Beyond the Stars: The Welfare Effects of Ratings under Homogeneous and Heterogeneous Consumer Tastes*

Noah Bohren Rustamdjan Hakimov Luís Santos-Pinto

July 24, 2026

Abstract

We study how rating systems affect welfare when consumer tastes are homogeneous or heterogeneous. A Bayesian model predicts that aggregate ratings improve welfare when consumers share a common ranking, but need not do so when product value is type-specific. In a pre-registered online experiment, ratings raise earnings in homogeneous-taste markets by improving match quality. In heterogeneous-taste markets, aggregate ratings leave average earnings unchanged but redistribute earnings toward the consumer type favored by early rating dynamics. Type-specific filtered ratings and algorithmic recommendations restore welfare gains; delayed disclosure does not. Rating design and market structure jointly determine welfare.

JEL Codes: *C90, D83, L15, L86*

Keywords: *Ratings, Algorithmic Recommendations, Preference Heterogeneity, Online Experiment*

*We are thankful to Pedro Freitas for his outstanding research assistance, and to Larbi Alaoui, Jean-Michel Benkert, Johannes Johnen, Robin Ng, Or Avishay-Rishi, Armin Schmutzler, Joel Sobel, Joao Montez, and participants of internal seminars at University of Lausanne and Universitat Pompeu Fabra for their valuable comments. We gratefully acknowledge financial support from the Swiss National Science Foundation, project number 00018-232179. The project received IRB approval from the LABEX Ethic Committee of HEC, University of Lausanne (#RAILER,02.05.23) and was pre-registered on AEA Registry (AEARCTR-0011386).
Noah Bohren: University of Lausanne, noah.bohren[at]unil[dot]ch
Rustamdjan Hakimov: University of Lausanne, rustamdjan.hakimov[at]unil[dot]ch
Luís Santos-Pinto: University of Lausanne, luispedro.santospinto[at]unil[dot]ch

1 Introduction

Online rating systems are a defining feature of digital markets. By aggregating individual experiences into simple metrics such as stars, likes, or reviews, they aim to reduce search costs, build trust, and help consumers identify desirable products. Yet the informational value of ratings depends critically on how they are generated and interpreted. Ratings are typically submitted by a self-selected subset of consumers, and early reviews may anchor beliefs and distort subsequent behavior. These challenges are exacerbated when consumers differ in their preferences: when products vary along dimensions that are not valued in the same way by all consumers, a single aggregated rating may conflate type-specific fit with commonly valued quality. In such settings, especially when consumers arrive sequentially, rating systems can mislead rather than inform, depending on their design and timing.

A growing literature documents both the promise and the limitations of rating systems. Reviews have been shown to increase trust, boost sales, and improve product visibility (Resnick et al., 2006; Anderson and Magruder, 2012; Luca, 2016; Tadelis, 2016; Chevalier and Mayzlin, 2006; Li et al., 2020; Reimers and Waldfogel, 2021; Cabral and Hortacsu, 2010; Bolton et al., 2004). The common view holds that aggregated ratings provide a reliable signal of product quality. However, researchers have also identified systematic biases that undermine their effectiveness (Tadelis, 2016). These include selection effects, strategic manipulation, fake reviews, cold-start problems, and inflated ratings resulting from harvesting strategies (Hu et al., 2006; Luca and Zervas, 2016; Hu et al., 2017, 2009; Acemoglu et al., 2022; Mayzlin et al., 2014; Carnehl et al., 2022; Che and Hörner, 2018; Vellodi, 2018; Dendorfer and Seibel, 2024; Johnen and Ng, 2024; Bolton et al., 2013).

Our contribution is threefold. First, we show that the welfare impact of ratings depends fundamentally on the structure of consumer tastes. When consumers have homogeneous tastes and share a common ranking of products, aggregate ratings can improve welfare by revealing which product has the higher expected payoff. When consumers have heterogeneous tastes and product value depends on consumer type, aggregate ratings may fail to improve welfare because they pool experiences across consumers for whom different products are optimal. Second, we provide controlled experimental evidence that isolates this mechanism, documents path dependence driven by the tastes of early reviewers, and shows that ratings redistribute earnings across consumer types even when they leave average earnings unchanged. Third, we evaluate three practical information policies: type-specific rating filters, delayed rating disclosure, and algorithmic recommendations. Filtering and algorithmic recommendations improve average welfare in the heterogeneous-taste environment, but algorithms introduce a new distortion through selective training data. Delayed disclosure has no effect on average welfare and does not

reduce the variability of market outcomes across rating sequences; it changes the source of that variability, from the tastes of early consumers to the order in which products accumulate enough ratings to be displayed.

To examine how the welfare effects of online rating systems vary with the structure of consumer tastes, we develop a Bayesian model in which risk-neutral consumers update their beliefs about product payoffs using average ratings and review counts before choosing between two products and an outside option. The model distinguishes between two environments. In the homogeneous-taste environment, consumers share the same ranking of products, so ratings reveal information about a common payoff component. In this case, ratings increase welfare through two channels: (i) they improve match rates by guiding consumers toward the product with the higher expected payoff, and (ii) they increase allocative efficiency by selectively raising purchase rates when quality is high and lowering them when quality is low, thus encouraging only those consumers who stand to gain a surplus to buy. In the heterogeneous-taste environment, different consumer types prefer different products and aggregate ratings direct consumers toward one product even though no single product is universally best. We show that, when ratings favor one of the two products, their welfare effect is in general ambiguous: ratings may either raise or lower welfare relative to a no-ratings baseline, depending on how they reallocate consumers across products and the outside option. In particular, ratings can improve welfare when they steer the right consumers toward their preferred product, but they can reduce welfare when they instead induce inefficient switching or exit. Finally, we show that introducing filtering-enhanced ratings in a heterogeneous-taste environment, that is, exposing each consumer type to ratings from their own type, raises welfare.

We then conduct a series of pre-registered experiments in a controlled online setting to test the model's predictions and alternative rating designs used in practice. We implement an online marketplace in which participants recruited from the Prolific platform choose between a safe outside option and one of two paid tasks, with earnings determined by performance. This setting is close to the decision environment participants regularly face on Prolific, where they select among tasks for monetary compensation. The tasks are pop-culture quizzes about real and fictional public figures and characters,[1] designed to implement either homogeneous or heterogeneous tastes. In the homogeneous-taste condition, one quiz is easier for both age groups and therefore has a higher expected payoff. In the heterogeneous-taste condition, the quiz with the higher expected payoff depends on the participant's age. We designed the quizzes based on a pretest: while age is irrelevant in the homogeneous-taste market, participants below 30 and above 50 years old exhibit opposing quiz preferences in the heterogeneous-taste market.[2]

[1] In what follows, we use "task" and "quiz" interchangeably.

[2] To ensure comparability, we restricted participation to individuals below 30 and above 50 years old. We targeted an equal share of participants below 30 and above 50 in each treatment; in the

The experimental design intentionally strips away several confounds present in field data, including unobserved product attributes, endogenous pricing (Carnehl et al., 2022), and selection into purchasing and rating. This allows us to isolate how aggregate ratings perform when tastes are either common across consumers or systematically type-specific. The cost of this control is that the environment is stylized: participants receive limited prior information about the two tasks before ratings are introduced. The estimated welfare gains should therefore be interpreted as benchmark effects in a setting where ratings are a central source of information, rather than as direct estimates of the magnitude of rating effects in markets with richer prior signals.

In the Baseline treatments, we establish benchmarks for homogeneous- and heterogeneous-taste environments by observing participants' task choices in the absence of ratings. In the Rating treatments, participants enter sequentially and observe the average rating on a 1–5 star scale and the number of ratings submitted by earlier participants. For each Rating treatment, we implement 14 independent sequences of 30 participants—15 young and 15 old, randomly ordered—each constituting a separate market with its own rating trajectory. This design mirrors real-world platforms in which early ratings may influence subsequent choices. Crucially, it enables us to study path dependence: in the heterogeneous-taste environment, age differences generate systematically opposed preferences, and variation in the age composition of early entrants allows us to test whether initial majority tastes anchor subsequent market outcomes.

Comparing outcomes between Rating and Baseline, we find that ratings significantly increase earnings in homogeneous-taste markets but have no effect on average earnings in heterogeneous-taste markets. The gains in homogeneous-taste markets are driven primarily by higher match rates: ratings help participants select the task with the higher expected payoff. However, we find no evidence that ratings increase the overall likelihood of purchasing a task rather than choosing the outside option. In heterogeneous-taste markets, the absence of an average effect masks a systematic reallocation: in market sequences whose ratings favor one task, participants of the type that prefers this task earn more relative to participants of the other type.

To address the limits of standard aggregate ratings in heterogeneous-taste markets, we introduce three policy treatments. In the Filtering treatment, ratings are segmented by consumer type. Specifically, ratings are displayed separately for participants below 30 and above 50 years old, as age is the primary dimension of preference heterogeneity in our setting. The model predicts that filtering is effective when it partitions consumers into groups within which tastes are sufficiently homogeneous.

sequential treatments, the split was exactly balanced within each sequence. Participants aged 30 to 50 were excluded to create a clear generational distinction between groups and to limit potential spillovers in quiz familiarity or preferences across age cohorts.

In the Freezing treatment, ratings are withheld until at least five reviews have accumulated for a task, a design inspired by platform practices that delay or suppress early ratings when early feedback is likely to be noisy or strategically distorted. This treatment is not formally analyzed in the theoretical model and is primarily motivated by platform practice. While it should not affect the long-run effectiveness of ratings, it may stabilize early ratings and reduce herding, thereby lowering the variance of market outcomes driven by early path dependence.

In the Algorithm treatment, ratings are replaced by personalized recommendations based on performance patterns observed in the Baseline treatment among participants with similar observable characteristics. We use ordinary least squares to predict expected earnings from each task based on age, gender, confidence, and risk aversion, and the algorithm recommends the task with the highest predicted earnings. This treatment is not derived from the theoretical model. Instead, it provides an empirical comparison between rating-based information and a direct-guidance policy in which participants observe a recommendation but not the underlying prediction rule.

Consistent with the model's prediction, the Filtering treatment significantly increases earnings in heterogeneous-taste markets. Earnings in the Algorithm treatment are also significantly higher than in the Baseline and not statistically different from Filtering, indicating that direct guidance can substitute for ratings. Although the two treatments deliver similar average earnings, they operate through different channels. The Algorithm treatment significantly increases the probability of purchase relative to the Baseline and raises the match rate; however, among matched participants, those in the Algorithm treatment earn significantly less than their counterparts in Filtering. This shortfall is consistent with the selective-labels problem (Kleinberg et al., 2018): the algorithm is trained on the biased subset of participants who self-select into purchasing the task. The Freezing treatment does not affect average earnings and does not reduce the dispersion of outcomes across sequences; it removes the dependence of market outcomes on the tastes of early participants but introduces a new dependence on which task is the first to reach the display threshold.

Beyond treatment effects, our setting allows descriptive analysis of rating behavior along the extensive and intensive margins. We document four main findings. First, 97% of participants rated the task they completed, despite receiving no monetary incentive, with no differences across treatments.[3] This near-universal willingness to rate suggests that the classic selection biases observed in consumer reviews using observational data (Dellarocas and Wood, 2008; Cabral and Li, 2015; Burtch et al., 2018; Fradkin and Holtz, 2023) are unlikely here, likely because the rating prompt appeared immediately after task completion and required minimal effort. Second, rating likelihood was systematic:

[3]Lafky (2014) similarly reports high rating participation in experimental settings when providing feedback is costless. Introducing even small rating costs significantly reduces propensity to give feedback.

participants with higher earnings and male participants were more likely to rate, while the average rating and number of existing reviews had no effect. Third, conditional on rating, assigned scores were strongly and positively correlated with quiz earnings, indicating that participants evaluated tasks in proportion to the payoffs received. Fourth, younger participants gave significantly lower ratings than older ones, conditional on earnings and the match.

