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

Sarah H. Cen
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, shcen@mit.edu

Andrew Ilyas
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, ailyas@mit.edu

Hannah Li
Decision, Risk, and Operations Division, Columbia University, hannah.li@columbia.edu

Jennifer Allen
Sloan School of Management, Massachusetts Institute of Technology, jnallen@mit.edu

David G. Rand
Sloan School of Management, Massachusetts Institute of Technology, drand@mit.edu

Aleksander Madry
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, madry@mit.edu

Recommendation algorithms shape what users see on platforms ranging from search engines to social media to online marketplaces. It is typically assumed that users behave exogenously—that is, their response to a recommendation depends on that recommendation and *not* on the algorithm that generated it. In this work, we investigate whether users break this assumption; specifically, whether they *strategize*. For example, a user may not click on a video on YouTube not because they are uninterested, but because they believe YouTube's algorithm will overfit to the click. To determine whether users strategize, we conduct a lab experiment and survey. We test two hypotheses consistent with user strategization and find strong support for both. Our findings suggest that users strategize; users are able to strategize their explicit feedback (e.g., clicks) moreso than implicit feedback (e.g., dwell time). Our survey further reveals that a large majority of users admit to strategizing. In response to an open-ended question, we also gain insights into why and how users strategizing, surfacing intriguing behaviors (e.g., creating multiple accounts, leaving Spotify playing for days, and searching for content in private browsing mode).

1. Introduction

Recommendation platforms—like TikTok, Netflix, and Amazon—attract and retain users by sifitng through a large, often messy, set of options (e.g., videos, shows, and products) and presenting each user with suggestions that are tailored to their interests. In order to provide personalized suggestions, these platforms employ data-driven algorithms trained on past user behavior. For instance, Netflix pays close attention to the shows that users choose to watch and the ratings that users leave in order to algorithmically generate recommendations for each user.

Platforms generally 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 (Ricci et al. 2011). 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 (their engagement behavior) remains consistent across recommendation algorithms as long as their true preferences (their unknown utility function) remains the same.

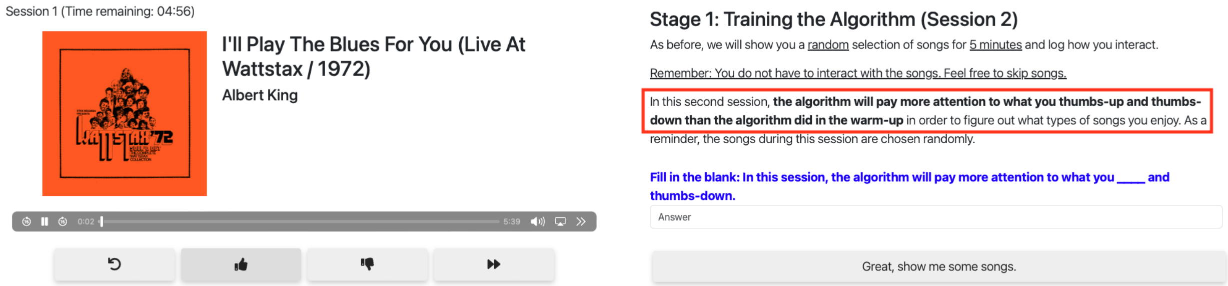
What this exogeneity assumption fails to capture is *strategic* behavior: that users may attempt to shape their future recommendations by adapting their revealed preferences to their recommendation algorithm, even if their true preferences do not change. For example, a TikTok user might "heart" a video 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 on recommendation algorithm design. Since recommendation algorithms are continually trained on user data, strategization can lead to unintended effects (such as feedback loops). 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 impact on the data that platforms gather, there has not been an experimental study to confirm user strategization in recommendation to the best of our knowledge. The goal of this work is to fill this gap. In order to do so, we conduct a survey and lab experiment that uncover user strategization and insights into why users strategize.

1.1. Our Contributions

In this work, we test for user strategization in recommendation. We begin with a formal definition of strategization in Section 2, which we adopt from Cen et al. (2023). Intuitively, 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, preferences*, which are unknown to the platform. Typically, it is assumed that users behave naively, e.g., play an action $B^*(Z) \in \arg\max U(Z, B)$ that maximizes their payoff under recommendation Z. This assumption is convenient because it implies


Session 1 (Time remaining: 04:56)
I'll Play The Blues For You (Live At Wattstax / 1972)
Albert King
0:02
5:39

Stage 1: Training the Algorithm (Session 2)

As before, we will show you a random selection of songs for 5 minutes and log how you interact.

Remember: You do not have to interact with the songs. Feel free to skip songs.

In this second session, **the algorithm will pay more attention to what you thumbs-up and thumbs-down than the algorithm did in the warm-up** in order to figure out what types of songs you enjoy. As a reminder, the songs during this session are chosen randomly.

Fill in the blank: In this session, the algorithm will pay more attention to what you ____ and thumbs-down.

Answer

Great, show me some songs.

(a) Our interface

(b) Example information condition

Figure 1 (a) On the left, a screenshot of the interface with which users interact. (b) On the right, an example description that users are shown at the start of their second listening session (i.e., the information condition).

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 π. A strategic user therefore chooses an action $B^*(Z,\pi)$ that maximizes their long-term payoff, as formalized in Section 2.[1] In other words, a strategic user's revealed preferences would be *algorithm-dependent*, which would complicate the platform's ability to estimate U.

Testing for user strategization is challenging. Not only are users' true preferences unknown, but users also have heterogeneous preferences (i.e., there is a different, hidden U for each user). As a result, there is no ground-truth information about each user's true preferences, which makes it difficult to determine how a user's observed behavior (their revealed preferences) reflect their underlying true preferences, which is at the core of differentiating whether users are strategic.

Despite these challenges, the definition provided by Cen et al. (2023) suggests that there are two hypotheses that we can use to test for strategization.

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

HYPOTHESIS 2 **(Incentive Hypothesis, informal)**. *User who are told that they will receive recommendations behave differently than users who are not told they will receive recommendations.*

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 actions now 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.

[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.

To test these hypotheses, 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). 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 1b). Furthermore, half of the participants are told that they will receive recommendations at the end of the study, and the other half are only told that their behavior is used to learn people's music preferences (but not that they will receive recommendations), as per Hypothesis 2 (see Figure 2).

We find 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 Hypotheses 1 and 2. We further survey participants at the end of the study to understand whether users strategize "in the wild" and whether they do so intentionally. The participant responses reveal that users are highly aware of their online algorithms. In addition, while a sizeable portion do not consciously strategize—stating that they do not wish to manipulate their algorithm—many others admit to "gaming" their algorithm. Several users confess that they often perform actions in "Incognito" mode (e.g., search for a link in a private browsing window) in order to hide certain interactions from their platform. Others intentionally obfuscate their algorithm (e.g., by clicking randomly) to prevent the algorithm from overfitting to spurious interests.

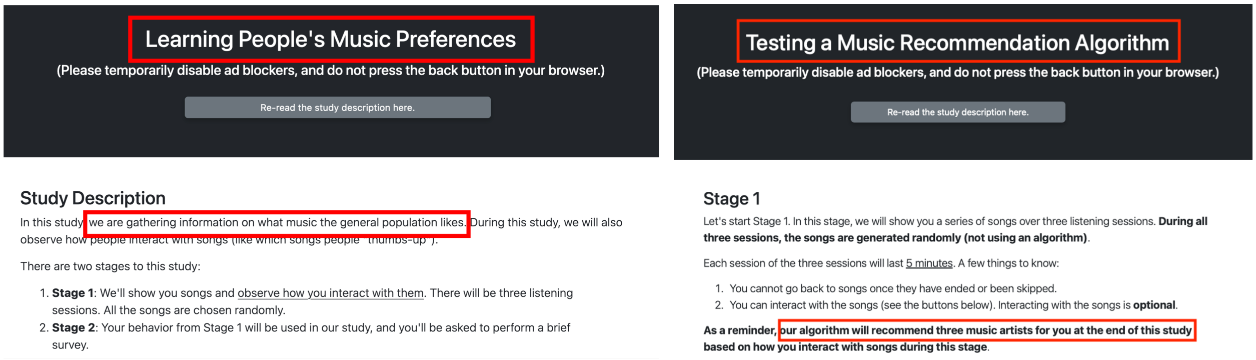
These findings provide a first step in documenting user strategization. Ultimately, user strategization can hurt platforms, as their observations of user behavior in response to recommendations would not depend on the recommendations alone, but also on the algorithm that generated the recommendations. Recommendation platforms would therefore need to be mindful of this effect (e.g., when using data to train a different algorithm).

