \epsilon_i \quad (3)$$

With Warmup Session Controls

$$\log \frac{Y_i}{S_i} \sim \beta_0 + \beta_1 \text{incent}_i + \beta_2 \text{likes}_i + \beta_3 \text{dwell}_i + \beta_4 \text{incent}_i \times \text{likes}_i + \beta_5 \text{incent}_i \times \text{dwell}_i + Y_i^0 + \epsilon_i \quad (4)$$

A.3 Regression Tables, Average Treatment Effect

Below we show the regression tables for the Average Treatment Effects (ATEs) for our outcome variables. Table 6 shows the results for our count variables (i) Likes + Dislikes, (ii) Fast Skips, and (iii) Dashboard Clicks. Table 7 shows the results for our continuous dwell time variables (i) Log Dwell Time and (ii) Log St. Dev. Dwell Time. Table 8 shows the results for our proportion variables (i) Proportion of Likes + Dislikes and (ii) Proportion of Fast Skips.

A.4 Subgroup Analysis

Figure 7 shows the CATEs for the Incentive condition on our count variables for participants above and below 35, respectively. In the“Treatment" Incentive condition, younger participants produced substantially more “Fast Skips" than those in the Incentive “Control" condition. We find that participants below 35 in the had 5.64 more “Fast Skips" in the “Treatment" vs. “Control" Incentive conditions; whereas participants above 35 had only 1.29 more “Fast Skips" in the “Treatment" vs. “Control" Incentive conditions, for a significant difference-in-CATEs of 4.36 ($p = .03$). In contrast, in the Incentive “Treatment" condition (vs. Incentive “Control") older participants have 4.4 more likes and dislikes (compared to 2.7 for younger participants), and 8.5 more Dashboard Clicks (compared to 3.25 for younger participants) – although the difference in CATEs for these two measures is not significantly different from zero. These findings, although exploratory, suggest that dwell time might be more salient metric for younger users than for older users. For example, TikTok, which has a particularly young userbase, uses watch time as the most important metric for generating new

Table 6. Quasi-Poisson Regression

Dependent Var	Likes + Dislikes		Fast Skips (5 sec)		Dashboard Clicks	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Constant	2.1***	1.6***	1.7***	1.2***	2.7***	2.1***
	(0.05)	(0.07)	(0.10)	(0.08)	(0.05)	(0.05)
1(Incentive)	0.39***	0.25***	0.51***	0.19	0.35***	0.15*
	(0.09)	(0.07)	(0.19)	(0.15)	(0.11)	(0.08)
1(Likes Info)	0.27***	0.34***	-0.22	-0.09	0.21*	0.33***
	(0.09)	(0.08)	(0.18)	(0.16)	(0.11)	(0.08)
1(Dwell Info)	-0.37***	-0.33***	-0.19	-0.20	-0.29***	-0.28***
	(0.10)	(0.09)	(0.18)	(0.15)	(0.11)	(0.09)
1(Incentive) × 1(Likes Info)	-0.07	-0.09	0.17	-0.30	0.005	-0.17
	(0.18)	(0.17)	(0.37)	(0.32)	(0.21)	(0.16)
1(Incentive) × 1(Dwell Info)	-0.09	-0.04	-0.21	-0.48	-0.10	-0.19
	(0.20)	(0.18)	(0.37)	(0.30)	(0.22)	(0.17)
Likes + Dislikes, Warmup		0.05***				
		(0.004)				
Fast Skips (5 sec), Warmup				0.05***		
				(0.003)		
Dashboard Clicks, Warmup						0.03***
						(0.002)
Fit statistics						
Observations	657	657	657	657	657	657
Squared Correlation	0.09	0.41	0.02	0.35	0.05	0.46

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

video recommendations [Staff, 2021]. On the other hand, older users show effects of substantial magnitude when it comes to more explicit forms of feedback, e.g. liking and disliking behavior.