We next analyze how participants incorporate rating information into their purchase decisions. Across all treatments, 23% of participants choose the safe outside option, with this share remaining stable even as more rating information becomes available. Less confident and more risk-averse participants are significantly more likely to choose the outside option. Among those who purchase, 70.5% select the quiz with the higher average rating, and choices respond systematically to both rating level and the number of reviews: larger gaps in either dimension increase the likelihood of choosing the higher-rated option. Following the higher average rating yields 25.5% higher earnings, on average, relative to not following it.

Finally, rating dynamics in heterogeneous-taste markets are sensitive to the order in which participants enter the market. The preferences expressed by early participants shape the choices of those who arrive later, creating herding patterns that reflect initial conditions rather than intrinsic product qualities. This path dependence highlights how ratings can amplify early signals and lock markets into outcomes that may not be welfare-maximizing.

The studies most closely related to ours are two recent papers that examine rating systems under consumer heterogeneity. Benkert and Schmutzler (2024) develop a theoretical framework to study the informativeness and value of ratings when preferences differ across consumers, characterizing optimal recommendation systems when homogeneous and heterogeneous tastes coexist. In contrast, we isolate these two dimensions and focus on their distinct welfare implications, allowing for a transparent comparison across market structures. Lafky and Ng (2024) show experimentally that raters project their own preferences when leaving ratings and that consumers interpret ratings through the lens of perceived rater similarity. We complement this work by studying how alternative rating designs—filtering, delayed disclosure, and algorithmic recommendations—can restore welfare in markets with heterogeneous tastes despite such preference-driven biases.

We also contribute to the literature on rating and reputation systems, which studies how informational signals aggregate over time and shape beliefs and incentives in markets with asymmetric information (Bar-Isaac et al., 2008; Klein et al., 2016), by showing that the informativeness of ratings depends fundamentally on the heterogeneity of tastes. Prior work documents several sources of bias that distort ratings, including selection effects that generate disproportionately extreme reviews (Hu et al., 2006, 2017, 2009), strategic

manipulation through fake reviews (Mayzlin et al., 2014; Luca and Zervas, 2016; He et al., 2022). Price effects can also bias ratings, as consumers who pay higher prices tend to leave lower ratings due to higher expectations (Carnehl et al., 2022), and firms may strategically shape early information—through pricing, marketing, or selective disclosure—to harvest favorable ratings from early consumers (Bar-Isaac et al., 2010; Johnen and Ng, 2024). Related work shows that recommender systems and algorithmic interventions can mitigate some of these biases and improve efficiency (Che and Hörner, 2018; Bénabou and Vellodi, 2024), while rating systems may also reinforce incumbency advantages (Vellodi, 2018). Our contribution is to show that even unbiased ratings can fail to improve welfare in heterogeneous-taste markets, and to identify rating designs that recover the welfare gains of ratings in such environments.

This study also contributes to the literature on social learning and reputation systems. Consumers update beliefs based on observed reviews, yet selection effects and early realizations can distort long-run outcomes (Acemoglu et al., 2022; Ifrach et al., 2019; Salganik et al., 2006). Herding and early rating biases can generate path dependence, making market outcomes sensitive to initial conditions. Our multiple-market experimental design provides clean evidence of this mechanism: early participants exert disproportionate influence on later choices in heterogeneous-taste markets. While delaying the disclosure of ratings attenuates path dependence, it does not translate into measurable welfare gains.

Finally, this paper contributes to the design of alternative rating mechanisms that correct for the limitations of traditional ratings (Bolton et al., 2013; Che and Hörner, 2018; Vellodi, 2018; Bénabou and Vellodi, 2024; Dendorfer and Seibel, 2024; Lafky and Ng, 2024). Platforms have implemented filtering systems to help consumers interpret ratings more effectively. Some platforms also delay early reviews to mitigate the impact of biased initial ratings. Our findings show that filtering ratings and algorithmic recommendations provide an effective alternative to ratings in heterogeneous-taste markets.

The rest of the paper is structured as follows. Section 2 introduces our stylized model and derives its main predictions. Section 3 describes the experiment. Section 4 discusses the results on welfare effects of ratings, ratings' determinants, and responses to ratings. Section 5 concludes. The Supplemental Appendix contains the proofs of the theoretical results, additional tables and figures, and the experimental instructions.

2 Theoretical Model

This section presents a Bayesian decision-theoretic framework to investigate how product ratings influence welfare in markets with homogeneous and heterogeneous consumer tastes. The model's purpose is not to show that ratings can reveal a universally preferred product—this is immediate under common tastes—but to separate the channels through

which rating information affects welfare and to clarify why the same aggregation rule can fail when ratings pool type-specific experiences.

2.1 Homogeneous-Taste Market

Risk-neutral subjects choose between a safe outside option and one of two quiz-based tasks: an Easy quiz, denoted by E, and a Hard quiz, denoted by H. Each quiz $k \in \{E, H\}$ has an underlying quality parameter q_k. Quality should be interpreted as the objective feature of the quiz that affects the probability of successfully completing it. We assume that $0 < q_H < q_E$, so that the Easy quiz has higher quality than the Hard quiz. Higher quality makes success more likely.

Each subject is characterized by a skill level $\theta \in [0, 1]$, distributed in the population according to cumulative distribution function $F(\theta)$ and density $f(\theta)$. A subject knows her own skill θ when making her choice. The distribution F is common knowledge. If a subject with skill θ takes a quiz of quality $q > 0$, her probability of success is

$$p(\theta, q) = \theta(1 - e^{-q}).$$

This specification has four useful properties:

1. Boundedness. For all skill $\theta \in [0, 1]$ and quality $q > 0$ we have $0 \leq p(\theta, q) \leq \theta \leq 1$.
2. Monotonicity. The probability of success p is increasing in both skill θ and quality q: $\partial p/\partial \theta = 1 - e^{-q} > 0$, and $\partial p/\partial q = \theta e^{-q} > 0$.
3. Diminishing returns in quality q: $\partial^2 p/\partial q^2 = -\theta e^{-q} < 0$.
4. Complementarity between skill θ and quality q: $\partial^2 p/\partial \theta \, \partial q = e^{-q} > 0$.

Before choosing, subjects observe information about the quality of the quizzes. For each quiz $k \in \{E, H\}$, there are n_k previous ratings $R_1^k, \ldots, R_{n_k}^k$. These ratings are informative signals about the quiz's quality. Conditional on the true quality q_k, ratings are independently drawn according to $R_i^k \sim N(q_k, \sigma^2)$. Hence, the ratings depend only on the quiz's quality and not on the skill of the subject who generated the rating. For each quiz k, subjects observe the number of ratings n_k and the sample mean rating

$$\bar{r}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} R_i^k.$$

Subjects do not observe the true quality q_k. Instead, they form beliefs about it. Let Q_k denote the unknown quality of quiz k. The prior belief about each quiz quality is

$Q_k \sim N(\mu, \tau^2)$, with $\mu > 0$ and $\tau > 0$. The prior is common knowledge. Conditional on observing $\bar{r}_k$ and n_k, Bayesian updating gives the posterior mean

$$E(Q_k \mid \bar{r}_k) = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n_k}} \bar{r}_k + \frac{\frac{\sigma^2}{n_k}}{\frac{\sigma^2}{n_k} + \tau^2} \mu, \tag{1}$$

and posterior variance

$$V(Q_k \mid \bar{r}_k) = \left(\frac{1}{\tau^2} + \frac{n_k}{\sigma^2} \right)^{-1}.$$

Let $g_k(q \mid \bar{r}_k, n_k)$ denote the posterior density of Q_k, with posterior mean $E(Q_k \mid \bar{r}_k)$ and posterior variance $V(Q_k \mid \bar{r}_k)$.

The timing is as follows.

1. Each subject learns her own skill θ.

2. For each quiz $k \in \{E, H\}$, the subject observes the number of previous ratings n_k and the mean rating $\bar{r}_k$.

3. Using this information, the subject forms posterior beliefs about each quiz quality Q_k.

4. The subject chooses one of three options: the outside option, the Easy quiz E, or the Hard quiz H.

5. If the subject chooses a quiz, her success probability depends on her skill and the quiz's true quality according to $p(\theta, q_k)$. If she chooses the outside option, she receives a fixed payoff.

The payoff from taking quiz k has two components: a fixed payment $a > 0$ and a piece-rate payment $b > 0$ paid upon success. Therefore, conditional on the true quiz quality q_k, a subject of skill θ has expected payoff

$$a + bp(\theta, q_k) = a + b\theta(1 - e^{-q_k}).$$

Since the subject does not observe q_k, she evaluates quiz k using her posterior belief about Q_k. Her subjective expected payoff from quiz k, after observing $(\bar{r}_k, n_k)$, is

$$U_k(\theta) = a + b\theta \int (1 - e^{-q}) g_k(q \mid \bar{r}_k, n_k) dq.$$

The outside option pays a certain amount $z \in (a, a + b)$. The lower bound $z > a$ implies that the outside option is not so low that quizzes are always strictly better. The upper bound $z < a + b$ implies that the outside option is not so high that quizzes are always unattractive, since a sufficiently skilled subject facing a sufficiently favorable quiz may prefer a quiz to the outside option.

A subject with skill θ prefers quiz k to the outside option if

$$a + b\theta \int (1 - e^{-q}) g_k(q \mid \bar{r}_k, n_k) dq > z.$$

Equivalently, quiz k is preferred to the outside option whenever

$$\theta > \frac{z - a}{b \int (1 - e^{-q}) g_k(q \mid \bar{r}_k, n_k) dq}.$$

A subject who compares the two quizzes prefers the Easy quiz E to the Hard quiz H if

$$\int (1 - e^{-q}) g_E(q \mid \bar{r}_E, n_E) dq > \int (1 - e^{-q}) g_H(q \mid \bar{r}_H, n_H) dq.$$

Similarly, the subject prefers H to E if the inequality is reversed. Notice that this comparison does not depend on the subject's skill θ, because skill enters the expected payoff from each quiz multiplicatively. Skill affects whether the subject prefers a quiz to the outside option, but it does not affect which quiz she prefers conditional on choosing a quiz.

Finally, if a subject is indifferent among several available alternatives, the subject randomizes uniformly over all alternatives that maximize expected payoff.

Our first result shows that ratings improve welfare in a homogeneous-taste market.

Proposition 1: *In a homogeneous-taste market such that $\mu > \tau^2/2$ and $z < a + b(1 - e^{-\mu+\tau^2/2})$, average earnings are higher with ratings than without, i.e.,*

$$E\big[I_{HOM}(\text{Ratings})\big] > E\big[I_{HOM}(\text{Baseline})\big].$$

Proposition 1 shows that in a homogeneous-taste market with an easy (higher-quality) quiz and a hard (lower-quality) quiz, ratings raise welfare relative to a no-ratings baseline through two channels that depend on subjects' prior beliefs. As shown by inequality (4) in the Supplemental Appendix, when prior beliefs about quiz quality are low, $\mu < q_E + \tau^2/2$, ratings (i) increase take-up by inducing intermediate-skill subjects with $\theta \in \big[\bar{\theta}_E, \bar{\theta}\big]$ who would otherwise opt out to purchase the easy quiz, and (ii) improve matching by redirecting high-skill subjects with $\theta \in \big[\bar{\theta}, 1\big]$ who would otherwise choose the hard quiz toward the easy quiz. As shown by inequality (5) in the Supplemental Appendix, when prior beliefs about quiz quality are high, $\mu \geq q_E + \tau^2/2$, ratings still raise welfare: they (i) reduce take-up by prompting intermediate-skill subjects with $\theta \in \big[\bar{\theta}, \bar{\theta}_E\big]$ who would otherwise purchase a quiz to opt out, and (ii) improve matching for high-skill subjects with $\theta \in \big[\bar{\theta}_E, 1\big]$. This yields our first hypothesis.