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

This work tests for user strategization on data-driven recommender systems. In this section, we briefly highlight a few lines of related work both on the theoretical front and the practical front.

On the theoretical side, our work is most closely related to that of Cen et al. (2023), which characterizes user strategization on data-driven platforms as "long-term planning" (see Section 2.1).



Learning People's Music Preferences

(Please temporarily disable ad blockers, and do not press the back button in your browser.)

Study Description

In this study, we are gathering information on what music the general population likes. During this study, we will also observe how people interact with songs (like which songs people "thumbs-up").

There are two stages to this study:

1. **Stage 1**: We'll show you songs and observe how you interact with them. There will be three listening sessions. All the songs are chosen randomly.
2. **Stage 2**: Your behavior from Stage 1 will be used in our study, and you'll be asked to perform a brief survey.

(a) Low-Incentive Condition

Testing a Music Recommendation Algorithm

(Please temporarily disable ad blockers, and do not press the back button in your browser.)

Stage 1

Let's start Stage 1. In this stage, we will show you a series of songs over three listening sessions. **During all three sessions, the songs are generated randomly (not using an algorithm)**.

Each session of the three sessions will last 5 minutes. A few things to know:

1. You cannot go back to songs once they have ended or been skipped.
2. You can interact with the songs (see the buttons below). Interacting with the songs is **optional**.

As a reminder, our algorithm will recommend three music artists for you at the end of this study based on how you interact with songs during this stage.

(b) High-Incentive Condition

Figure 2 (a) On the left, screenshot of the Low-Incentive study description. (b) On the right, screenshot of the High-Incentive description at Stage 1.

Our work can be viewed as providing empirical evidence for this notion of user strategization. There are also other models of user strategization on recommender systems (Haupt et al. 2023, Cen et al. 2022) that focus on slightly different aspects of strategization and are thus not as directly testable by our work. A related model is that of Kleinberg et al. (2022), which concerns users who have *inconsistent preferences*. Inconsistent preferences and strategization are intuitively similar (e.g., Kleinberg et al. (2022) also propose "myopic" and "long-term" users, though in a slightly different sense), and lead to many of the same implications. However, inconsistent preferences neither imply nor follow from the hypotheses tested in this work. (In particular, a user with consistent preferences can still strategize, and a non-strategic user can have inconsistent preferences).

There are also several more general models of multi-agent interaction under which the same kind of endogeneity induced by user strategization can arise. In the Economics literature, for example, there are models of equilibria arising from endogenous learning (Esponda and Pouzo 2016, Heidhues et al. 2021, Fudenberg et al. 2021). In the Computer Science literature, the most closely related models are that of performative prediction (Perdomo et al. 2020), strategic classification (Hardt et al. 2016), and alternating games (Roth et al. 2010). Exploring the implications of our findings through the lens of these models is left to future work.

On the practical side, our work adds to a mounting body of evidence that users are *aware* of their recommendation algorithms and *behave* accordingly. Several distinct lines of work have studied user awareness of recommendation algorithms—for example, algorithmic awareness has been studied in the context of digital literacy (Newman et al. 2018, Sirlin et al. 2021, DeVito 2021), human-computer interaction (DeVito et al. 2018, Taylor and Choi 2022), and beyond. More recently, there has been a growing qualitative account of users "taming" or "training" their social media algorithms—that is, explicitly changing their behavior to get more desirable outcomes. For example, (Narayanan 2023b,a) explain the mechanism behind recommendation systems like TikTok and provides explicit guidelines on how to induce a better feed. In a similar vein, a rapidly-growing line of work has documented users adapting to what they increasingly recognize as data-driven social media algorithms (Petre et al. 2019, Shin 2020, Lee et al. 2022, Simpson et al. 2022).

Finally, our work is related to studies of strategization in other domains. For example, strategization has been long-documented in online auctions (Edelman and Ostrovsky 2007), and has more recently been studied in the context of freelancing (Rahman 2021) and on ride-share platforms such as Uber (Marshall 2020).

2. Hypotheses

Before describing our methodology, we begin by defining user strategization, adopting the formalization given by Cen et al. (2023). While it is difficult to directly test for user strategization, as discussed in Section 1.1, this definition suggests two hypotheses that, if confirmed, are strong indications of strategization. We propose an experiment for testing these hypotheses in Section 3.

2.1. Definition of User Strategization

We begin by describing the model proposed by Cen et al. (2023) in which we situate our hypotheses. Cen et al. (2023) model the interactions between a user and their recommendation platform as a repeated, two-player game. At each time step t, the platform plays a *recommendation* $Z_t \in \mathcal{Z}$, and the user responds with *behavior* $B_t \in \mathcal{B}$ (e.g., a click or "like"). At the end of each time step, the user and platform collect payoffs. In this work, we are only concerned with the user's payoff function denoted by U, which maps an interaction (Z_t, B_t) to a real-valued payoff.

To formalize user strategization, this model must capture how users *adapt* under their understanding of a recommendation algorithm. Cen et al. (2023) do so using a model similar to a *Stackelberg game*. In a Stackelberg game, one player (the platform) goes first and declares their game-play strategy (the recommendation algorithm). The second player (the user) utilizes this newfound information to select their own strategy (how they reveal their true preferences U). Translating this setup to our setting, the platform first declares a *recommendation algorithm* π which maps "histories" $\mathcal{H}_t = \{(Z_1, B_1), \ldots, (Z_t, B_t)\}$ to new recommendation Z_{t+1}. The algorithm π captures how the platform uses past user behavior to provide personalized recommendations. Next, the user chooses how they will respond to recommendations. We denote the user's chosen strategy by $\rho^{\pi,U}$, as it may depend on the platform's algorithm π and the user's true preferences U. Formally, game play proceeds as follows:

1. The platform declares their recommendation algorithm π.
2. The user sets their strategy $\rho^{\pi,U}$.

3. At every time step t,
   (i) The platform applies algorithm π to the history $\mathcal{H}_t$ to generate recommendation Z_t.
   (ii) The user responds with behavior B_t which is drawn from the distribution $\rho^{\pi,U}(\,\cdot\,; Z_t)$.
   (iii) The user collects payoff $U(Z_t, B_t)$.

(Note that, in practice, users have a partial, often erroneous understanding of their platform's algorithm. However, for ease of analysis, Cen et al. (2023) assume the complete-information setting, and we explain in Section 3 how our experiment accommodates this assumption.)

Naive and strategic behavior. Within this model, Cen et al. (2023) define two types of user behavior. If a user behaves *naively*, then they do not plan ahead, i.e., do not anticipate how their actions affect future recommendations. Specifically, they maximize their immediate payoff by choosing the strategy ρ^U_{naive}, where

$$\rho^U_{\text{naive}}(Z) \in \arg\max_B \; U(Z, B). \tag{1}$$

In the context of recommendation, naive behavior correspond to the exogenous preference assumption: the user's chosen actions depend only on the content Z they are shown, but *not* on the platform's declared strategy π.

On the other hand, if a user behaves *strategically*, they plan ahead. The user recognizes that their behavior at time t influences the platform's future recommendations because each interaction (Z_t, B_t) helps to train the recommendation algorithm π at future time steps. They consider how their current action might affect future recommendations and choose the strategy $\rho^{\pi,U}_{\text{strat}}$ that maximizes their *long-term utility* at $t \to \infty$:

$$\rho^{\pi,U}_{\text{strat}} \in \arg\max_{\rho\in\Delta(\mathcal{B})} \lim_{t\to\infty} \mathbb{E}_{\mathcal{H}_t\sim\pi\times\rho}[U(Z_t, B_t)], \tag{2}$$

where $\Delta(\mathcal{B})$ denotes the probability simplex over possible user behaviors, and $\mathcal{H}_t \sim \pi \times \rho$ is a shorthand for the distribution of histories that would result from game play under recommendation algorithm π and user strategy ρ. For details, we refer the interested reader to (Cen et al. 2023).