In addition, we report additional subgroup analysis for (i) age (below or above 35) and (ii) Spotify use ("often", coded as greater than once per week vs. "rare", coded as less than or equal to once per week) for our continuous dwell time variables.

We also examine self-reported behavior change ("yes", coded as "probably" or "definitely" yes, vs. "no", coded as "probably" or "definitely" no) for our count variables and continuous dwell time variables as well.

Interestingly, we do not observe evidence of substantial moderation by whether or not participants self-reported changing their behavior in our post-survey. For example, even participants who did not report changing their behavior showed significant increased numbers of likes and dislikes in the "Likes Information" Condition and fewer likes and dislikes in the "Dwell Condition – and the difference-in-CATEs between the two subgroups were not significant. This suggests that users might engage in "strategization" without consciously acknowledging that they are doing so.

Table 7. OLS Regression

Dependent Var	Log Dwell Time		Log Std Dev Dwell Time	
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	10.1***	2.5***	9.9***	4.9***
	(0.04)	(0.36)	(0.04)	(0.46)
1(Incentive)	-0.27***	-0.02	-0.34***	-0.15**
	(0.08)	(0.06)	(0.08)	(0.07)
1(Likes Info)	-0.16*	-0.04	-0.31***	-0.18**
	(0.09)	(0.06)	(0.09)	(0.07)
1(Dwell Info)	-0.04	0.07	0.06	0.17**
	(0.09)	(0.06)	(0.08)	(0.07)
1(Incentive) × 1(Likes Info)	-0.24	0.10	-0.16	0.12
	(0.19)	(0.12)	(0.18)	(0.15)
1(Incentive) × 1(Dwell Info)	0.10	0.22*	-0.03	0.09
	(0.18)	(0.12)	(0.17)	(0.14)
Log Dwell Time, Warmup		0.74***		
		(0.04)		
Log Std Dev Dwell Time, Warmup				0.50***
				(0.05)
Fit statistics				
Observations	657	657	657	651
R^2	0.03	0.56	0.05	0.31
Adjusted R^2	0.02	0.56	0.05	0.30

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

A.5 ECDFs, Dwell Variables

In Figure 8, we plot the ECDFs for our incentive and information conditions, respectively, for our continuous dwell time variables. We see evidence that the Treatment Incentive Condition decreases Log Average dwell time

Table 8. Proportion Regression

Dependent Var	Proportion of Likes + Dislikes		Proportion of Fast Skips (5 sec)	
Model:	(1)	(2)	(3)	(4)
Variables				
Constant	0.09	-1.9***	-0.72***	-2.0***
	(0.09)	(0.13)	(0.10)	(0.10)
1(Incentive)	0.17	0.23*	0.32*	0.11
	(0.17)	(0.13)	(0.19)	(0.14)
1(Likes Info)	0.59***	0.80***	-0.36*	-0.29**
	(0.17)	(0.14)	(0.18)	(0.14)
1(Dwell Info)	-0.48***	-0.40***	-0.14	-0.19
	(0.17)	(0.15)	(0.19)	(0.13)
1(Incentive) × 1(Likes Info)	-0.32	-0.06	0.09	-0.03
	(0.34)	(0.30)	(0.36)	(0.29)
1(Incentive) × 1(Dwell Info)	-0.005	-0.006	-0.14	-0.19
	(0.34)	(0.29)	(0.37)	(0.27)
Proportion of Likes + Dislikes, Warmup		3.8***		
		(0.22)		
Proportion of Fast Skips (5 sec), Warmup				4.4***
				(0.21)
Fit statistics				
Observations	657	657	657	657
Squared Correlation	0.07	0.50	0.02	0.59

Heteroskedasticity-robust standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Table 9. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Incentive Condition, by Age, Without Controls, pooled across Information Conditions

Outcome	**ATE**		**DIC**
	Below-35	Above-35	
Log Dwell Time	-0.31**(0.12)	-0.22*(0.09)	-0.09 (0.15)
Log Std Dev Dwell Time	-0.2†(0.11)	-0.39***(0.09)	0.19 (0.14)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