H1: *Average earnings in homogeneous-taste markets with ratings are significantly higher than without ratings.*

2.2 Heterogeneous-Taste Market

We now consider two quizzes, Y and O, with type-specific qualities: q_{YY} and q_{YO} for the "young" quiz, and q_{OY} and q_{OO} for the "old" quiz. Young subjects comprise a fraction α of the population, while old subjects comprise the remaining fraction $1-\alpha$, where $\alpha \in (0,1)$. To introduce heterogeneous tastes, we assume $0 < q_{YO} < q_{YY}$ and $0 < q_{OY} < q_{OO}$, that is, the young quiz has higher quality for young subjects and the old quiz has higher quality for old subjects. The success probability in quiz $k \in \{Y, O\}$ by a subject of type $j \in \{Y, O\}$ is

$$p(\theta, q_{kj}) \;=\; \theta(1 - e^{-q_{kj}}).$$

Subjects of type j who take quiz k provide ratings $R_1^{kj}, \ldots, R_{n_{kj}}^{kj}$ drawn from $N(q_{kj}, \sigma^2)$. A subject with prior beliefs $N(\mu, \tau^2)$ who observes that the mean rating of quiz k by n_{kj} subjects of type j is $\bar{r}_{kj}$ infers the posterior mean

$$E(Q_{kj} \mid \bar{r}_{kj}) \;=\; \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n_{kj}}}\, \bar{r}_{kj} \;+\; \frac{\frac{\sigma^2}{n_{kj}}}{\frac{\sigma^2}{n_{kj}} + \tau^2}\, \mu, \tag{2}$$

and the posterior variance

$$V(Q_{kj} | \bar{r}_{kj}) = \left(\frac{1}{\tau^2} + \frac{n_{kj}}{\sigma^2} \right)^{-1}.$$

The expected payoff of a subject of type j with skill θ of taking quiz k after observing mean rating $\bar{r}_{kj}$ by n_{kj} subjects of type j is obtained in an analogous way as in the homogeneous-taste market.

In the heterogeneous-taste market with ratings, we assume that subjects observe only the aggregate rating information associated with each quiz. Specifically, for both the young and the old quiz, they see the mean rating and the total number of reviews, but not the composition of reviewers by type. In particular, subjects do not observe that an α share of reviews is provided by young subjects and a $1-\alpha$ share by old subjects, or that the corresponding mean ratings differ across reviewer types and quizzes. Accordingly, subjects base their decisions solely on the aggregate mean ratings and volume of reviews of each quiz.

Proposition 2: *Consider a heterogeneous-taste market such that* $\mu > \tau^2/2$, $z < a + b(1 - e^{-\mu+\tau^2/2})$, *and where* $q_Y = \alpha q_{YY} + (1-\alpha)q_{YO} > \alpha q_{OY} + (1-\alpha)q_{OO} = q_O$, *i.e., ratings induce subjects to purchase the young quiz rather than the old quiz. Then the effect of ratings on average earnings is as follows.*
(i) If $q_{YO} + \tau^2/2 \leq \mu < q_Y + \tau^2/2$, *ratings raise average earnings if and only if inequality (8) in the Supplemental Appendix holds; otherwise, ratings reduce average earnings.*
(ii) If $q_Y + \tau^2/2 \leq \mu \leq q_{YY} + \tau^2/2$, *ratings raise average earnings if and only if inequality (9) in the Supplemental Appendix holds; otherwise, ratings reduce average earnings.*

Proposition 2 shows that, in a heterogeneous-taste market with two age-specific quizzes (young and old), where ratings favor the young quiz, the introduction of ratings may either enhance or reduce welfare relative to a no-ratings baseline. As shown by inequality (8) in the Supplemental Appendix, when prior beliefs about quiz quality are relatively low, that is, when $q_{YO} + \tau^2/2 \leq \mu < q_Y + \tau^2/2$, ratings operate through two margins: they induce intermediate-skill subjects who would otherwise opt out to purchase the young quiz, and they redirect high-skill subjects who would otherwise choose the old quiz toward the young quiz. Both effects increase the welfare of the young and decrease the welfare of the old. Hence, ratings raise aggregate welfare if the gain for the young exceeds the loss for the old; otherwise, they lower it. As shown by inequality (9) in the Supplemental Appendix, when prior beliefs about quiz quality are relatively high, that is, when $q_Y + \tau^2/2 \leq \mu \leq q_{YY} + \tau^2/2$, ratings affect welfare through two offsetting channels. First, they induce intermediate-skill subjects who would otherwise purchase the young quiz to opt out. Second, they redirect high-skill subjects who would otherwise choose the old quiz toward the young quiz. The first channel reduces the welfare of the young but increases the welfare of the old, whereas the second increases the welfare of the young but reduces the welfare of the old. Hence, the overall effect of ratings on welfare is ambiguous and depends on the relative strength of these two opposing forces.[4]

Proposition 2 implies that ratings do not have a uniform welfare effect in heterogeneous-taste markets. When ratings favor the young quiz, they shift demand toward the young quiz; when ratings favor the old quiz, they shift demand toward the old quiz. This reallocation benefits the group whose preferred quiz is favored by ratings but may harm the other group. Hence, whether ratings raise or lower average earnings depends on the relative size of these gains and losses. Since the experimental markets differ in the direction and strength of the reallocation induced by ratings, the theory predicts heterogeneous treatment effects across markets rather than a uniform welfare gain. This leads to our

[4] A fully symmetric result obtains when ratings favor the old quiz rather than the young quiz, that is, when $q_O > q_Y$, with the roles of the two quizzes reversed. By contrast, when $q_Y = q_O$, ratings do not reallocate demand between the young and old quizzes and leave welfare unchanged only if $q_Y = q_O = \mu - \tau^2/2$.

second hypothesis.

H2: *In heterogeneous-taste markets, ratings do not have a uniform effect on average earnings. Depending on the market, ratings may increase or decrease average earnings relative to the no-ratings baseline; therefore, pooled across heterogeneous-taste markets, ratings are not expected to generate a systematic earnings gain.*

The reallocation logic behind Proposition 2 also has a distributional implication that can be tested directly. When the ratings observed in a market favor the young quiz, young subjects are more likely to be guided toward their preferred quiz, and old subjects are more likely to be guided away from theirs; the reverse holds when ratings favor the old quiz. The earnings of each type should therefore increase in the rating advantage of that type's preferred quiz.

Prediction 1. *In a heterogeneous-taste market with ratings, the earnings of each consumer type increase in the rating advantage of the type's preferred quiz: a rating advantage of the young quiz raises the expected earnings of young subjects and lowers those of old subjects, and symmetrically when the old quiz is favored.*

Prediction 1 was not part of the pre-analysis plan; we derive it from Proposition 2 and test it in Section 4.1.1 as a supplementary, non-pre-registered analysis.

Finally, we show that introducing ratings enhanced by filtering, that is, letting subjects see the ratings from their own type, helps them pick the right quiz and hence raises welfare.

Proposition 3: *In a heterogeneous-taste market such that* $\mu > \tau^2/2$ *and* $z < a + b(1 - e^{-\mu+\tau^2/2})$*, average earnings with filtering-enhanced ratings exceed those without, i.e.,*

$$E\big[I_{HET}(\text{Filtering})\big] > E\big[I_{HET}(\text{Baseline})\big].$$

Proposition 3 shows that in a heterogeneous-taste market with two age-specific quizzes (young and old), filtering-enhanced ratings raise welfare relative to a no-ratings baseline. Filtering of ratings allows for improved matching of high-skill young and old subjects to their age-specific quiz.[5] In addition, when prior beliefs about quiz quality are low, filtering of ratings increases take-up by guiding intermediate-skill young and old subjects to take their age-specific quiz rather than opting out. When prior beliefs about quiz quality are high, filtering of ratings reduces take-up by steering intermediate-skill young

[5]The model assumes that young and old subjects choose their information structures optimally. In the Filtering treatment in the experiment, participants observe the ratings by age groups but they may make mistakes when selecting an information structure (Charness et al., 2021).

and old subjects to the outside option rather than buying their age-specific quiz. In short, filtering of ratings essentially splits the market into two sub-markets with homogeneous age-specific tastes. This leads to our third hypothesis.

H3: *Average earnings in heterogeneous-taste markets with filtering-enhanced ratings are significantly higher than without filtering-enhanced ratings.*

The model treats ratings as exogenous, whereas in the experiment they are endogenously generated by participants who base their decisions on earlier ratings. Despite this difference, the model serves three purposes for our study. First, it gives us a simple benchmark for how participants should react to higher vs. lower ratings, holding everything else fixed. Second, it illustrates the two channels whereby ratings can improve welfare, namely, improving match rates and buying rates. Third, it provides testable hypotheses for the different impact of ratings on welfare in markets with homogeneous and heterogeneous consumer tastes.

3 Experimental Design

This section presents the experimental protocol, which received ethical approval and was pre-registered prior to data collection.[6] We begin by outlining the main objectives of the study, followed by a detailed description of the design, implementation procedures, and treatment structure.

3.1 Primary Objectives

The experiment simulates an online marketplace where consumers choose between purchasing one of two goods or opting out. A key feature of the design is the controlled field setup: participants are online workers from the Prolific platform whose primary goal is to earn money by completing various tasks. We simulate an online market for tasks within the experiment, closely following our theoretical model.

Participants choose between two primary tasks (quizzes), which represent products in either homogeneous- or heterogeneous-taste markets, and a fallback task, which represents the option not to buy. Upon completing a primary task, participants evaluate it on a 1-to-5-star scale. The experimental design varies both consumer tastes (homogeneous vs. heterogeneous) and the availability of ratings (no ratings vs. ratings). Additionally,

[6]The project received IRB approval from the LABEX Ethic Committee of HEC, University of Lausanne (RAILER,02.05.23) and was pre-registered on AEA Registry (AEARCTR-0011386).

we introduce variations in rating and recommendation systems for heterogeneous-taste markets, testing filtering, freezing, and algorithmic recommendations.

The primary objectives of the study are: (1) to evaluate the causal effects of the presence of rating systems on consumer welfare[7] in homogeneous- and heterogeneous-taste markets, (2) to compare the effectiveness of alternative rating and recommendation systems in heterogeneous-taste markets, and (3) to examine how participants incorporate rating information in their decision-making process.

3.2 Stages and tasks of experiment

The experiment was conducted on Qualtrics, using the sample from Prolific, based in the US. The experiment consisted of three stages: (1) the buying stage, where participants reviewed task descriptions and selected one to complete; (2) the task stage, where participants completed the chosen task; and (3) the rating stage, where participants who selected one of the quizzes could rate it.

Each participant received an initial endowment of £2.50. The primary tasks were quizzes, with participants selecting and completing one of two available quizzes. Taking a quiz required a £1.70 fee, deducted from the initial endowment.[8]

Each quiz consisted of ten multiple-choice questions about public figures and characters, with six answer options per question, only one of which was correct. Earnings were performance-based, with £0.45 awarded per correct answer, leading to possible final payouts ranging from £0.80 (no correct answers) to £5.30 (all correct answers). To ensure participants understood the payment structure, comprehensive examples were provided before the main experiment. Participants had 110 seconds to complete the quiz, with a strict time limit of 11 seconds per question. This constraint minimized the likelihood of searching for answers online. Participants were not allowed to skip or revisit questions.[9] Participants did not complete a practice round. This may have increased uncertainty about the quiz task, but the information and timing rules were common across treatments, and the absence of practice preserves the intended pre-choice uncertainty faced by all participants.

Alternatively, participants could select a fallback task, which involved counting zeros in 10 distinct matrices for 110 seconds. This option was free, allowing participants to retain their full initial endowment of £2.50 but without the opportunity to earn additional income. The inclusion of this outside option aligns with the theoretical model and enables

[7]Welfare is measured by participant payoffs. While completing a task may provide additional intrinsic utility, we abstract from this consideration.

[8]Participants were compensated in GBP. They were informed that the exchange rate at the time of the study was £1 = $1.25.

[9]Comprehensive instructions and screenshots of the experiment's interface are available in the Supplemental Appendix.

us to examine the dual role of online ratings: (1) assisting customers in deciding whether to purchase a product and (2) guiding customers in selecting the highest-quality product.