As an example, suppose that a certain video recommendation seems interesting to the user such that the user's payoff if they click on the recommendation is higher than that if they do not. Then, behaving naively would lead the user to click on the recommendation. Suppose further that the user generally dislikes videos that are made by the video's creator, but the platform's declared algorithm π primarily recommends content from creators that users have previously interacted with. Then, a strategic user may anticipate that clicking on the video will cause the platform to over-recommend bad content in the long term, and ultimately choose to ignore the video.

The main goal of our work is to determine whether users are strategic in this way, i.e., whether they behave according to Eq. (2) rather than Eq. (1). Because we lack access to the user's payoff function U, we can't test Eq. (2) directly—we opt instead to test the *implications of strategization*.

2.2. Hypotheses

We formulate two hypotheses based on this definition. The first is inspired by the fact, under the model above, behaviors B_t are independent of the platform's strategy π and instead depend only on the recommendation Z_t if a user is naive. On the other hand, if a user is strategic, they will use their knowledge of π in order to pick their behaviors B_t. Thus, if users are strategic, changing their *perception* of π should change their behavior, as captured below.

Hypothesis 1 **(Information Condition)**. *Holding all else constant, providing users with different descriptions of how users preferences are learned (i.e., of algorithm π) changes the way they behave. Formally, there exist algorithms π_1 and π_2 as well as recommendation $Z \in \mathcal{Z}$ such that a user's response to recommendation Z if the platform declares algorithm π_1 in Step 1 of Section 2.1 is different from the user's response to the same Z if the platform declares algorithm π_2.*

Recalling Eqs. (1) and (2), if a user behaves naively, then their actions cannot depend on π. On the other hand, if a user behaves strategically, then their actions are algorithm-dependent. Therefore, confirming Hypothesis 1 would indicate that the user adapts their behavior to the algorithm, as would hold true under user strategization.

On its own, confirming Hypothesis 1 indicates that users adapt to their algorithms, but not necessarily that they strategize under our definition in Section 2.1. In particular, it does not indicate that users adapt to elicit better recommendations.[2] This additional component is captured by our second hypothesis. In particular, the hypothesis below posits that users behave differently when told that their actions will be used to provide personalized recommendations for the user than when told that their actions will be used to learn preferences (in a general sense).

Hypothesis 2 **(Incentive Condition)**. *A user's behavior when they are told that their actions are used to provide personalized recommendations should differ from that if they are not told their actions are used to provide personalized recommendations. Formally, there exists a recommendation $Z \in \mathcal{Z}$ such that, for a fixed π, the user's response to Z under the knowledge that π affects the user's future payoffs via future recommendations is different from their response to Z if they do not believe the algorithm affects their future payoffs.*

[2] It is possible, for instance, that U depends on π, under which Hypothesis 1 would hold true even if users behave "naively," but Hypothesis 2 would not.

Hypothesis 2 provides clarity on Hypothesis 1. That is, a user may adapt, but it is unclear for the motivation behind adaptation. Hypothesis 2 posits that adaptation is temporally-motivated: it is built on the understanding that current behaviors affect future payoffs, as given in Eq. (2). Together, these two hypothesis capture the algorithm-dependence and long-term nature of strategization. Evidence that they hold would strongly suggest that users strategize in a way similar to the formalization given by Cen et al. (2023).

3. Methodology

In this section, we describe our experimental methodology and analysis. In short, we build a basic music recommendation platform on which users can listen to and interact with songs, as they similarly would on Spotify or Pandora. We execute 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 user'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. In other words, we sought to answer the questions: Do different Information and Incentive conditions affect user 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 users 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 47 participants who ran into technical issues. Of the remaining participants, 48 failed at least two audio-visual attention checks or written attention checks. Finally, another 12 participants had metrics (likes, dislikes, skips, and dwell time) that were more than four standard deviations away from the average. The final sample has 643 participants, of which 48% are female, 52% are male; 46% are 35 years-old and below, and 70% use Spotify or music recommendations platforms at least a few times a week.

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 XX songs. The songs presented in the listening sessions are random for all participants and therefore do *not* depend on the users' behavior. Users are randomly exposed to different treatment conditions during their two listening sessions, 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 1. The second factor tests, which we refer to as the "Incentive condition," tests Hypothesis 2. 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.

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

- **Incentive control (low):** 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.
- **Incentive treatment (high)**: 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 2. Note that participants under the Incentive control 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 2.

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. 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*:

- **Information control:** 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" condition:** 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" condition:** 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 users (all songs during the listening sessions are generated randomly), only the information that participants receive. In this way, the "actual" algorithm used to learn user preferences is irrelevant, as we control the users' perception of the algorithm using the Information condition and are primarily interested in the treatment effects of these conditions (i.e., the difference in outcomes between an Information treatment and Information control) rather than the raw outcomes.

3.4. Summary of Procedure

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 (i.e., users in the Incentive control are told they are participating in a general-interest survey, and users in the Incentive treatment 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 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. They then interact with the music player for five minutes. Next, every participant undergoes a test listening session: each participant is shown their assigned Information condition, after which they interact with the music player for five minutes.[3]

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 Section A.1. In addition to demographic information, we ask participants several multiple-choice/checkbox questions to query: (1) whether they changed the way they interacted

[3] 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.

across sessions and, if so, how; (2) general questions about 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

In our analysis, we examine the data collected from our experimental procedure for signs of strategization. To test Hypotheses 1 and 2 from Section 2, we look at *average treatment effects* across Information conditions and across Incentive conditions. In particular, we examine whether participants' feedback in the test (i.e., second) listening session differs across our six experimental conditions. We focus in particular on the use of (dis)likes and the dwell time length across experimental conditions.

3.6.1. Likes, dislikes, and skips. First, we investigate whether our treatment conditions affect the users' explicit feedback, where we define explicit feedback to be the total number of users "likes" and "dislikes" in the test listening session. To do so, we run a quasi-Poisson count regression with the total number of likes and dislikes as the outcome variable, dummy variables for the (i) Incentive condition, (ii) Information condition, and (iii) their interaction as treatment variables, and heteroskedasticity-robust standard errors. Formally, let $Y^{(\text{dis})\text{likes}}(i)$ denote the total number of (dis)likes for participant i (our outcome variable) and

$$\begin{aligned}
\textsc{Incen}(i) &= \text{Ind}(\text{participant } i \text{ is assigned the Incentive treatment}),\\
\textsc{Likes}(i) &= \text{Ind}(\text{participant } i \text{ is assigned the likes Information condition}),\\
\textsc{Dwell}(i) &= \text{Ind}(\text{participant } i \text{ is assigned the dwell Information condition}),
\end{aligned}$$

denote the dummy variables. Then, we fit a quasi-Poisson count regression to:

$$\begin{aligned}
Y^{(\text{dis})\text{likes}}(i) \sim \beta_0 &+ \beta_1 \textsc{Incen}(i) + \beta_2 \textsc{Likes}(i) + \beta_3 \textsc{Dwell}(i)\\
&+ \sum_{j\in[2],k\in[3]} \gamma_{j,k} \cdot \text{Ind}(i \text{ is given Incentive condition } j \text{ and Information condition } k),
\end{aligned}$$

where we enumerate the Incentive and Information conditions in the order given in Section 3.3. We use a Poisson quasi-maximum-likelihood model in order to account for potential overdispersion in the engagement data (Wooldridge 1999). For both the Incentive condition and the Information conditions, we report the Average Marginal Effect of each experimental condition compared to the respective control condition.

To assess whether any increase in engagement is due to users increasing the overall number of songs they listen to, we run an additional quasi-Poisson model with the same specification as above, except with the total number of songs listened to as the outcome variable instead of the total number of likes and dislikes. To assess whether any increase in engagement is due to an increased *rate* of engagement—that is, users liking or disliking a higher proportion of songs—we use a quasi-binomial generalized linear model (GLM) with the number of songs as the number of "trials" and the number of likes and dislikes as the number of "successes." We use a quasi-maximum-likelihood model to account for potential overdispersion in the data. As before, we include treatment dummies for the (i) Incentive condition, (ii) Information condition, and (iii) their interactions, with heteroskedascity-robust standard errors.