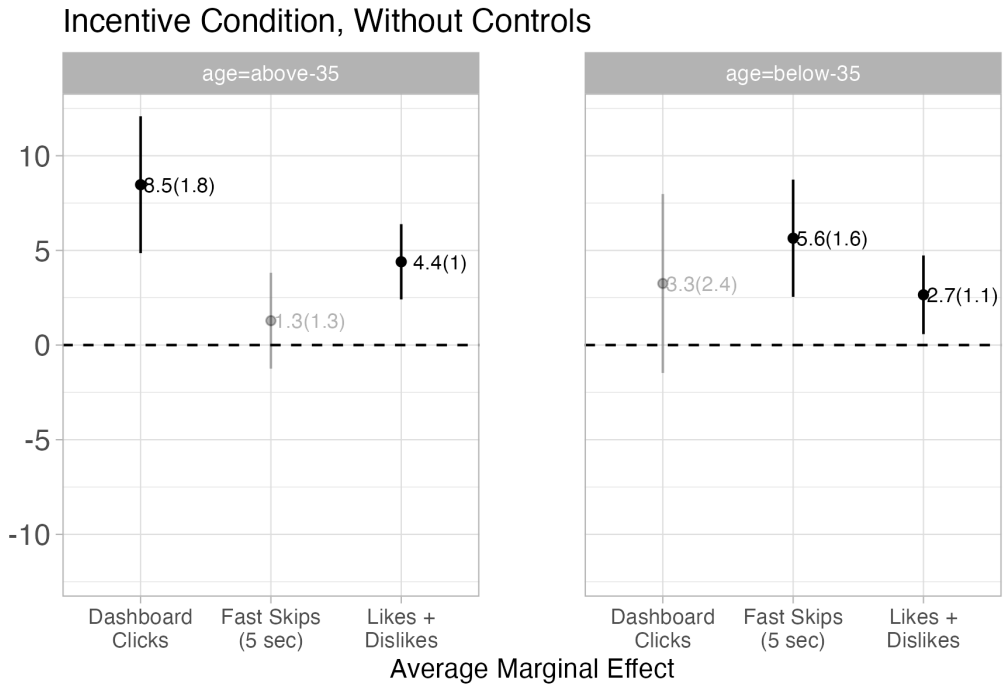


Fig. 7. Conditional Average Treatment Effects (CATEs) for the Incentive condition (i) Dashboard Clicks, (ii) Fast Skips, and (iii) Likes and Dislikes for the Incentive condition. Left: CATEs for participants above 35. Right: CATEs for participants below 35

Table 10. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Information Condition, by Age, With Controls, pooled across Incentive Conditions

Outcome	**Info Condition**	**ATE**		**DIC**
		Below-35	Above-35	
Log Dwell Time	Likes	0.02(0.1)	-0.06(0.07)	0.08 (0.12)
Log Dwell Time	Dwell	0.08(0.1)	0.08(0.07)	-0.01 (0.12)
Log Std Dev Dwell Time	Likes	-0.15(0.12)	-0.18†(0.09)	0.03 (0.15)
Log Std Dev Dwell Time	Dwell	0.15(0.12)	0.22*(0.09)	-0.07 (0.14)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

Table 11. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Incentive Condition, by Spotify Use, Without Controls, pooled across Information Conditions

Outcome	**ATE**		**DIC**
	Spotify Use=Rare	Spotify Use=Often	
Log Dwell Time	-0.17(0.11)	-0.31***(0.09)	0.14 (0.15)
Log Std Dev Dwell Time	-0.37**(0.12)	-0.31***(0.09)	-0.06 (0.15)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

Table 12. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Information Condition, by Spotify Use, With Controls, pooled across Incentive Conditions