After completing a quiz, participants had the option to rate it on a 1-to-5 scale. The rating prompt stated: "Please provide your opinion on QUIZ [name of the quiz] on a scale of 1 to 5. This information can be helpful for future participants. You may skip this section if you choose to do so." Ratings were entirely optional.

Additionally, we collected individual characteristics of participants. Recruitment on Prolific was restricted ex ante to individuals below 30 years old or above 50 years old, and this screening criterion determined assignment to the corresponding age category in the experiment. For control variables, however, we rely on self-reported demographic information collected within the survey, including age and gender. The two age measures are highly consistent: only 24 out of 2,279 participants (1.1%) reported an age inconsistent with their Prolific-screened category, indicating minimal measurement error in the recruitment procedure.

Prior to the buying stage, we also elicited participants' confidence in their expected quiz performance. Specifically, participants were asked: *"Imagine taking a quiz about celebrities. Out of 100 randomly selected people who also took the same quiz, how many do you think would perform worse than you?"* We use this non-incentivized measure as a proxy for confidence and, therefore, for participants' perceived likelihood of benefiting from purchasing a quiz.

Following the rating stage, participants completed a standard non-incentivized measure of risk preferences on a 0–10 scale, following Falk et al. (2018). We intentionally placed this measure at the end of the experiment to minimize potential priming effects that could distort earlier purchase decisions by making risk considerations salient before the buying stage. We acknowledge that eliciting risk preferences after earnings were realized may introduce some measurement endogeneity if outcomes affect subsequent self-reports. However, given the modest monetary stakes of the experiment, any such bias is likely limited.

3.3 Treatment Variation 1: Consumer Tastes

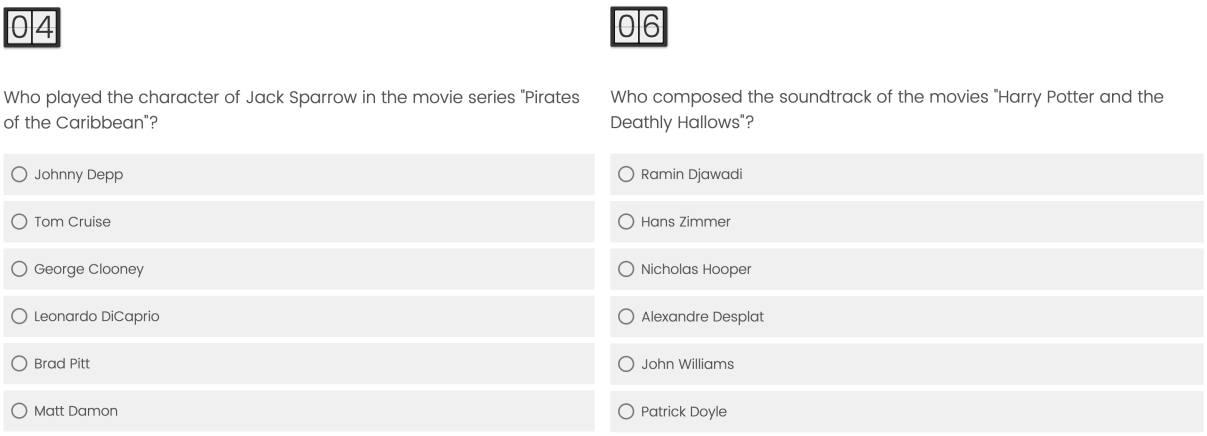
The first key dimension of treatment variation is consumer tastes. In some treatments, we establish a homogeneous-taste market using hard-easy quizzes, while in the other treatments, we create a heterogeneous-taste market using quizzes targeted to different age groups. The next two subsections detail the differences between the two types of markets.

3.3.1 Homogeneous-Taste Market

To establish a homogeneous-taste market we designed two public-figure and pop-culture quizzes with distinct difficulty levels: one *easy* and one *hard*. The *easy* quiz represents the higher-quality product, as participants pay the same fee for either quiz but have a higher likelihood of earning greater rewards by selecting the *easy* quiz over the *hard* one.

To construct the *easy* and *hard* quizzes, we pretested a pool of 120 public-figure and pop-culture quiz questions with 260 Prolific participants similar to those recruited for the main experiment. Based on these pretests, we selected 10 questions for each quiz to reflect the intended difficulty levels. Figure 1 provides an example of selected questions for the homogeneous-taste market.

Figure 1: Example of selected questions for homogeneous-taste market



(a) *easy* quiz **(b)** *hard* quiz

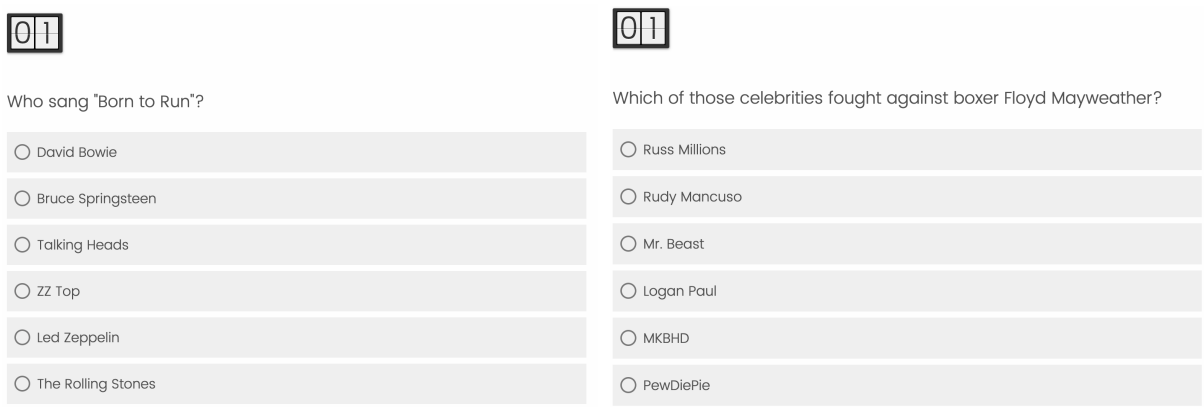
To confirm that the *easy* and *hard* quizzes represent a homogeneous-taste market, we compared pretest participants' performance across both quizzes. The *easy* quiz had an average correct response rate of 69%, yielding expected earnings of £3.90, while the *hard* quiz had a correct response rate of 36%, with expected earnings of £2.40. Among pretest participants who completed both quizzes, 98.5% scored the same or higher on the *easy* quiz than on the *hard* quiz. These results confirm that the *easy* and *hard* quizzes effectively capture a homogeneous-taste environment.

3.3.2 Heterogeneous-Taste Market

To create a heterogeneous-taste market in the experiment we designed two generation-specific quizzes: one tailored for the older generation (*50+* quiz) and one for the younger generation (*30-* quiz). The quizzes were constructed using the same pretest population

as in the homogeneous-taste market, with ten questions selected for each quiz.[10] Figure 2 provides an example of selected questions.

Figure 2: Example of selected questions for the heterogeneous-taste market



(a) *50+* quiz **(b)** *30-* quiz

To confirm that the generation-specific quizzes capture a heterogeneous-taste market, we analyzed the performance of two distinct pretest age groups: the *old* group (aged 50 and above) and the *young* group (aged 18 to 30). Among older participants, 89% performed better or equally well on the *50+* quiz compared to the *30-* quiz. Similarly, 91% of younger participants performed better on the *30-* quiz than on the *50+* quiz. Table 1 summarizes mean earnings by age group and quiz type.[11]

Table 1: Mean earnings by age group and quiz type during pretesting

Age Group	50+ Quiz	30- Quiz
Old	£4.00	£2.38
Young	£2.46	£3.89

These results confirm that the generation-specific quizzes successfully create heterogeneous preferences over quizzes.

3.4 Treatment Variation 2: Availability of Ratings

The second key dimension of treatment variation is the availability of ratings. Two main treatments, *Baseline* and *Rating*, were implemented across both homogeneous and

[10]For the homogeneous-taste market, a standard t-test reveals no significant performance differences between the *old* and *young* groups for either the *easy* ($p = 0.68$) or *hard* quiz ($p = 0.81$).

[11]For the heterogeneous-taste market, earnings along the diagonal of Table 1—where participants are matched to their generation's quiz—do not significantly differ between age groups ($p = 0.27$). The same holds for the off-diagonal mismatched cases ($p = 0.46$).

heterogeneous-taste markets to assess the causal effects of the presence of rating systems on welfare. The following subsections detail the differences between these treatments.

3.4.1 Baseline

To establish a benchmark for welfare in the absence of a rating system, we recruited 199 participants for each Baseline market type (homogeneous and heterogeneous), with 100 participants below 30 and 99 participants above 50. Participants selected between two quizzes or the fallback counting task. Those who chose a quiz could rate it, but these ratings were not displayed to subsequent participants.

To prevent any implicit signaling about quiz difficulty or content, quizzes were labeled neutrally as Quiz BLUE and Quiz YELLOW. Figure 3a summarizes the information provided at the buying stage.

3.4.2 Rating

In this treatment, participants entered the market sequentially and observed, for each quiz, the average rating and the number of ratings submitted by prior participants. Sequential entry was implemented by recruiting participants one at a time within each of the 14 parallel sequences. Each sequence consisted of 30 participants—15 below 30 years old and 15 above 50 years old—arranged in a randomly predetermined order. After each participant completed the task, average ratings were manually verified and updated before recruiting the next participant. Depending on the experimental condition, participants have either homogeneous or heterogeneous tastes over quizzes. This design mirrors rating aggregation and display practices on major online platforms such as Google, Amazon, and eBay, where both the average rating and the number of reviews are prominently displayed. Figure 3b summarizes the information available at the buying stage.

Figure 3: Screenshot of task information provided to participants

Below you can find a table summarizing the choices you can make. Please select one option.

	Quiz BLUE	Quiz YELLOW	Counting Task
Topic	Celebrities (cinema, music, politics, sports, ...)	Celebrities (cinema, music, politics, sports, ...)	Counting zeros in tables
Time	110s	110s	110s
Possible £ outcome	£0.8-£5.3	£0.8-£5.3	£2.5

(a) *Baseline*

Below you can find a table summarizing the choices you can make. Please select one option. Note that earlier participants had the opportunity to rate quizzes on a 1-5 scale. The table displays both the average rating and the number of reviews (indicated in parentheses).

	Quiz BLUE	Quiz YELLOW	Counting Task
Topic	Celebrities (cinema, music, politics, sports, ...)	Celebrities (cinema, music, politics, sports, ...)	Counting zeros in tables
Time	110s	110s	110s
Possible £ outcome	£0.8-£5.3	£0.8-£5.3	£2.5
Average rating from previous participants	2.9 (7)	4.8 (12)	

(b) *Rating*

Comparing participant earnings between the Baseline and Rating treatments allows us to identify the causal effect of the introduction of ratings on earnings in both homogeneous and heterogeneous-taste markets.

The Rating treatments allow for an individual-level analysis of purchase behavior, which we exploit to address two additional questions. First, we examine how participants' purchase decisions respond to rating information and to individual characteristics. Second, conditional on making a purchase, we analyze how the selection of a quiz depends on both its average rating and the number of reviews.

Beyond individual behavior, rating systems may introduce path dependence in market outcomes. Because each rating affects subsequent decisions, early fluctuations—potentially unrelated to product quality—can steer later choices. The Rating treatments allow us to investigate whether markets with similar early ratings converge toward comparable welfare outcomes and whether markets with divergent early ratings evolve differently due to such path dependencies. This question is particularly important, as it speaks to the risk that a product's eventual success may hinge more on early randomness than on intrinsic quality.

To study this, we create 14 independent markets in each consumer-taste environment (homogeneous and heterogeneous) with 30 participants in each market making choices sequentially. Participants in each group observe only the ratings generated within their own market, ensuring that each sequence evolves independently. This setup provides a clean environment to assess whether rating systems facilitate herding behavior and how much the characteristics of early participants shape the decisions of those who follow.