Then, we examine whether our treatment condition affects the average number of times the user "Early skips" a song, where a "Early skip" is defined as the user moving to the next song within 5 seconds of the song starting. This definition is for analysis purposes only; we do not provide participants with a definition of a "Early skip." As with our specification for likes and dislikes, we run a quasi-Poisson count regression with the total skips as the outcome variable, dummy variables for the (i) Incentive condition, (ii) Information condition, and (iii) their interaction as treatment variables, and heteroskedasticity-robust standard errors.

3.6.2. Dwell time. Additionally, we examine whether the our treatment conditions affect the amount of time that users spend listening to each song, measured as the users' average log dwell time per song. We run an ordinary least squares (OLS) model with average log dwell time as the outcome variable, (i) Incentive condition, (ii) Information condition, and (iii) their interaction as treatment variables, and heteroskedasticity-robust standard errors.

For intepretability, for our quasi-Poisson count models, we report the Average Marginal Effect (AME) of (i) Incentive and (ii) Information treatments, pooled across conditions, on our outcome variables (calculated using the `marginaleffects` package in R, by Arel-Bundock (2023)) in addition to providing the coefficients from the quasi-Poisson model in Table format. In particular, although the coefficients of the quasi-Poisson model do not have a straightforward interpretation, the models allow us to test whether there are significant interactions between the Incentive and Information conditions.

3.6.3. Additional details. Unless otherwise specified, we report the results from our models without additional controls variables. However, in addition to the baseline specification, we also test an additional specification that includes the participants' behavior in the warm-up session as a control. For example, for a model with the total number of likes and dislikes in the test session as the outcome variable, we add the total number of likes and dislikes in the warm-up session as a control variable. This specification allows us to better control for individual-level differences in engagement behavior. The full results are reported in Appendix B.2.

In a deviation from our pre-registration, we also report the results examining the effect of the Incentive condition on user explicit feedback in the warm-up condition. The reason for this deviation is because participants were randomized into the Incentive condition before starting the warm-up session, so analyzing the warm-up session further demonstrates the effects of the Incentive condition on user behavior. These analyses provide additional support for our hypothesized effect; they do not change the overall interpretation of our results.

3.6.4. Post-experiment survey. For the post-experiment survey, we compute simple statistics across answers. For demographic information, we compute the proportion of participants in each category. We similarly compute the proportion of participants who report changing their behavior across the warm-up and test sessions, grouped by experiment conditions. Finally, we manually analyze all open-ended responses in order to determine whether users self-report strategizing "in the wild" (i.e., outside of our lab experiment) and their reasons for doing so.

4. Results

Our results provide strong evidence that user strategize when interacting with recommender algorithms. We find that minor variations in the (i) the incentives users have to engage and (ii) information that users receive about the recommender system cause significant downstream differences in their engagement patterns.

4.1. How information and incentives affect (dis)liking behavior

We first examine whether users change their patterns of liking and disliking under different Incentive and Information conditions. Figure 3 shows the average number of likes and dislikes across conditions, shown with 95% confidence-intervals. Results from the models specified in Section 3 are shown in Table 2 and reported below.

First, we find that increasing the users' incentives for improving their long-term recommendations substantially increases the amount they engage on the platform. We find that users in the High-Incentive condition, who were told that the algorithm was learning their preferences and they would have to listen and rate three of the algorithm's recommendations, engaged much more compared to users in the Low-Incentive condition, who were told the algorithm was trying to learn the general public's preferences. Pooling across all "Information" conditions, we find that participants in the High-Incentive condition provided 2.71 (SE: 0.694, $p < .001$) more likes and dislikes than those in the Low-Incentive condition in the Warmup session, and 4.32 (SE: .799, $p < .001$) more likes and dislikes in the Test session. That is, while High-Incentive participants engaged at higher levels in the

Warmup session, the effect is even more pronounced in the Test session when participants are told the algorithm is tracking their behavior in order to learn their preferences. This pattern is consistent with our hypothesized mechanism of strategization; participants engage in more strategic behavior when they believe their short-term engagement can improve their long-term recommendations. The magnitude of this effect is substantial – in the Test session, participants in the High-Incentive condition submit 54% more likes and dislikes than those in the Low-Incentive condition.

Next, we turn to analyze the Information intervention. For this analysis, we only examine behavior in the Test session, during which participants were given different information about which behaviors the algorithm was tracking. We pool results across Incentive conditions. We find that that informing participants that the algorithm tracks certain behaviors induces participants to engage in those behaviors at higher rates. Participants in the "Likes" condition, who were told the algorithm was tracking their likes and dislikes, produced 1.85 (SE: 1.06, $p = .08$) more likes compared to the Control condition, in which we gave no information about how the algorithm worked. While this difference is not significant in the baseline model, when we add the participant's number of likes and dislikes in the Warmup session as a control variable to our specification (thus controlling for existing individual-level heterogeneity in engagement behavior), we see that the Likes condition actually increases the number of additional likes and dislikes in the Test session by 3.5 (SE: .79, p <.001).

Conversely, informing participants that the algorithm was tracking the time spent on each song (the Duration condition) – and thus, implicitly suggesting that the algorithm was *not* tracking like and dislike behavior – *decreased* the average number of (dis)likes by 3.4($SE = .95$, $p < 0.001$) compared to Control (or 2.86 ($SE = .67$, $p < .001$) in a model with additional controls.) Again, the magnitude of these effects are large. Participants in the Likes condition engage 80% more than those in the Duration condition.

Interestingly, our model (shown in Table 2) finds no significant interaction between our Incentive and Information conditions – we find only significant *main* effects for both of these conditions. Nonetheless, the cumulative effects of the conditions are striking. Participants in the High-Incentive, Likes condition submitted 12.0 (95% CI: [10.7,13.5]) likes and dislikes compared to the 4.6 (95% CI: [3.8, 5.6]) likes and dislikes submitted in the Low-Incentive, Duration condition – a 260% difference. These results suggest that a large portion of the users' explicit engagement behavior can be attributed to strategization.

A natural follow-up question is whether these increased levels of total engagement are due to users engaging at higher rates (i.e. increased liking or disliking a song, conditional on having listened to it) or users listening to more songs during the 5 minute session, but liking songs at the same rate. That is, are users spurred to express their preferences in a strategic way to the algorithm, or

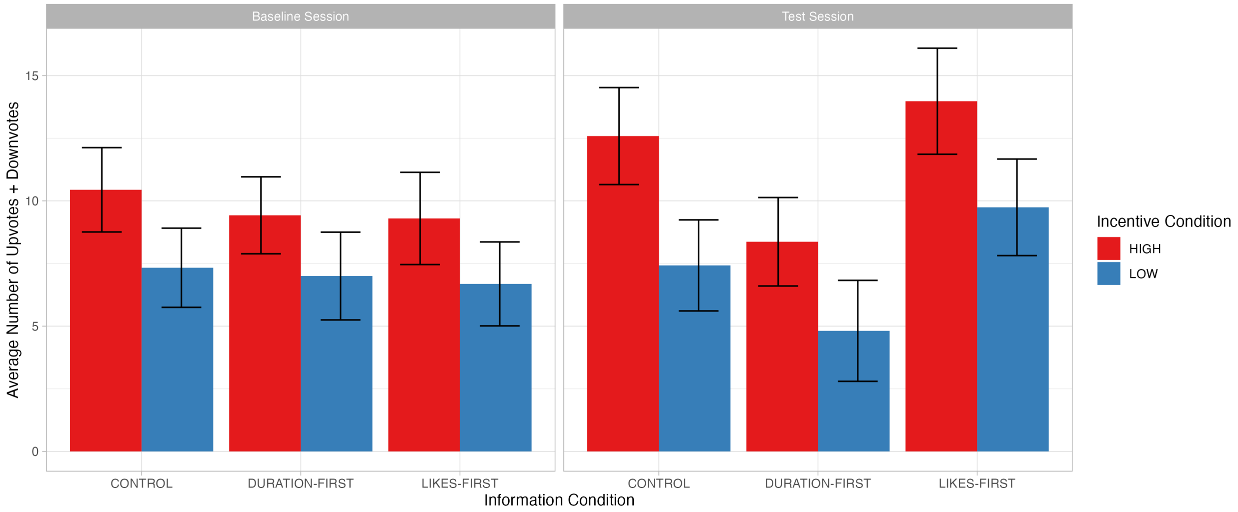


Figure 3 Average Number of Likes + Dislikes, across Incentive and Information Conditions. Means are shown with 95% confidence intervals.

strategically searching more for songs on which to provide feedback? Our analysis suggests that both of these mechanisms play a role, although the effects differ by intervention.