Outcome	**Info Condition**	**ATE**		**DIC**
		Spotify Use=Rare	Spotify Use=Often	
Log Dwell Time	Likes	0.03(0.1)	-0.05(0.07)	0.08 (0.12)
Log Dwell Time	Dwell	0.13(0.09)	0.05(0.08)	0.08 (0.12)
Log Std Dev Dwell Time	Likes	0.04(0.13)	-0.24**(0.09)	0.28† (0.16)
Log Std Dev Dwell Time	Dwell	0.35**(0.12)	0.12(0.09)	0.22 (0.15)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

Table 13. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Incentive Condition, by Changed Interaction, Without Controls, pooled across Information Conditions

Outcome	**ATE**		**DIC**
	No	Yes	
Likes + Dislikes	4.61***(1.23)	2.84**(0.87)	1.78 (1.51)
Fast Skips (5 sec)	4.21*(1.68)	2.22†(1.22)	1.99 (2.08)
Dashboard Clicks	7.29**(2.31)	4.6*(1.9)	2.69 (2.99)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

Table 14. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Information Condition, by Changed Interaction, With Controls, pooled across Incentive Conditions

Outcome	**Info Condition**	**ATE**		**DIC**
		No	Yes	
Likes + Dislikes	Likes	2.27*(0.94)	3.68***(1.01)	-1.4 (1.4)
Likes + Dislikes	Dwell	-3.08**(1.01)	-2.66*(1.12)	-0.42 (1.49)
Fast Skips (5 sec)	Likes	-0.35(1.48)	-0.33(1.3)	-0.02 (1.97)
Fast Skips (5 sec)	Dwell	-1.46(1.34)	-0.55(1.23)	-0.9 (1.82)
Dashboard Clicks	Likes	3.74†(1.98)	5.19**(1.86)	-1.45 (2.71)
Dashboard Clicks	Dwell	-4.49*(2.14)	-5.06**(1.94)	0.57 (2.89)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

Table 15. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Incentive Condition, by Changed Interaction, Without Controls, pooled across Information Conditions

Outcome	**ATE**		**DIC**
	No	Yes	
Log Dwell Time	-0.32**(0.1)	-0.19†(0.1)	-0.13 (0.14)
Log Std Dev Dwell Time	-0.45***(0.1)	-0.2*(0.09)	-0.25† (0.14)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

Table 16. Conditional Average Treatment Effects (CATEs) and Difference-in-CATEs (DIC) of the Information Condition, by Changed Interaction, With Controls, pooled across Incentive Conditions

Outcome	**Info Condition**	**ATE**		**DIC**
		No	Yes	
Log Dwell Time	Likes	0.06(0.08)	-0.14(0.09)	0.21† (0.12)
Log Dwell Time	Dwell	0.1(0.08)	0(0.1)	0.1 (0.12)
Log Std Dev Dwell Time	Likes	-0.04(0.12)	-0.27**(0.11)	0.23 (0.16)
Log Std Dev Dwell Time	Dwell	0.14(0.1)	0.19†(0.1)	-0.05 (0.14)

[a] Heteroskedasticity Robust Standard Errors in Parentheses.
[b] Signif. Codes: ***: .001, **: .01, *: .05, †: .1