3.5 Policy Treatments: Enhanced Rating and Algorithm Recommendation Systems

Within the heterogeneous-taste market, we examine the welfare effects of two enhanced rating systems, filtering and freezing, alongside one algorithmic recommendation system. We label these policy treatments as *Filtering*, *Freezing*, and *Algorithm*. These enhancements address the potential insufficiency of ratings alone in improving welfare in heterogeneous-taste markets. Each treatment is described in detail below.

3.5.1 Filtering

As Proposition 2 shows, the effect of ratings alone on welfare is generally ambiguous in heterogeneous-taste markets. To address this, platforms such as Booking.com allow users to sort or filter reviews along several dimensions, illustrating how platforms can expose consumers to more targeted rating information. Such an enhanced rating system exploits the structure of heterogeneous-taste markets, which are composed of subgroups

with distinct preferences. As Proposition 3 shows, filtering ratings by subgroups increases the informativeness of the ratings that are most relevant to each subgroup. In theoretical terms, this mechanism is conceptually related to the idea in Benkert and Schmutzler (2024) that platform designers can influence sender–receiver alignment by selecting the sender population.

To evaluate whether filtering ratings by subgroups can improve welfare, we introduce the Filtering treatment. This treatment replicates the conditions of the Rating treatment with 14 independent markets, each comprising 30 participants interacting sequentially. However, participants observe the ratings by age groups (below 30 years old and above 50 years old), capturing the two subgroups with opposing quiz preferences. This treatment allows us to assess whether filtering by age enhances welfare in the heterogeneous-taste market. Figure 4a summarizes the information provided at the buying stage.

3.5.2 Freezing

In heterogeneous-taste markets, early ratings can disproportionately shape subsequent choices. When initial reviews misrepresent product quality and discourage future purchases, learning may stall, especially if better products are prematurely ignored. Acemoglu et al. (2022) show that Bayesian agents can eventually uncover true quality from reviews, but only if some consumers continue purchasing poorly rated products. When products are in competition, demand might halt completely and misperceptions may persist indefinitely. This concern is supported by real-world evidence showing that early reviews can have long-lasting effects on business outcomes. For instance, early negative reviews have been shown to harm restaurants' long-term success, even when later performance improves.[12]

To mitigate the risk of early misrepresentation, we introduce the Freezing treatment. It mirrors the structure of the Rating treatment, with 14 independent markets of 30 sequential participants each, but suppresses the display of ratings until a task has received at least five reviews. This delay is intended to reduce the impact of noisy early feedback and provide participants with more stable, representative signals.[13]

3.5.3 Algorithm

Many online platforms selling products in heterogeneous-taste markets—such as Netflix, Spotify, and Amazon—rely on algorithmic recommendation systems. These systems, of-

[12]See: https://news.osu.edu/how-a-few-negative-online-reviews-early-on-can-hurt-a-restaurant/ (last accessed June 1, 2025).

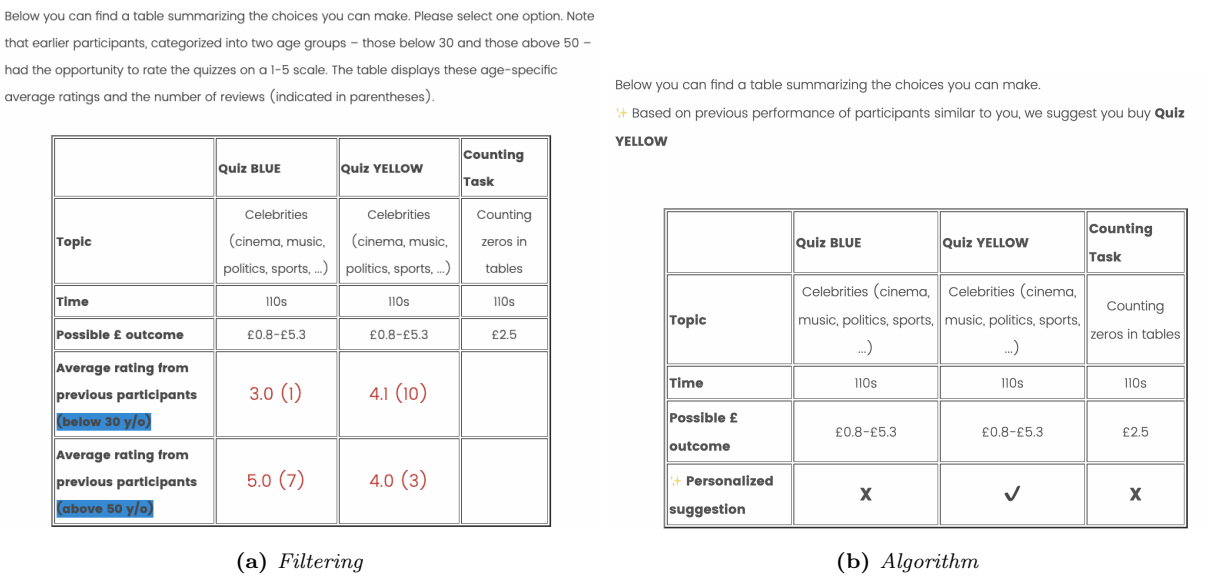
[13]This approach is inspired by Amazon's decision to delay ratings for new releases—such as The Rings of Power—to counteract "review bombing" and ensure more informative feedback from a broader audience (Forbes, 2022b,a).

ten based on collaborative filtering or knowledge-based methods, complement or replace traditional star ratings to help consumers navigate complex choice environments. While filtering can improve the informativeness of ratings in heterogeneous-taste markets, algorithms may further enhance consumer guidance by directly incorporating individual-level characteristics.

To evaluate whether algorithmic recommendations outperform traditional rating and filtering systems, we introduce the Algorithm treatment. Using data from the Baseline condition, we estimate an OLS model that predicts individual earnings from each quiz based on participants' age, gender, confidence, and risk aversion. The algorithm then recommends the quiz associated with the highest predicted payoff.[14] In this treatment, participants did not observe any ratings. Instead, they received a personalized recommendation with the message: "Based on previous performance of participants similar to you, we suggest you buy XX quiz."[15]

We recruited 201 participants for this treatment, with 100 participants above 50 and 101 participants below 30. Because recommendations were not based on prior participants' behavior within the session, purchase decisions were not path-dependent. Therefore, we did not implement the independent market structure used in the Rating treatment. Figure 4b summarizes the information provided at the buying stage.

Figure 4: Screenshot of task information provided to participants



Below you can find a table summarizing the choices you can make. Please select one option. Note that earlier participants, categorized into two age groups – those below 30 and those above 50 – had the opportunity to rate the quizzes on a 1-5 scale. The table displays these age-specific average ratings and the number of reviews (indicated in parentheses).

	Quiz BLUE	Quiz YELLOW	Counting Task
Topic	Celebrities (cinema, music, politics, sports, ...)	Celebrities (cinema, music, politics, sports, ...)	Counting zeros in tables
Time	110s	110s	110s
Possible £ outcome	£0.8-£5.3	£0.8-£5.3	£2.5
Average rating from previous participants (below 30 y/o)	3.0 (1)	4.1 (10)	
Average rating from previous participants (above 50 y/o)	5.0 (7)	4.0 (3)	

(a) *Filtering*

	Quiz BLUE	Quiz YELLOW	Counting Task
Topic	Celebrities (cinema, music, politics, sports, ...)	Celebrities (cinema, music, politics, sports, ...)	Counting zeros in tables
Time	110s	110s	110s
Possible £ outcome	£0.8-£5.3	£0.8-£5.3	£2.5
✨ Personalized suggestion	✗	✓	✗

(b) *Algorithm*

[14] Age is the primary predictor. All participants were recommended the quiz corresponding to their age group. The model never predicted selection of the Counting Zeros task for a matched individual. Full estimation results are provided in Table A1 in the Supplemental Appendix.

[15] Participants were not informed about the specific inputs or structure of the algorithm, in line with standard practice on commercial platforms.

3.6 Implementation details

For the main experiment, we recruited a total of 2,279 participants. We targeted a balanced age composition in each treatment. In the sequential treatments, the split was exactly balanced within each sequence, with 15 participants below 30 and 15 above 50.[16] In the nonsequential treatments, final cell counts differ by at most one participant.[17] We refer to the Baseline and Rating treatments, implemented in both homogeneous- and heterogeneous-taste markets, as the four main treatments. We refer to Filtering, Freezing, and Algorithm as policy treatments, implemented in the heterogeneous-taste markets to address potential inefficiencies of standard rating systems arising from taste heterogeneity. Each participant could take part in the experiment only once.

After accessing the survey, participants first reported their self-assessed confidence in public-figure and pop-culture quizzes (Figure A3), were informed of the payment structure (Figures A4–A5), and then viewed the decision screen presenting available quizzes with ratings displayed according to treatment (Figure A6). They subsequently completed their chosen task, observed their performance outcome, and were invited to submit a rating (Figure A10). Finally, they reported their risk preferences following Falk et al. (2018) (Figure A11) before exiting the platform.

Table 2 presents a detailed breakdown of participant numbers and average ages across treatments. On average, participants completed the experiment in 6.95 minutes, with the average payoff of £3.18—equivalent to an hourly rate of £27.40, well above Prolific's recommended guidelines.

Table 2: Number of participants per treatment by age group

	Main treatments				Policy treatments		
Age Group	Homogeneous Baseline	Homogeneous Rating	Heterogeneous Baseline	Heterogeneous Rating	Heterogeneous Filtering	Heterogeneous Freezing	Heterogeneous Algorithm
Older than 50 y.o.	99	210	99	210	210	210	100
	(58.5)	*(60.3)*	*(59.3)*	*(60.5)*	*(58.2)*	*(57.2)*	*(58.7)*
Younger than 30 y.o.	100	210	100	210	210	210	101
	(25.3)	*(24.9)*	*(24.6)*	*(24.8)*	*(24.7)*	*(25.3)*	*(24.7)*
All	199	420	199	420	420	420	201

Average age in parentheses

[16] For each sequence, we generated a random ordering with 15 younger and 15 older participants and invited participants according to that order.

[17] As mentioned, we do not have participants between 31 and 49 years old, which ensures clear heterogeneous-taste markets. We kept the same age composition in the homogeneous-taste markets to ensure a clean comparison between the two types of consumer tastes.

4 Results

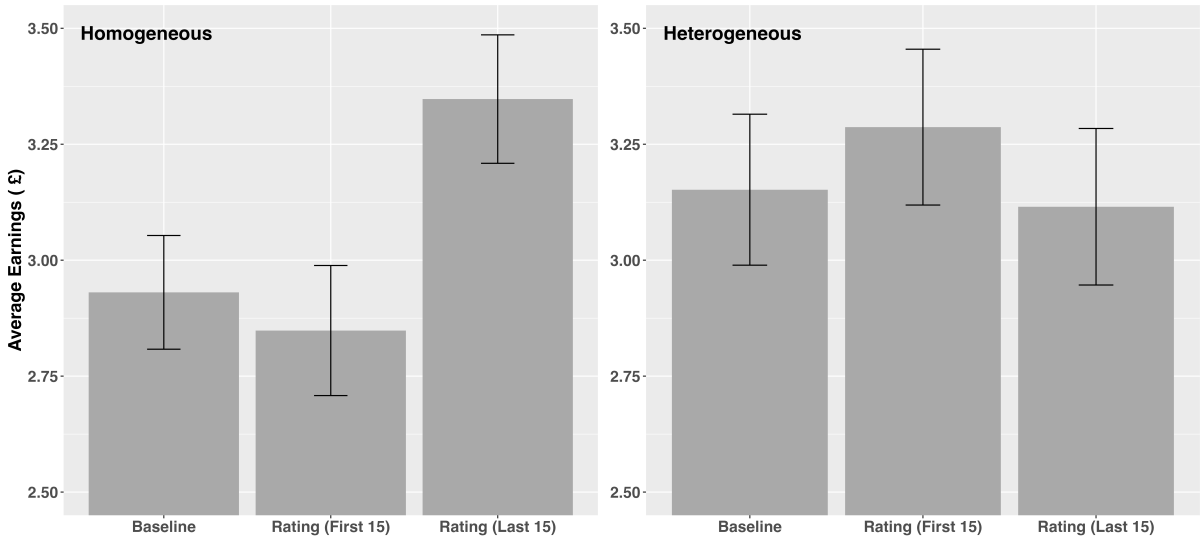
This section presents the experimental results. We begin by analyzing earnings in the two primary treatments, focusing on the effects of consumer tastes and the availability of ratings. We then examine the impact of the policy treatments, followed by an analysis of the determinants of individual rating and purchase behavior.