Regarding an increase in the rate of engagement, we find that the High-Incentive condition increases the proportion of engagement per song by 8.0 percentage points (SE: .03, $p = .018$) the Like condition increase proportion of engagement per song by 10.8 percentage points (SE: .04, $p = .007$). Conversely, we find that the Duration condition significantly decreases the rate of engagement by 14.1 percentage points (SE: .04, $p < .001$). As shown in Table 4, we find no significant interaction between conditions.

Additionally, we find evidence that the High-Incentive condition also increases the *number* of songs that the participants listen to. Participants in the High-Incentive condition listen to 5.4 (SE: 1.19, $p < .001$) more songs than participants in the Low-Incentive condition. That is, the High-Incentive condition not only induces participants to reveal their preferences to the algorithm through explicitly providing feedback, it also spurs users to explore a greater number of songs. We do not find a significant differences in on the overall number of songs listened to in either Information condition, or a significant interaction (see Table 2). Additional figures and analyses of exploring these results can be found in Appendix B.1.

4.2. How information and incentives affect dwell-time

We next examine whether the users' song dwell-time under different Information and Incentive conditions, since the time spent listening to songs is an important form feedback used by music recommender systems Mehrotra et al. (2020, 2019). For simplicity, we only examine users' behavior in the Test session.

Q: How do social media platforms decide what to show you?	**%**
Based on what's currently trending across the platform	22
Based on your age, gender, and location	17
By analyzing how long you watch videos and how you scroll down your feed	20
By analyzing what posts you've liked/commented on/etc.	28
By randomly selecting posts that editors at the platform pick	6
By randomly selecting recent posts on the platform	6
I don't know	1

Table 1 Post-experiment user survey responses. Users were asked how they believe social media algorithms select the content they see on their feeds or homepages.

We find that increasing users' incentives for engagement changes the amount of time they spend listening to songs. In particular, we find that the High-Incentive condition significantly increased the number of times a user skipped a song – users in the High-Incentive condition skipped 3.29 (SE: .987, $p < .001$) more songs than users in the Low-Incentive condition. Similarly, users in the High-Incentive condition spend less time listening to each song than users in the Low-Incentive condition. Users in the High-Incentive condition listen to songs for approximately 24% less time than users in the Low-Incentive condition ($\beta = -.28$, SE: .07, $p < .001$).

We do not find that either of the Information conditions alter the number of skips, nor do we find a significant interaction between the Incentive and Information conditions, as can be seen in Tables 2 and 3.

4.3. Post-Experiment Survey

Of the High-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 Low-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 the Appendix.

We manually analyzed the open-ended responses, in which we asked users whether they strategize on their own recommendation platforms. We found that around 20 percent of users reported definitive strategization, 42 percent reported not strategizing, and 38 percent provided information for which it was unclear. Finally, we analyzed why and how users strategize. We identified several trends that persist across users and include example responses below.

Being pigeonholed by algorithms: Some users expressed 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 keep 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: "Yeah, totally. For instance during 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 indicated 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 found that several users 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 confessed 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 users confessed 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

In order to test for user strategization, we designed a lab experiment to test for Hypotheses 1 and 2. Our results verify that both hypotheses, although there are some subtleties. In particular, we find strong support for Hypotheses 1 and 2 in terms of the participants’ use of likes and dislikes, i.e., strong support that users strategize their explicit feedback. The evidence is less definitive for implicit feedback mechanisms, such as dwell time. That is, we find trends in how users strategize their dwell time, but we treatment effects are less significant. In combination, these results suggest that users can more easily strategize explicit feedback than implicit feedback. It is unclear, however, whether this observation motivates greater use of implicit feedback mechanisms by recommendation platforms. Although it may be more difficult for users to strategize their dwell time, dwell time is also a noisier feedback mechanism than likes and dislikes.

In testing Hypotheses 1 and 2, we surface evidence that users try to shape the content that they see. 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, we show that (at least some) users also play an active role in shaping what they see. We therefore provide a first step into documenting and measuring strategization through an online lab experiment. We find strong evidence of strategization, where users change their behaviors in response to their beliefs about how the recommendation algorithm. In particular, we find that users change their behavior in response to a) information about what types of feedback (likes or length of time listening) the algorithm pays attention to and b) information on whether their behavior will affect future recommendations.

This strategization implies that the behavior of users in response to these algorithms is not exogenous to the recommendation system. In other words, how the user interacts with their content (e.g., the probability of clicking on a post) does not simply depend on the piece of content itself—it also depends on the algorithm that recommended the content. This behavior creates feedback loops in the system, since the user's behavior depends on the algorithm, and the algorithm then learns from this behavior to provide future recommendations. Because users' behavior is now endogenous, strategization can hurt the platform's ability to repurpose data gathered under one algorithm for inference about a different algorithm, as discussed in Cen et al. (2023).

Below we outline three directions for future work: identifying the mechanism that leads users to strategize, measuring strategization on an existing platform, and better design of future algorithms.

The mechanism of why people strategize is still unknown. Do users want a more heterogeneous set of recommendations? Does strategization arise because users have inconsistent preferences Kleinberg et al. (2022)? For example, users may enjoy a piece of content but know that the content is not good for them and actively try to reduce how often see the content.

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 more challenging and likely requires different methods than an online lab experiment. Note that we were able to measure differences in users' behavior as a response to different beliefs about the recommendation algorithm, precisely because we were able to give information to influence their beliefs. On a real platform, however, users' beliefs are generally shaped through their previous interactions with the algorithm. When users interact with the algorithm, they observe their recommendations in subsequent time steps, and infer (correctly or incorrectly) how their interactions shaped the subsequent recommendations. Thus it is hard to find two comparable groups that differ only in their beliefs about the algorithm.

Finally, our work suggests that platforms should be cognizant of strategization in how they design their recommendation algorithms for two reasons. First, strategization may bias training data that future algorithms are trained on. User engagement data depends not only on the users' preferences, but also the users' beliefs about the current recommendation algorithm. This can also bias the (offline) evaluation of new algorithms, on the available training data. Second, the presence of strategization may indicate that some algorithms are giving suboptimal recommendations. Some users in the post-experiment free response express that they strategize because they may see too much of a type of content that they like. Such feedback would suggest that these algorithms should recommend a variety of content, and limit how much of any one type is shown.

References

Arel-Bundock V (2023) marginaleffects: Predictions, comparisons, slopes, marginal means, and hypothesis tests.

Cen S, Ilyas A, Madry A (2023) User strategization and trust on data-driven platforms. *arXiv preprint*.

Cen SH, Ilyas A, Madry A (2022) A Game-Theoretic perspective on trust in recommendation. *NeurIPS Workshop on Responsible Decision Making in Dynamic Environments (RDD)*.

DeVito MA (2021) Adaptive folk theorization as a path to algorithmic literacy on changing platforms. *Proc. ACM Hum.-Comput. Interact.* 5(CSCW2):1–38.

DeVito MA, Hancock JT, French M, Birnholtz J, Antin J, Karahalios K, Tong S, Shklovski I (2018) The algorithm and the user: How can HCI use lay understandings of algorithmic systems? *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–6, number Paper panel04 in CHI EA '18 (New York, NY, USA: Association for Computing Machinery).

Edelman B, Ostrovsky M (2007) Strategic bidder behavior in sponsored search auctions. *Decision support systems* 43(1):192–198.

Esponda I, Pouzo D (2016) Berk–Nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica: journal of the Econometric Society* .

Fudenberg D, Lanzani G, Strack P (2021) Limit points of endogenous misspecified learning. *Econometrica: journal of the Econometric Society* 89(3):1065–1098.