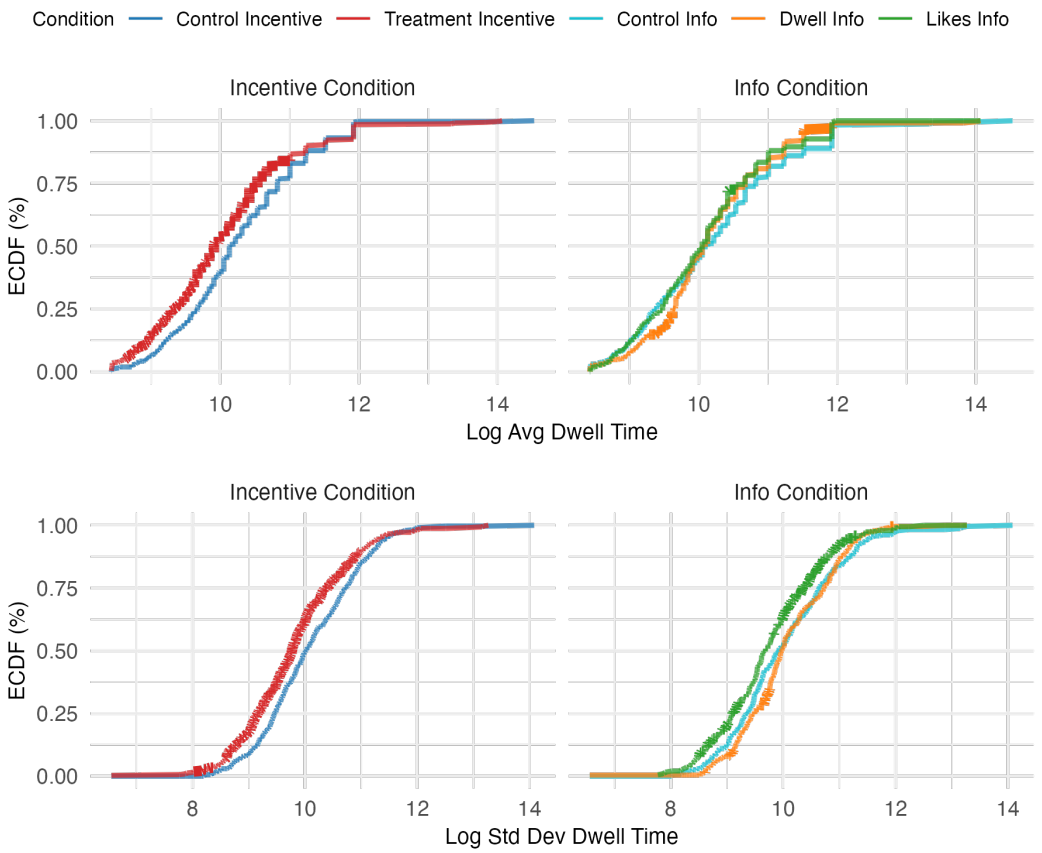


Fig. 8. ECDF for Dwell Time

B ADDITIONAL POST-EXPERIMENT SURVEY RESULTS

B.1 Post-Experiment Survey Questions

(1) *Did the way that you interacted with songs change across the three listening sessions?* Answers: (a) Definitely yes, (b) Probably yes, (c) Probably no, (d) Definitely no, (e) I don't know.
(2) *If yes, how did your interactions change across listening sessions? (CHECK ALL THAT APPLY)* Answers: (a) I changed how much I used the thumbs-up/down buttons, (b) I changed how much I used the skip button, (c) I changed how much I used the restart button, (d) I changed how long I spent on each song, (e) I'm not sure.
(3) *How do you think platforms like Spotify choose what to show you on your homepage? (CHECK ALL THAT APPLY)* Answers: (a) Based on what's most popular across the platform, (b) By randomly selecting songs you've recently listened to, (c) By analyzing what you've liked or skipped on the platform, (d) By randomly selecting songs that editors have picked, (e) Based on your age, gender, and location, (f) I don't know.
(4) *How do you think social media platforms like Facebook, Twitter, or TikTok choose what to show you? (CHECK ALL THAT APPLY)* Answers: (a) Based on what's currently trending across the platform, (b) By randomly selecting recent posts on the platform, (c) By analyzing what posts you've liked/commented on/etc., (d) By analyzing how long you watch videos and how you scroll down your feed, (e) By randomly selecting posts that editors at the platform pick, (f) Based on your age, gender, and location, (g) I don't know.
(5) *Do you ever try to "talk" to your algorithm or "hide" things from it? For example, do you ever give a song a "thumbs-up" just to Spotify that you want to see similar songs? Or do you sometimes avoid clicking on an advertisement just because you're worried about getting many similar advertisements in the future? If you do, tell us how and why.* Participants are permitted to provide open-ended, text answers to this question.
(6) *Are you concerned about data privacy online?* Answers: (a) Yes, I'm very concerned, (b) I'm sometimes concerned, (c) I'm rarely concerned, (d) No, I'm not concerned at all, (e) I don't know what data privacy is.
(7) *How often do you use music recommendation platforms, like Spotify?* Answers: (a) A few hours everyday, (b) A few hours each week, (c) A few hours each month, (d) Less than a few hours each month, (e) Never.
(8) *How old are you?* Answers: (a) 18-25, (b) 25-35, (c) 45-55, (d) 55+.
(9) *What is the highest level of education you have completed?* Answers: (a) Some high school or less, (b) High school diploma or GED, (c) Some college but no degree, (d) Associates or technical degre, (e) Bachelor's degree, (f) Graduate or professional degree, (g) Prefer not to say.
(10) *Any comments, questions, or feedback?* Participants are permitted to provide open-ended, text answers to this question.