4.1 Earnings: Main Treatments

We first compare welfare along two experimental dimensions: (i) consumer tastes (homogeneous vs. heterogeneous) and (ii) availability of ratings (no ratings vs. ratings). Figure 5 displays average participant earnings across the main treatments, disaggregated by the first 15 and last 15 participants in each sequence.

To estimate the causal effect of the presence of ratings, we compare earnings in the Baseline treatment with those of the last 15 participants in the Rating treatment.[18] The results show that ratings significantly improve earnings in homogeneous-taste markets, but have no measurable effect in heterogeneous-taste markets.

Figure 5: Average earnings in main treatments



Error bars are 95% confidence intervals. The vertical axis starts at £2.50, the payoff of the safe outside option, so bar heights show average earnings in excess of the safe payoff.

[18] Focusing on the last 15 participants follows the pre-analysis plan, as these participants are exposed to more fully developed rating information.

Table 3: Treatment effects on earnings, match rate, and buy rate

Dependent Variable:	Earnings (£)		Match (%)		Buy (%)	
	(Homogeneous)	(Heterogeneous)	(Homogeneous)	(Heterogeneous)	(Homogeneous)	(Heterogeneous)
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Constant	2.876***	3.020***	0.341***	0.563***	0.779***	0.804***
	(0.151)	(0.134)	(0.091)	(0.073)	(0.055)	(0.071)
Rating first 15	-0.066	0.137	0.186**	0.050	0.019	-0.025
	(0.116)	(0.124)	(0.087)	(0.054)	(0.042)	(0.040)
Rating last 15	0.433***	-0.044	0.417***	-0.058	0.039	0.023
	(0.101)	(0.104)	(0.061)	(0.055)	(0.046)	(0.047)
Gender = Male	-0.046	0.102	-0.043	0.040	-0.050	-0.030
	(0.078)	(0.106)	(0.066)	(0.048)	(0.044)	(0.038)
Risk aversion (0-10)	0.028	0.011	-0.003	-0.004	-0.018***	-0.026***
	(0.018)	(0.018)	(0.011)	(0.009)	(0.006)	(0.009)
Confidence (0-100)	-0.001	0.001	0.000	-0.002*	0.001*	0.002**
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Gender	✓	✓	✓	✓	✓	✓
Risk Aversion	✓	✓	✓	✓	✓	✓
Confidence	✓	✓	✓	✓	✓	✓
Observations	616	617	468	470	616	617
R^2	0.058	0.006	0.119	0.017	0.023	0.046
Adjusted R^2	0.050	-0.002	0.109	0.007	0.015	0.038

Results of OLS regressions with standard errors using the CR2 bias-reduced cluster-robust estimator with Satterthwaite degrees-of-freedom adjustment (Pustejovsky and Tipton, 2018), clustered at the individual level for Baseline treatments and at the independent sequence level for Rating treatments. Controls include gender, risk aversion, and self-assessed confidence in the task. Significantly different from zero at 1% (***), 5% (**), 10% (*).

Models (1) and (2) in Table 3 report regression estimates of treatment effects on participant earnings. In homogeneous-taste markets, the average earnings of the last 15 participants in each sequence of the Rating treatment are significantly higher than those in the Baseline treatment. By contrast, the first 15 participants in the Rating sequences do not earn significantly more than their Baseline counterparts, suggesting that ratings become effective only once sufficient information has accumulated. Consistent with this interpretation, earnings within the Rating treatment are themselves significantly higher for the last 15 participants than for the first 15 participants ($\Delta = 0.49$, $p < 0.01$). Relative to the Baseline treatment, the resulting welfare gain for the last 15 participants corresponds to a 14% increase in average earnings.[19] These findings support **Hypothesis H1** (*average earnings in homogeneous-taste markets are significantly higher with ratings than without*).

In heterogeneous-taste markets, however, the average earnings of the last 15 participants in the Rating treatment are not significantly different from those in the Baseline. This is consistent with **Hypothesis H2** (*pooled across heterogeneous-taste markets, rat-*

[19] All reported differences include the fixed participation fee of £0.80 received by all participants. Excluding this fixed component would yield larger proportional treatment effects.

ings do not generate a systematic earnings gain).[20] The null result is informative about magnitudes. Given the standard error of 0.104 on the Rating-last-15 coefficient in model (2), the design detects effects of £0.29 or larger—about 10 percent of Baseline earnings, and two-thirds of the corresponding effect in homogeneous-taste markets—with 80 percent power. We can therefore rule out aggregate welfare gains in heterogeneous-taste markets comparable to those observed under homogeneous tastes. As we show in Section 4.1.1, however, the absence of an average effect does not mean that ratings leave individual outcomes unchanged.

To further illustrate these dynamics, Figure A1 in the Supplemental Appendix plots average earnings by sequence position. Although position-specific averages are naturally noisy, as each sequence rank contains only 14 observations, the broader pattern is informative. In the homogeneous-taste market, earnings in the later part of the sequence are systematically higher: all of the final 15 positions lie above the Baseline benchmark. This pattern is consistent with ratings becoming more informative as additional feedback accumulates. By contrast, in the heterogeneous-taste market, only 5 of the final 15 positions exceed the Baseline benchmark, with no comparable upward pattern. The figure therefore complements the regression evidence by showing that welfare gains from ratings emerge broadly across later positions only in the homogeneous-taste market.

As the theoretical model shows, the observed earnings gains in the homogeneous-taste market may be driven by two main channels:

1. **Better match rate** – Ratings help guide participants toward selecting the easier quiz, thereby increasing their expected earnings.

2. **Better buying rate** - Ratings modify the purchase decision of intermediate-skill participants. When prior beliefs about quiz quality are low, ratings encourage such participants to purchase the easy quiz rather than opt for the outside option (the counting-zeros task). When prior beliefs are high, ratings instead discourage intermediate skill participants who would otherwise purchase a quiz from doing so, redirecting them toward the outside option.

Models (3) and (4) in Table 3 analyze match rates, while models (5) and (6) examine buying rates across treatments. In the homogeneous-taste market, we observe a substantial increase in match rates, from 32% in the Baseline treatment to 73% for the last

[20]We observe higher earnings in the first 15 rounds of the heterogeneous-taste market relative to the homogeneous-taste market. This is due to participants' preference for the Blue quiz, which is the suboptimal option for all participants in the homogeneous-taste market and therefore causes a larger earnings loss there. In heterogeneous-taste markets, the cost of this preference is lower because Blue is the preferred quiz for half of the participants. This pattern also explains why Baseline earnings are higher in heterogeneous than in homogeneous-taste markets. More details of the determinants of the quiz choice are presented in section 4.3.1.

15 participants in the Rating treatment.[21] Within the Rating treatment itself, match rates are significantly higher for the last 15 participants than for the first 15 participants ($\Delta = 0.23$, $p < 0.01$), consistent with the gradual accumulation of informative ratings over time.

By contrast, the introduction of ratings did not significantly increase the rate at which participants chose to purchase a quiz. This pattern is consistent with the theoretical predictions under particular parameterizations of prior beliefs, where the positive and negative extensive-margin effects of ratings may offset in the aggregate. It may also indicate that participants primarily used ratings to choose between tasks rather than to decide whether any task was worth purchasing. Consistent with the absence of strong extensive-margin effects, we do not observe meaningful differences in average self-reported confidence—our proxy for expected quiz performance—between the first 15 participants (47.54) and the last 15 participants (44.78; $\Delta = 2.76, p = 0.21$). This comparison should be interpreted cautiously, however, as self-reported confidence is an imperfect and potentially noisy proxy for underlying skill.[22] We return to this question in Section 4.3.1.

In the heterogeneous-taste market, we find no significant improvement in either match rates or buying rates relative to the Baseline treatment. If anything, match rates are modestly higher for the first 15 participants than for the last 15 participants within the Rating treatment ($\Delta = 0.11$, $p < 0.05$). Taken together, these results reinforce the conclusion that the effectiveness of ratings depends fundamentally on the structure of consumer tastes.

4.1.1 Redistribution of Earnings Across Consumer Types

The absence of an average effect in heterogeneous-taste markets does not mean that ratings leave outcomes unchanged. Proposition 2 implies that ratings reallocate consumers between the two quizzes, benefiting the type whose preferred quiz the ratings favor and harming the other type (Prediction 1). We test this implication directly. As stated in Section 2, this analysis was not part of the pre-analysis plan.

For every participant in the last 15 positions of the heterogeneous-taste Rating sequences who observed at least one rating for each quiz, we define the *rating tilt* as the difference between the displayed average rating of the Yellow (*30-*) quiz and that of the Blue (*50+*) quiz at the moment of entry. Because the order of young and old participants

[21] The initially low match rate in the homogeneous-taste market reflects a baseline preference for the BLUE task in the absence of informative signals. This preference persists in the heterogeneous-taste market, where the BLUE task corresponds to the *50+* quiz. As color assignments were held constant across treatments, this bias does not compromise identification.

[22] A direct test finds no evidence that ratings shift the composition of buyers by confidence. Regressing the purchase decision on confidence, the highest displayed average rating, and their interaction—for participants in the homogeneous-taste Rating treatment who faced at least one rated quiz—yields an insignificant interaction (coefficient -0.15, standard error 0.27, $p = 0.59$, $N = 397$).

within a sequence is random and the age composition of every sequence is fixed at 15–15, the tilt a participant faces is unrelated to her own type. Table 4 reports OLS regressions of earnings, match, and purchase on the tilt, an indicator for being young, and their interaction.

Table 4: Rating tilt and the redistribution of earnings across consumer types

Dependent Variable:	Earnings (£)	Match (%)	Buy (%)
Model:	(1)	(2)	(3)
Constant	2.871***	0.447***	0.751***
	(0.245)	(0.131)	(0.117)
Tilt toward Yellow	−0.187	−0.268*	0.013
	(0.218)	(0.116)	(0.070)
Age group = Young	0.531***	0.037	0.087
	(0.124)	(0.085)	(0.062)
Tilt × Young	0.961**	0.785***	−0.046
	(0.320)	(0.188)	(0.099)
Tilt, young participants	0.774***	0.517***	−0.033
	(0.209)	(0.108)	(0.075)
Gender	✓	✓	✓
Risk Aversion	✓	✓	✓
Confidence	✓	✓	✓
Observations	207	165	207
R^2	0.102	0.196	0.074

Results of OLS regressions on the sample of the last 15 participants in the heterogeneous-taste Rating sequences who observed at least one rating for each quiz. Tilt toward Yellow is the difference between the displayed average ratings of the Yellow (*30-*) and Blue (*50+*) quizzes at the moment the participant entered the market. Match is defined for purchasers only. The row "Tilt, young participants" reports the linear combination of Tilt and Tilt × Young. Controls include gender, risk aversion, and self-assessed confidence in the task. Standard errors, in parentheses, use the CR2 bias-reduced cluster-robust estimator with Satterthwaite degrees-of-freedom adjustment (Pustejovsky and Tipton, 2018), clustered at the independent-sequence level. Significantly different from zero at 1% (***), 5% (**), 10% (*).