Hardt M, Megiddo N, Papadimitriou C, Wootters M (2016) Strategic classification. *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 111–122, ITCS '16 (New York, NY, USA: Association for Computing Machinery).

Haupt A, Hadfield-Menell D, Podimata C (2023) Recommending to strategic users .

Heidhues P, Koszegi B, Strack P (2021) Convergence in models of misspecified learning. *Theoretical economics* 16(1):73–99.

Kleinberg J, Mullainathan S, Raghavan M (2022) The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *Proceedings of the 23rd ACM Conference on Economics and Computation*, 29, EC '22 (New York, NY, USA: Association for Computing Machinery).

Lee AY, Mieczkowski H, Ellison NB, Hancock JT (2022) The algorithmic crystal: Conceptualizing the self through algorithmic personalization on TikTok. *Proc. ACM Hum.-Comput. Interact.* 6(CSCW2):1–22.

Marshall A (2020) Uber changes its rules, and drivers adjust their strategies. *Wired* .

Mehrotra R, Lalmas M, Kenney D, Lim-Meng T, Hashemian G (2019) Jointly Leveraging Intent and Interaction Signals to Predict User Satisfaction with Slate Recommendations. *The World Wide Web Conference*, 1256–1267, WWW '19 (New York, NY, USA: Association for Computing Machinery), ISBN 978-1-4503-6674-8, URL http://dx.doi.org/10.1145/3308558.3313613.

Mehrotra R, Xue N, Lalmas M (2020) Bandit based Optimization of Multiple Objectives on a Music Streaming Platform. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3224–3233 (Virtual Event CA USA: ACM), ISBN 978-1-4503-7998-4, URL http://dx.doi.org/10.1145/3394486.3403374.

Narayanan A (2023a) How to train your TikTok. https://knightcolumbia.org/blog/how-to-train-your-tiktok, accessed: 2023-11-10.

Narayanan A (2023b) Understanding social media recommendation algorithms. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms, accessed: 2023-11-10.

Newman N, Fletcher R, Kalogeropoulos A, Levy D, Nielsen RK (2018) Reuters institute digital news report 2018.

Perdomo J, Zrnic T, Mendler-Dünner C, Hardt M (2020) Performative prediction. Iii HD, Singh A, eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7599–7609 (PMLR).

Petre C, Duffy BE, Hund E (2019) “gaming the system”: Platform paternalism and the politics of algorithmic visibility. *Social Media + Society* 5(4):2056305119879995.

Rahman HA (2021) The invisible cage: Workers’ reactivity to opaque algorithmic evaluations. *Administrative science quarterly* 66(4):945–988.

Ricci F, Rokach L, Shapira B (2011) Introduction to recommender systems handbook. Ricci F, Rokach L, Shapira B, Kantor PB, eds., *Recommender Systems Handbook*, 1–35 (Boston, MA: Springer US).

Roth A, Balcan MF, Kalai A, Mansour Y (2010) On the equilibria of alternating move games. *Proceedings of the 2010 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 805–816, Proceedings (Society for Industrial and Applied Mathematics).

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

Simpson E, Hamann A, Semaan B (2022) How to tame “your” algorithm: LGBTQ+ users’ domestication of TikTok. *Proc. ACM Hum.-Comput. Interact.* 6(GROUP):1–27.

Sirlin N, Epstein Z, Arechar AA, Rand DG (2021) Digital literacy is associated with more discerning accuracy judgments but not sharing intentions .

Taylor SH, Choi M (2022) An initial conceptualization of algorithm responsiveness: Comparing perceptions of algorithms across social media platforms. *Social Media + Society* 8(4):20563051221144322.

Wooldridge JM (1999) Quasi-Likelihood Methods for Count Data. *Handbook of Applied Econometrics Volume 2: Microeconomics*, 321–368 (John Wiley & Sons, Ltd), ISBN 978-1-4051-6641-6, URL http://dx.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.

Appendix A: Additional Experimental Details

A.1. Post-Experiment Survey Questions

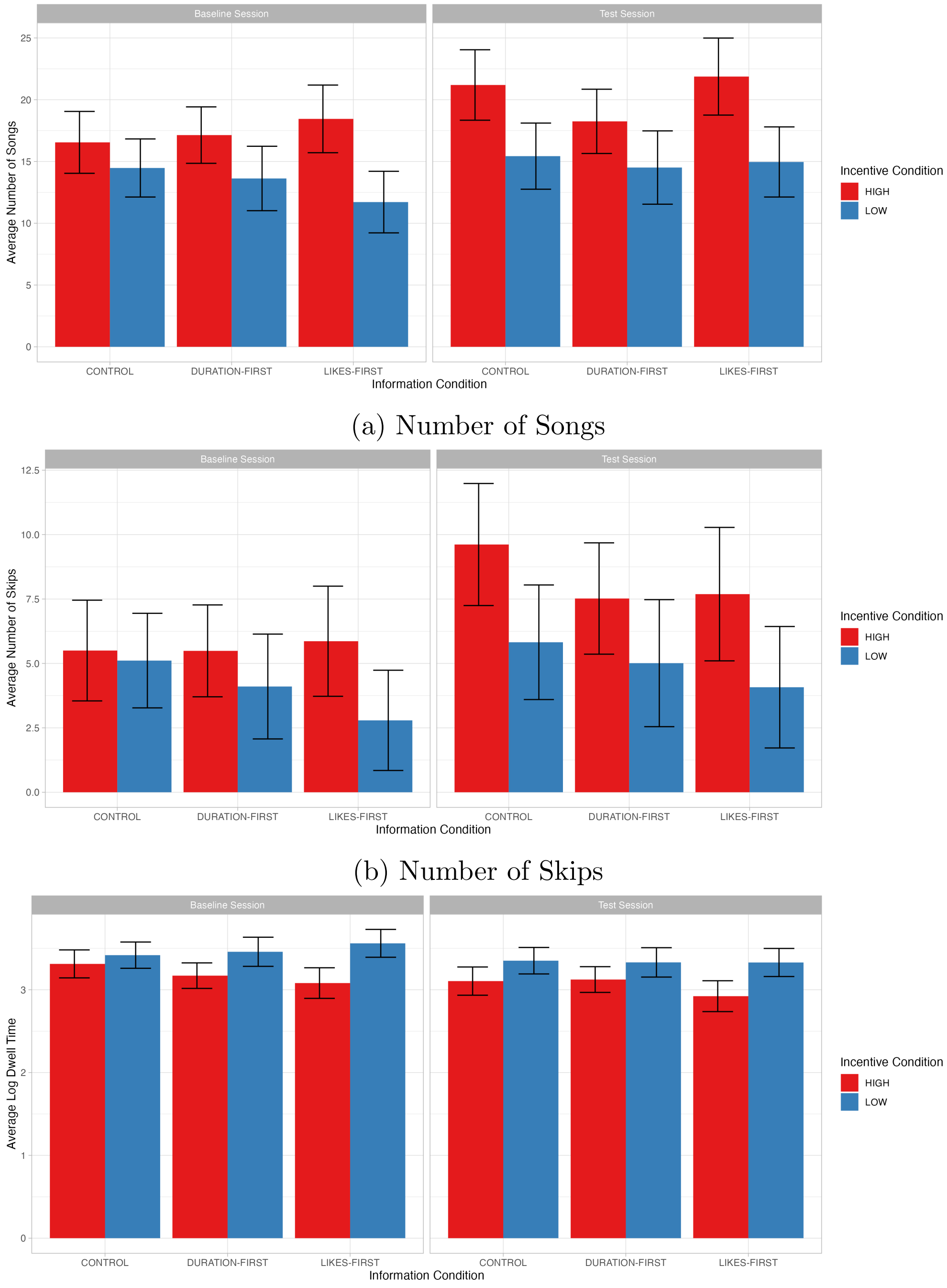
1. *Did the way that you interacted with songs change across the three listening sessions?* Answers: (a) Definitely yes, (b) Probably yes, (c) Probably no, (d) Definitely no, (e) I don't know.
2. *If yes, how did your interactions change across listening sessions? (CHECK ALL THAT APPLY)* Answers: (a) 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.