The order of the answers is randomized for Questions 2-4.

B.2 Demographics

The plots below summarize the demographic groups represented by our study. There are more plots showing the demographic split across different treatment groups (i.e., verify whether our randomization was effective) in the "figures/post_experiment_plots" folder.

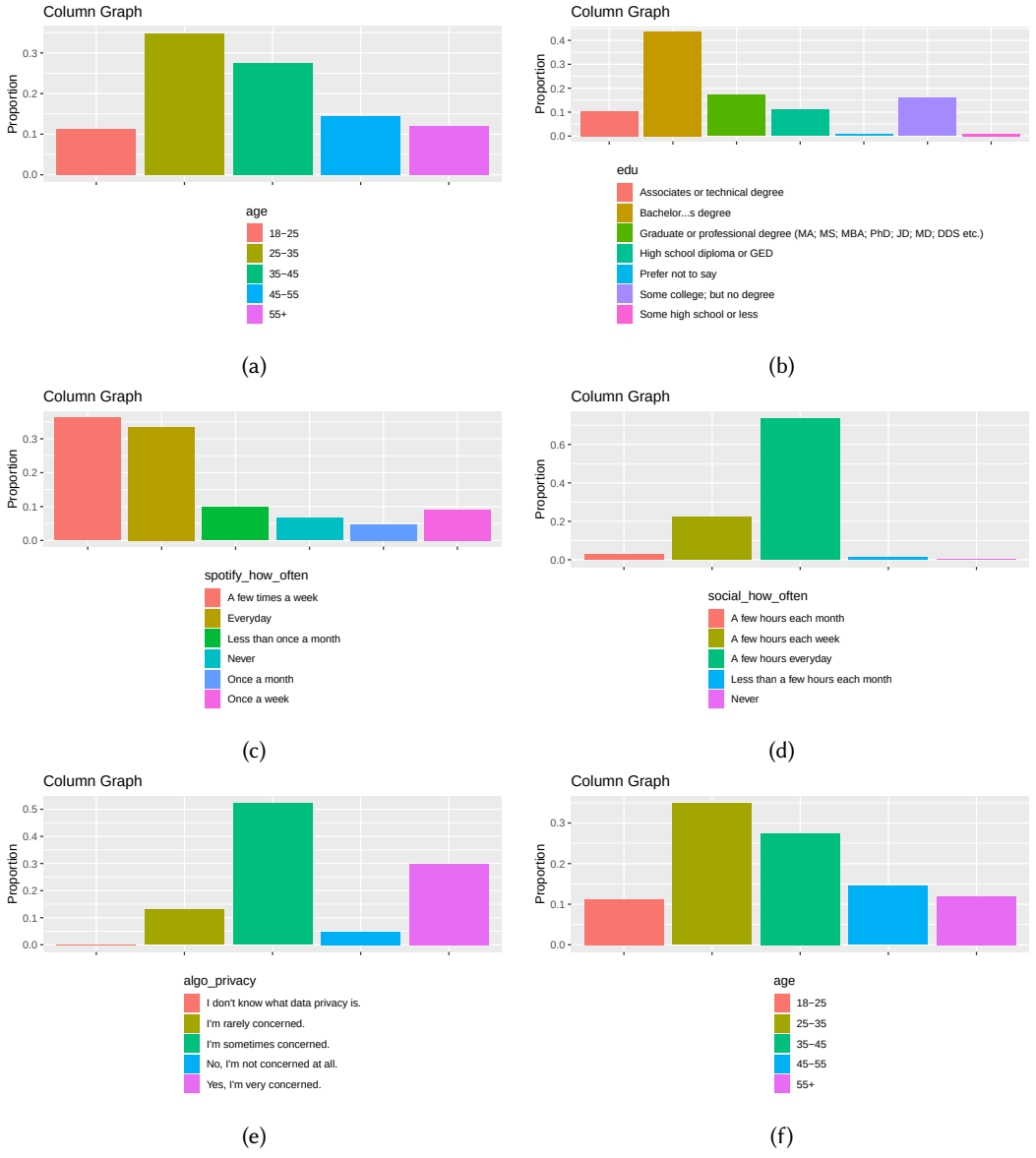


Fig. 9. Demographic responses

B.3 How participants believe online algorithms work

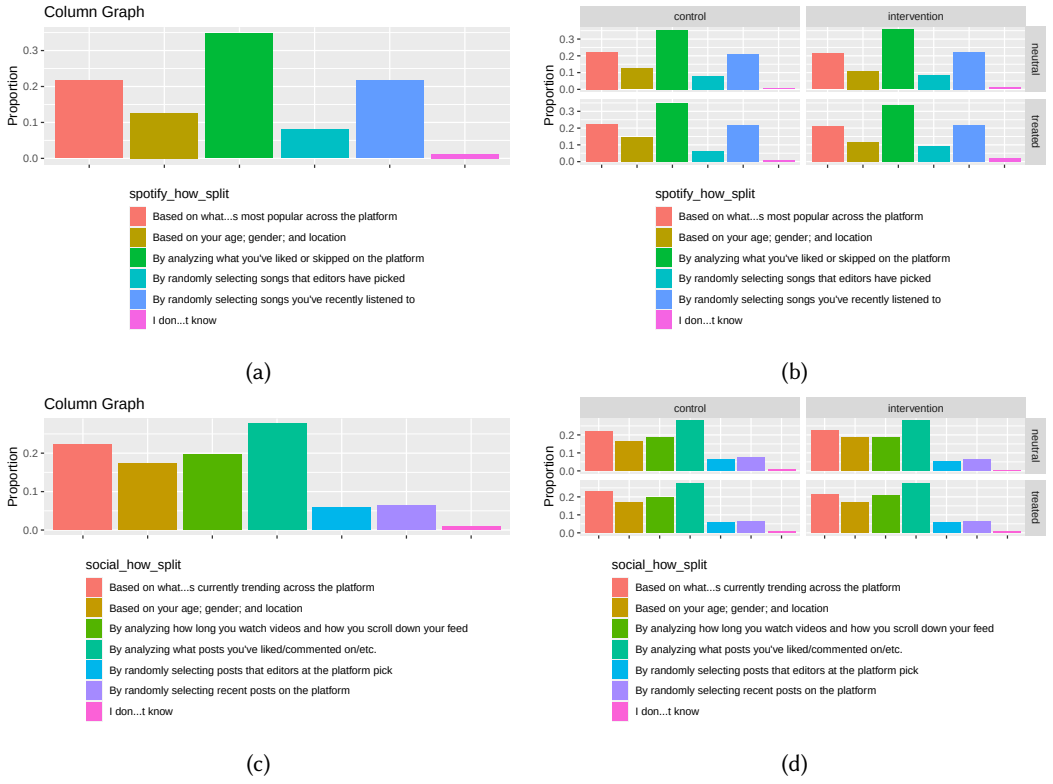


Fig. 10. How participants believe that Spotify (top) and social media algorithms (bottom) work.