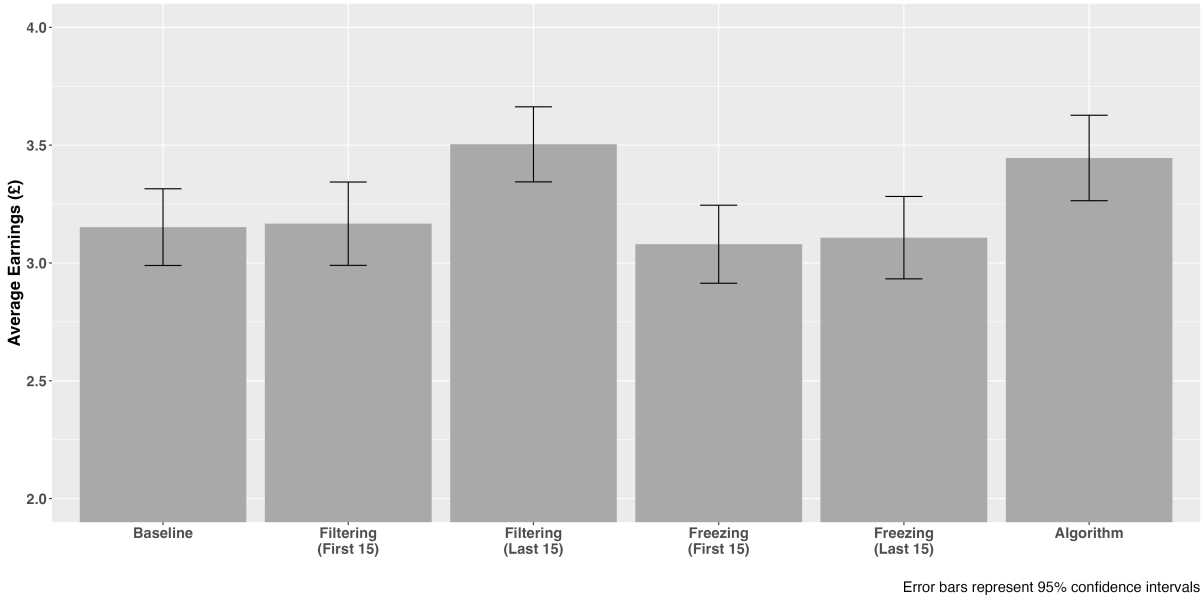
The results confirm the prediction. A one-point increase in the tilt toward the Yellow quiz raises the earnings of young participants by £0.77 ($p < 0.01$) while the point estimate for old participants is negative and imprecise (−£0.19, $p = 0.42$); the difference between the two responses is £0.96 and statistically significant ($p < 0.05$). The match-rate results in model (2) show the mechanism: the same one-point tilt raises the probability that a young participant selects her age-specific quiz by 52 percentage points ($p < 0.01$) and lowers this probability for old participants by 27 percentage points, although the latter effect is only marginally significant ($p = 0.06$). Purchase rates in model (3) do not respond to the tilt, so the reallocation operates through the choice between quizzes rather than through participation.

Together with the herding evidence in Section 4.4, these results show that in heterogeneous-taste markets ratings are not merely uninformative. Each market tips toward one quiz, and the direction of the tip determines which consumer type gains relative to the other. Average welfare is unchanged, but individual welfare depends on the early rating history of the market a consumer happens to enter.

4.2 Earnings: Policy Treatments

We now present the earnings results for the policy treatments in the heterogeneous-taste market. Figure 6 shows the average earnings across these treatments, while Table 5 reports the regression analyses for earnings, match rate, and buy rate relative to Baseline. Below, we discuss each treatment individually.

Figure 6: Average earnings in policy treatments



Error bars are 95% confidence intervals

Table 5: Treatment effects on earnings, match rate, and buy rate

Dependent Variable:	Earnings (£)	Match (%)	Buy (%)
Model:	(1)	(2)	(3)
Constant	3.004***	0.529***	0.846***
	(0.129)	(0.059)	(0.044)
Filtering first 15	0.021	0.004	-0.021
	(0.124)	(0.057)	(0.042)
Filtering last 15	0.347**	0.187***	0.065
	(0.137)	(0.051)	(0.040)
Freezing first 15	-0.072	-0.040	-0.025
	(0.114)	(0.050)	(0.044)
Freezing last 15	-0.046	0.047	-0.043
	(0.141)	(0.055)	(0.041)
Algorithm	0.309**	0.429***	0.080**
	(0.125)	(0.046)	(0.040)
Gender = Male	0.006	0.056*	-0.116***
	(0.075)	(0.031)	(0.025)
Risk aversion (0-10)	0.020	0.002	-0.029***
	(0.013)	(0.007)	(0.004)
Confidence (0-100)	0.001	-0.002***	0.002***
	(0.001)	(0.001)	(0.001)
Gender	✓	✓	✓
Risk Aversion	✓	✓	✓
Confidence	✓	✓	✓
Observations	1,238	958	1,238
R^2	0.020	0.125	0.073
Adjusted R^2	0.013	0.118	0.067

Results of OLS regressions. Standard errors use the CR2 bias-reduced cluster-robust estimator with Satterthwaite degrees-of-freedom adjustment (Pustejovsky and Tipton, 2018), clustered at the individual level for the Algorithm treatment and at the independent sequence level for the Freezing and Filtering treatments. Controls include gender, risk aversion, and self-assessed confidence in the task. Significantly different from zero at 1% (***), 5% (**), 10% (*).

Filtering

Filtering ratings by age group significantly improved earnings compared to the Baseline treatment. More importantly, the last 15 participants in each sequence of the Filtering treatment earned 13% more than those in the Rating treatment in heterogeneous-taste markets ($p = 0.01$).[23] These findings provide empirical support for **Hypothesis H3**

[23]Direct comparisons between treatments are conducted using OLS regressions with clustered standard errors. The specifications include individual controls for risk preferences, confidence, age group, and gender.

(average earnings in heterogeneous-taste markets with filtering-enhanced ratings are significantly higher than without).

Filtering works by identifying subpopulations with similar preferences and displaying the most relevant ratings. However, filtering entails a trade-off: fewer "relevant" ratings are displayed, which could slow the learning process. On the other hand, filtered ratings may carry greater perceived credibility, prompting participants to update their beliefs more quickly. To assess the net effect of filtering, we use the fact that average earnings in cases of task match and mismatch are statistically similar between homogeneous- and heterogeneous-taste markets.[24] Thus, if standard ratings increase earnings faster due to a larger number of "relevant" ratings, we would expect earnings improvements to emerge earlier in the sequence. However, we find the opposite: average earnings in the Filtering treatment for participants in positions 15–30 were 18% higher than in the homogeneous-taste market Rating treatment for participants in positions 8–15 ($p = 0.02$), reinforcing the welfare advantage of Filtering over Rating.These findings suggest that filtering not only makes ratings more informative but also increases their perceived credibility, and both channels contribute to the welfare gains. Despite providing a clearer signal about which quiz to buy, filtering had only a marginally significant positive effect on the buying rate.

Freezing

The *Freezing* treatment, which delayed the display of ratings until at least five reviews were accumulated, had no significant impact on earnings, match rates, or buying rates. This is in line with our expectation that delaying early ratings would not affect long-run average outcomes. We had also conjectured that freezing would stabilize initial signals and thereby reduce the variability of market outcomes across sequences. As we show in Section 4.4, this conjecture is only partly borne out: freezing removes the dependence of market outcomes on the tastes of early participants, but it does not reduce the overall dispersion of outcomes across sequences, because a new source of variation appears—which task is the first to reach the display threshold.

Algorithm

Providing participants with direct algorithmic recommendations significantly increased earnings relative to Baseline. This improvement can be attributed to both significant

[24] In Baseline (where ratings have no influence), participant earnings are statistically similar across homogeneous- and heterogeneous-taste markets, both when participants are matched ($p = 0.68$) and when they are mismatched ($p = 0.86$).

increase in the match rate, and significant increase in the buying rate. 91% of participants followed the algorithm's recommendation.[25]

Table 6 contrasts earnings (models (1)–(3)), match rate (model (4)), and buying rate (model (5)) across Filtering and Algorithm treatments.

Despite improvements relative to the Baseline, average earnings in the Algorithm treatment are not significantly different from those of the last 15 participants in the Filtering treatment. While the match rate is significantly higher in the Algorithm treatment compared to filtering, this did not translate into higher earnings. To understand why, we compare outcomes between matched and mismatched participants. Among matched participants, those in the Algorithm treatment earned significantly less than their counterparts in the Filtering treatment.

This outcome is consistent with a well-known limitation of algorithmic prediction: the selective labels problem (Kleinberg et al., 2018). The algorithm was trained on data from the Baseline treatment, which includes only participants who voluntarily selected into purchasing a quiz. In the Algorithm treatment, however, this selection margin shifts—participants who would otherwise abstain are now encouraged to purchase a quiz, even though the model lacks performance data on comparable individuals. As a result, the algorithm cannot learn when recommending not buying would be optimal. As shown in Table 5, algorithmic recommendations significantly increase the share of participants who purchase a quiz, particularly among those who, based on Baseline data, would have rationally abstained.[26] A more appropriate design would retrain the algorithm on data from the Algorithm treatment itself, allowing it to adjust to the new selection dynamics and incorporate abstention as a valid recommendation. Nonetheless, despite this limitation, the Algorithm treatment delivers welfare gains comparable to the Filtering treatment. This suggests that the practical relevance of the selective labels problem may be limited in our context, echoing recent findings in the hiring literature, where similar concerns have yielded modest effects (Dargnies et al., 2024b).

Buying-rate comparisons are consistent with this interpretation. The share purchasing a quiz is 83.6% in the Algorithm treatment and 81.9% among the last 15 participants of the Filtering treatment, both above the Baseline rate of 75.4%. The increase relative to Baseline is statistically significant ($p < 0.01$), while the difference between Algorithm and Filtering is not statistically significant ($p = 0.65$). Accordingly, the data do not support

[25]To assess whether followers differ systematically from non-followers, we estimate a linear probability model with observables (gender, risk aversion, and confidence) and test for joint significance. A Wald test in a linear probability model controlling for age, gender, risk aversion, and confidence yields no joint significance ($p = 0.77$), indicating no observable differences between followers and non-followers.

[26]Our data do not allow a sharp test of this mechanism. When we estimate purchase propensities from the Baseline treatment and compare matched buyers in the bottom quartile of estimated propensity—those most likely to have been induced to purchase by the recommendation—with the remaining matched buyers in the Algorithm treatment, their earnings are not significantly lower ($p = 0.39$). The selective-labels interpretation is therefore consistent with, but not established by, our data.

the claim that Algorithm induces significantly more buying than Filtering. Instead, the evidence indicates that both treatments encourage purchase relative to Baseline, but only Algorithm weakens earnings conditional on a successful match.

These findings suggest that both Filtering and Algorithm provide participants with additional information that encourages them to purchase the quiz. In the Algorithm treatment, however, part of this additional buying is unprofitable, as participants tend to earn less from the matched quiz. This is likely because the characteristics of the quiz remain opaque and are embedded in the recommendation, whereas they are more transparent and salient in the Filtering treatment.

Table 6: Algorithm vs. Filtering (last 15 participants)

Dependent Variable:	Earnings (£)			Match (%)	Buy (%)
	(All)	(Match)	(Mismatch)	(All)	
Model:	(1)	(2)	(3)	(4)	(5)
Constant	3.388***	3.952***	2.426***	0.678***	0.910***
	(0.237)	(0.264)	(0.358)	(0.058)	(0.059)
Algorithm	-0.019	-0.403**	-0.380	0.250***	0.017
	(0.146)	(0.159)	(0.298)	(0.040)	(0.037)
Gender = Male	-0.162	-0.039	-0.320	0.026	-0.084**
	(0.135)	(0.159)	(0.254)	(0.046)	(0.041)
Risk aversion (0-10)	0.041*	0.067**	0.082	0.009	-0.022***
	(0.023)	(0.028)	(0.052)	(0.007)	(0.007)
Confidence (0-100)	-0.001	-0.001	0.002	-0.001	0.001
	(0.003)	(0.004)	(0.004)	(0.001)	(0.001)
Gender	✓	✓	✓	✓	✓
Risk Aversion	✓	✓	✓	✓	✓
Confidence	✓	✓	✓	✓	✓
Observations	409	267	73	340	409
R^2	0.015	0.060	0.103	0.104	0.040
Adjusted R^2	0.005	0.045	0.050	0.093	0.030

Results of OLS regressions. Standard errors use the CR2 bias-reduced cluster-robust estimator with Satterthwaite degrees-of-freedom adjustment (Pustejovsky and Tipton, 2018), clustered at the individual level for the Algorithm treatment and at the independent sequence level for the Filtering treatment. Controls include gender, risk aversion, and self-assessed confidence in the task. Significantly different from zero at 1% (***), 5% (**), 10% (*).