Appendix B: Additional post-experiment survey results

An exhaustive list of plots is in the "figures/post_experiment_plots" folder.

B.1. Additional Figures, Average Treatment Effect



(a) Number of Songs

(b) Number of Skips

(c) Log Dwell Time

B.2. Additional Tables, Average Treatment Effect

Table 2 Quasipoisson

Dependent Var.	Likes + Dislikes		Songs		Skips	
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
1(High Incentive)	0.53^{***}	0.32^{**}	0.32^{***}	0.24^{**}	0.50^{**}	0.49^{**}
	(0.15)	(0.15)	(0.12)	(0.10)	(0.24)	(0.22)
1(Duration Info)	-0.43^{***}	-0.37^{**}	-0.06	-0.007	-0.15	-0.02
	(0.15)	(0.14)	(0.12)	(0.10)	(0.28)	(0.23)
1(Likes Info)	0.27^{*}	0.39^{***}	-0.03	0.14	-0.36	0.03
	(0.14)	(0.14)	(0.11)	(0.10)	(0.27)	(0.24)
1(High Incentive) × 1(Duration Info)	0.03	0.004	-0.09	-0.16	-0.10	-0.23
	(0.22)	(0.17)	(0.17)	(0.12)	(0.36)	(0.29)
1(High Incentive) × 1(Likes Info)	-0.17	-0.12	0.06	-0.17	0.13	-0.29
	(0.19)	(0.17)	(0.16)	(0.12)	(0.36)	(0.32)
Baseline Like + Dislikes Count		0.05^{***}				
		(0.003)				
Baseline Songs Count				0.03^{***}		
				(0.002)		
Baseline Skips Count						0.06^{***}
						(0.003)
Fit statistics						
Observations	635	635	635	635	635	635
Squared Correlation	0.08	0.52	0.04	0.51	0.02	0.41

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

B.3. 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.

Table 3 OLS Models

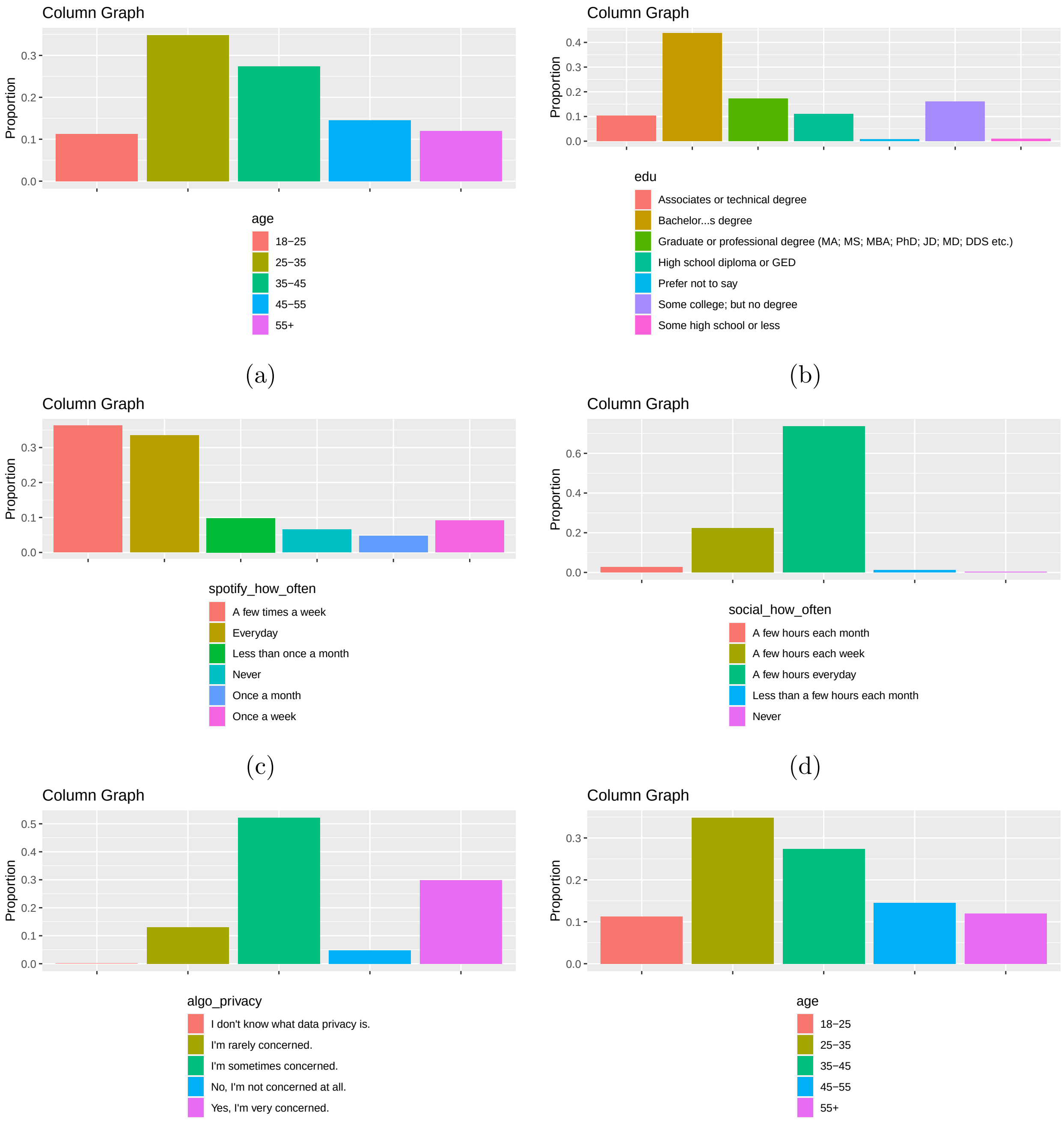
Dependent Var.	Likes + Dislikes		Dislikes		Likes		Songs	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
1(High Incentive)	5.2^{***}	2.4^{**}	3.7^{***}	1.9^{**}	1.5^{***}	0.67^{*}	5.8^{**}	4.0^{***}
	(1.6)	(1.0)	(1.3)	(0.85)	(0.48)	(0.36)	(2.2)	(1.5)
1(Duration Info)	-2.6^{***}	-2.3^{***}	-2.1^{**}	-1.6^{**}	-0.56^{*}	-0.64^{**}	-0.92	-0.18
	(0.95)	(0.85)	(0.81)	(0.75)	(0.31)	(0.28)	(1.8)	(1.4)
1(Likes Info)	2.3^{**}	2.9^{***}	1.6	2.2^{***}	0.75^{**}	0.77^{***}	-0.47	1.9
	(1.2)	(0.93)	(1.1)	(0.82)	(0.32)	(0.29)	(1.7)	(1.4)
1(High Incentive) × 1(Duration Info)	-1.6	-1.0	-0.84	-0.57	-0.76	-0.51	-2.0	-3.3^{*}
	(1.9)	(1.2)	(1.7)	(1.1)	(0.58)	(0.45)	(2.9)	(1.9)
1(High Incentive) × 1(Likes Info)	-0.93	-0.49	-0.97	-0.84	0.04	0.28	1.2	-2.9
	(2.1)	(1.4)	(1.8)	(1.2)	(0.74)	(0.54)	(3.0)	(1.9)
Baseline Like + Dislikes Count		0.88^{***}						
		(0.07)						
Baseline Dislikes Count				0.94^{***}				
				(0.07)				
Baseline Likes Count						0.65^{***}		
						(0.08)		
Songs Count								0.87^{***}
								(0.05)
Fit statistics								
Observations	635	635	635	635	635	635	635	635
R^2	0.08	0.62	0.05	0.59	0.07	0.47	0.04	0.60
Adjusted R^2	0.07	0.61	0.05	0.59	0.07	0.46	0.03	0.59