B.4 Change in user behavior

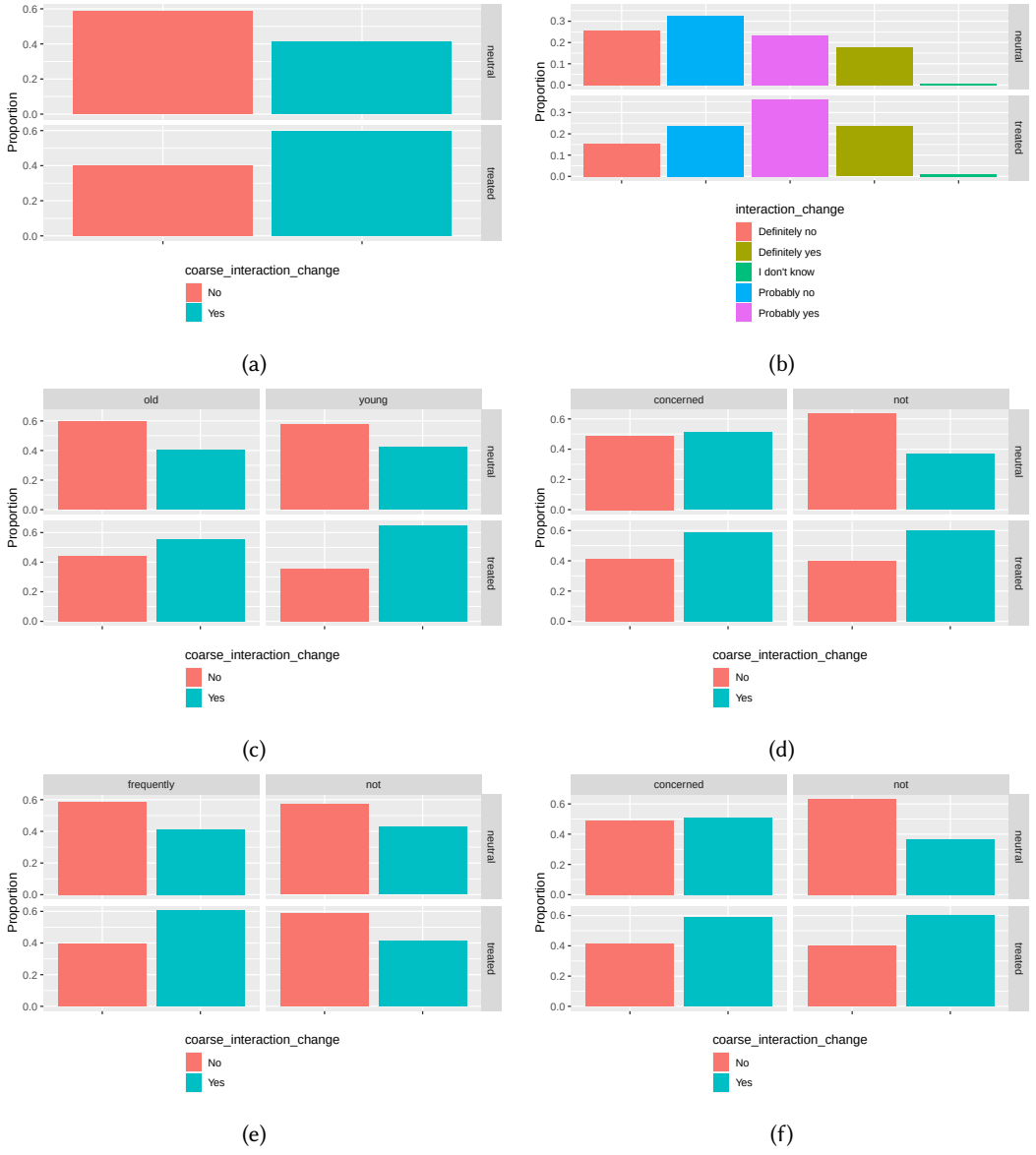


Fig. 11. Whether users changed their behavior across different splits.