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

Table 4 Quasi Binomial Model

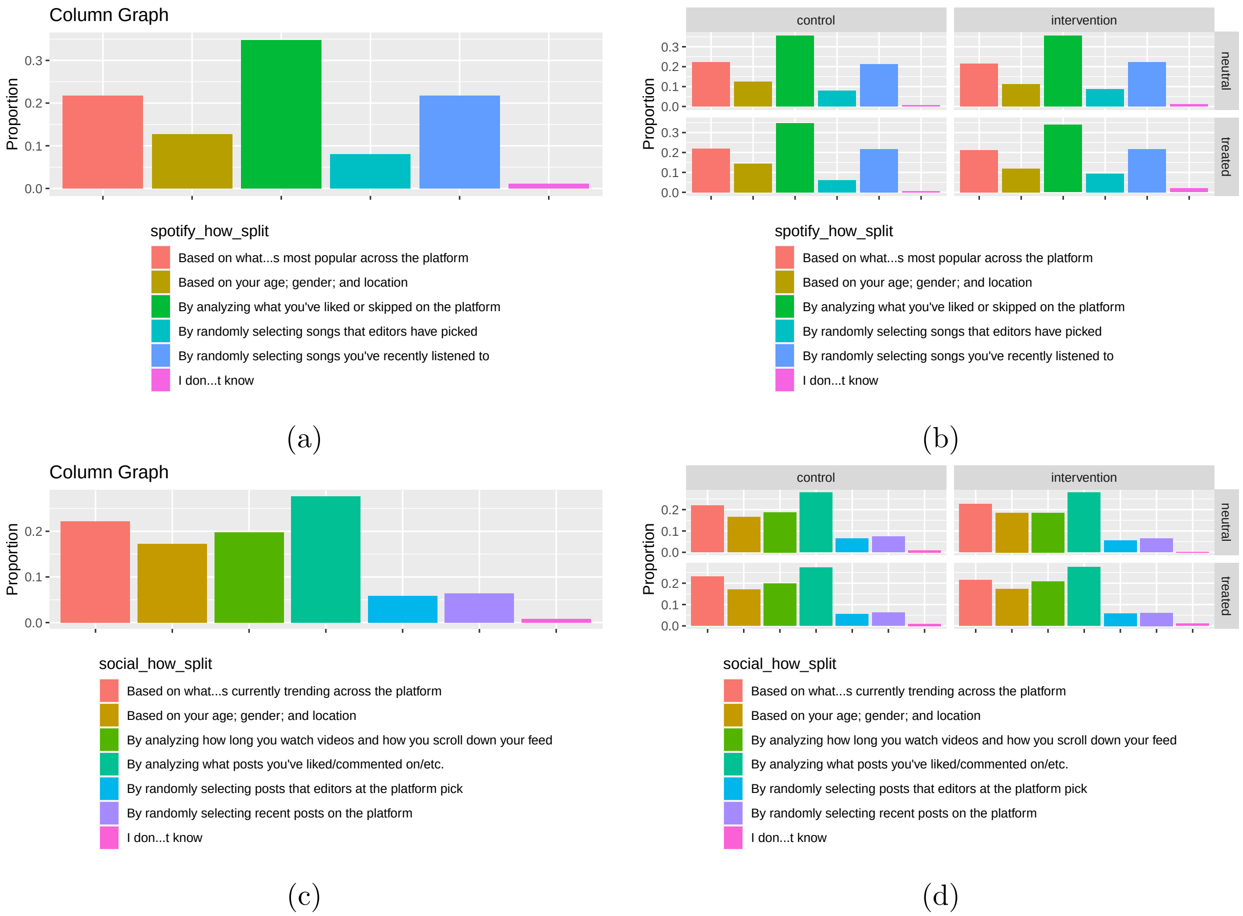
Dependent Var.	Likes + Dislikes Per Song	
Model:	(1)	(2)
Variables		
1(High Incentive)	0.46*	0.39*
	(0.25)	(0.23)
1(Duration Info)	-0.62***	-0.42*
	(0.24)	(0.24)
1(Likes Info)	0.70***	0.83***
	(0.25)	(0.24)
1(High Incentive) × 1(Duration Info)	0.08	0.002
	(0.35)	(0.29)
1(High Incentive) × 1(Likes Info)	-0.51	-0.13
	(0.34)	(0.31)
Baseline Likes + Dislikes Per Song		3.9***
		(0.22)
Fit statistics		
Observations	635	635
Squared Correlation	0.07	0.49

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



(a) (b) (c) (d) (e) (f)

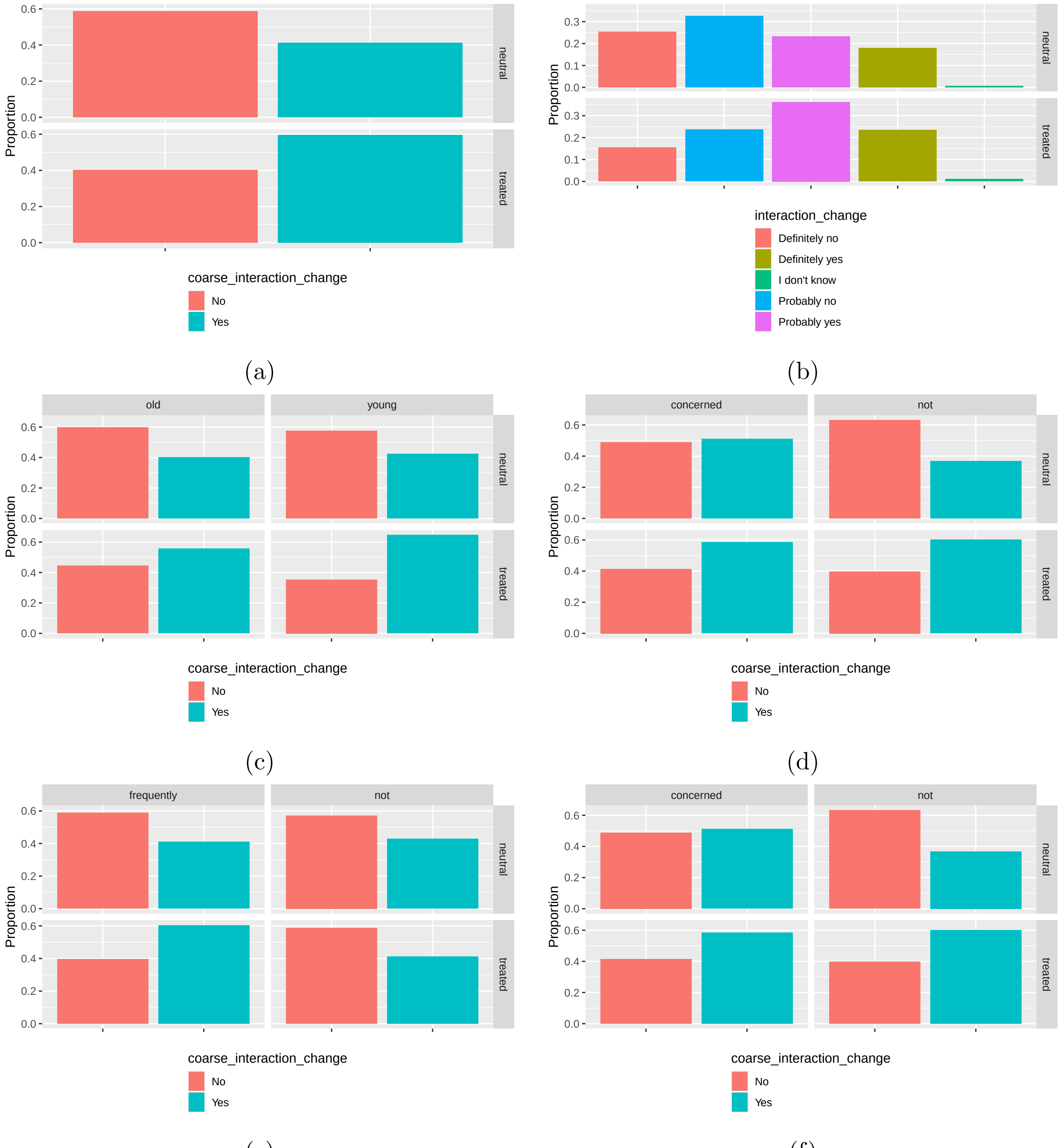
Figure 5 Demographic responses



(a) (b)

(c) (d)

Figure 6 How participants believe that Spotify (top) and social media algorithms (bottom) work.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7 **Whether users changed their behavior across different splits.**