4.3 Individual Ratings

We examine individual-level rating behavior along both the extensive and intensive margins. On the extensive margin, we study the decision to rate; on the intensive margin, we analyze the determinants of rating generosity. We further assess the extent to which these ratings reflect objective performance and individual characteristics. We then explore how ratings are incorporated in participants' purchase decisions.

Rating Provision (Extensive Margin). Among participants who purchased a quiz, 97% provided a rating. This near-universal take-up contrasts with the lower and more selective participation observed in most digital platforms, where rating behavior is endogenous and potentially biased (Hu et al. (2009); Lafky (2014)).

Table A2 in the Supplemental Appendix estimates the probability of providing a rating conditional on demographic and behavioral covariates. Two patterns emerge. First, rating provision increases with realized earnings: participants who earn more are significantly more likely to provide feedback.[27] This suggests that engagement is driven by task performance. Second, male participants are more likely to rate than female participants. In contrast, prior ratings of the selected task (average or number) at the time of purchase as well as individual characteristics such as confidence or risk preferences do not significantly influence the decision to rate.

Rating Generosity (Intensive Margin). Conditional on submitting a rating, we observe considerable heterogeneity in rating generosity. The average rating is 4.07 out of 5, with a variance of 1.2. As shown in Figure 7, the distribution exhibits a strong right skew, with the highest rating (5 stars) most frequently selected—a pattern consistent with prior evidence from online markets (Cabral and Hortacsu, 2010; Tadelis, 2016). However, the distribution is less polarized than the J-shape documented in Hu et al. (2009), likely due to the near-universal participation which mitigates selection on extreme views.

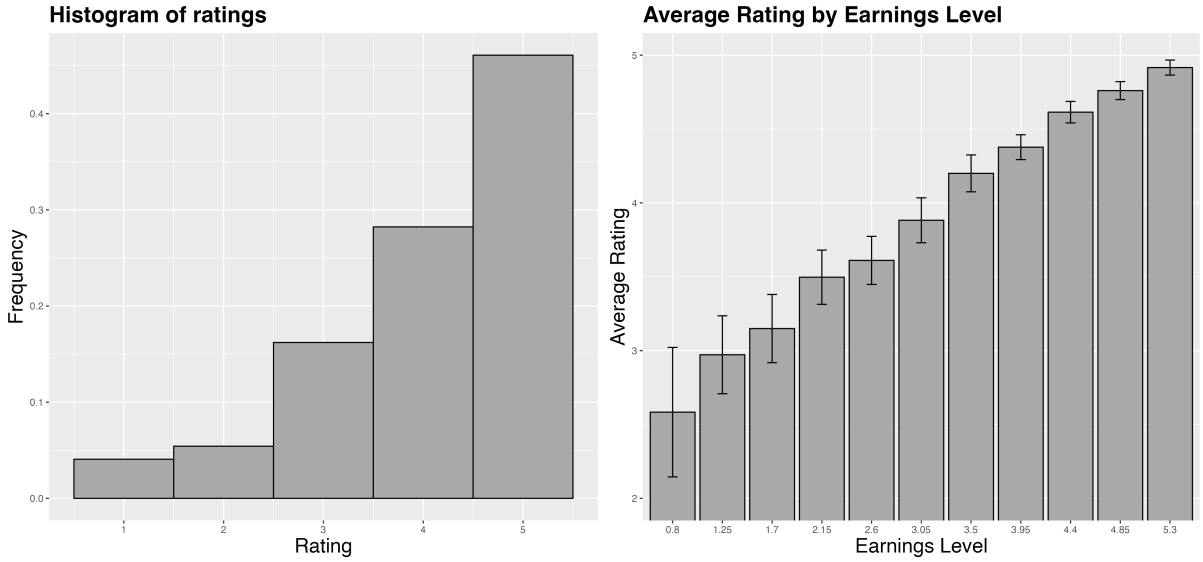
To quantify the determinants of rating generosity, we estimate OLS models of rating scores as a function of earnings and participant characteristics. The results, presented in Table A3 in the Supplemental Appendix, reveal several robust patterns.

First, each additional pound earned increases the average rating by about 0.5 points, indicating that participants evaluate quiz quality primarily through realized earnings. Second, conditional on earnings, participants matched to their age-specific quiz give higher ratings in the pooled sample and in heterogeneous-taste treatments. This effect does not appear in homogeneous-taste treatments, as expected, since match status is not

[27] This contrasts with the U-shaped pattern commonly documented in the literature, where both low and high earners are more likely to rate. In our case, lower earners are the least likely to provide a rating, resulting in a monotonic, upward-sloping relationship between earnings and rating provision.

type-specific there. Third, more risk-averse participants give lower ratings in the pooled sample and in heterogeneous-taste treatments, with no significant effect in homogeneous-taste treatments. Fourth, participants under 30 assign significantly lower ratings than those over 50, even after controlling for earnings and match status.

Figure 7: Histogram of ratings and average rating by earnings (£)



Error bars are 95% confidence intervals

Taken together, these findings indicate that participant ratings are driven by earnings but also by idiosyncratic components such as being matched, risk aversion and age.

4.3.1 Incorporating Ratings into Purchasing Decisions

We begin by analyzing the factors influencing participants' decisions to select the outside option instead of purchasing one of the two available quizzes. Across all treatments, 23% of participants chose the outside option. This rate remains largely stable across conditions, except for a marginally significant decrease in the Filtering treatment and a significant decrease in the Algorithm treatment. To identify the drivers of this choice, we estimate OLS models reported in Table A4 in the Supplemental Appendix. Participants who were less confident and more risk-averse were significantly more likely to opt for the safe outside option. By contrast, neither the presence of ratings, the number of ratings, nor the average rating of the highest-rated quiz significantly predicted opting out.

Turning to participants who did purchase a quiz, we examine how ratings influenced their choice between the Blue and Yellow options. In the absence of displayed ratings—that is, in the Baseline treatment or among participants in the other treatments who entered before any rating was displayed—Blue was chosen 66.7% of the time, reveal-

ing an underlying preference for that quiz.[28] However, once ratings became available, participants systematically adjusted their decisions in response.

When only one quiz was rated, participants responded to the rating's valence. If Blue had a high average rating (4 or above), 78% of participants chose it; this dropped to 61% when its rating was below 4. Similarly, when only Yellow was rated, a high rating led 63% of participants to switch to Yellow, while a low rating led only 27% to choose it.

When both quizzes had at least one rating, participants reliably chose the higher-rated option: 70.5% selected the quiz with the higher average rating. This behavior was associated with significantly better outcomes. Participants who followed the higher-rated option earned, on average, £3.69, compared to £2.94 for those who chose against it. Moreover, 39% of non-followers earned less than the guaranteed outside option (£2.50), compared to just 18% of those who followed the higher-rated quiz. These findings indicate that disregarding rating information leads to costly mistakes.

Breaking down the analysis by consumer tastes, we find that compliance rates are similar in homogeneous- and heterogeneous-taste Ratings treatments. However, the consequences of non-compliance differ sharply. In heterogeneous-taste markets, deviating from the higher-rated option does not entail a significant earnings loss ($p = 0.52$). By contrast, in homogeneous-taste markets, non-compliance is highly detrimental: earnings differ substantially between followers and non-followers ($p = 0.00$). Ratings therefore translate into welfare gains only when they convey information about a universally superior option; in heterogeneous-taste markets, following or disregarding ratings has limited payoff consequences on average.

To formally assess the effect of ratings on quiz choices, we estimate four OLS specifications reported in Table 7. The dependent variable is an indicator for choosing Blue, conditional on purchasing a quiz; the sample is restricted to decisions in which both quizzes had at least one rating, so that participants face a genuine rating comparison.[29] Because neither the mean nor the precision of subjects' prior beliefs is directly observed, we do not impose a fully structural Bayesian updating model. Instead, we estimate reduced-form choice models in which the effect of the displayed mean ratings of the two quizzes are allowed to depend on the number of reviews. This provides a flexible test of whether participants place greater weight on ratings when those ratings are supported by more information.

[28] Throughout this section, we classify rating information by what participants actually saw on the decision screen: in the Freezing treatment, ratings below the five-review threshold were not displayed and are treated as absent, and in the Filtering treatment we use the age-specific ratings shown to the participant.

[29] The estimations pool observations from the Rating Homogeneous, Rating Heterogeneous, Filtering, and Freezing treatments. In the Filtering treatment, only ratings relevant to each subgroup are included. The estimated coefficients are similar in both sign and magnitude when the regressions are estimated separately by treatment or on the pooled sample.

Table 7: Effect of ratings on quiz choice

Dependent Variable:	Task Selected = Blue (%)			
Model:	(1)	(2)	(3)	(4)
Constant	0.478***	0.485***	0.464***	0.534***
	(0.079)	(0.076)	(0.072)	(0.094)
Rating gap (Blue − Yellow)	0.146***	0.134***	0.137***	0.095***
	(0.023)	(0.024)	(0.021)	(0.036)
Volume gap (Blue − Yellow)		0.008**	0.014***	0.011***
		(0.004)	(0.003)	(0.003)
Rating gap (Blue − Yellow) × Volume gap (Blue − Yellow)			0.011***	0.012***
			(0.004)	(0.004)
Volume (Blue + Yellow)				-0.005
				(0.006)
Rating gap (Blue − Yellow) × Volume (Blue + Yellow)				0.006
				(0.004)
Controls	✓	✓	✓	✓
Treatment fixed eff.	✓	✓	✓	✓
Observations	827	827	827	827
R^2	0.115	0.122	0.131	0.139
Adjusted R^2	0.106	0.112	0.120	0.127

Results of OLS regressions of an indicator for choosing the Blue quiz on rating regressors, estimated on the sample of purchasers facing two rated quizzes. Rating gap is the difference between the mean ratings of the Blue and Yellow quizzes: $\bar{r}_{\text{Blue}} - \bar{r}_{\text{Yellow}}$. Volume gap is the difference in the number of reviews: $n_{\text{Blue}} - n_{\text{Yellow}}$. Column (3) adds the interaction between the rating gap and the volume gap, testing whether participants place more weight on ratings when the better-rated quiz is also supported by a larger number of reviews. Column (4) additionally controls for the total number of reviews across quizzes and interacts it with the rating gap, testing whether ratings become more influential when more review information is available overall. Controls (gender, age group, risk aversion, and self-assessed task confidence) and treatment fixed effects are included in all specifications but omitted from display. Standard errors, in parentheses, use CR2 small-sample cluster-robust corrections at the independent-sequence level Pustejovsky and Tipton (2018). Significance levels are based on Satterthwaite-adjusted inference. Significance: 1% (***), 5% (**), 10% (*).

Model (1) shows that a one-point increase in the *rating gap*, defined as the difference between the mean ratings of the Blue and Yellow quizzes ($\bar{r}_{\text{Blue}} - \bar{r}_{\text{Yellow}}$), increases the probability of choosing Blue by 14.6 percentage points ($p < 0.001$).

Model (2) augments the specification by including the *volume gap*, defined as the difference between the number of ratings of the two quizzes ($n_{\text{Blue}} - n_{\text{Yellow}}$). Conditional on the rating gap, participants are also more likely to choose Blue when it has accumulated more ratings than Yellow: a one-review increase in the relative review count for Blue increases the probability of choosing Blue by 0.8 percentage points ($\hat{\beta} = 0.008$, $p < 0.05$).

This suggests that participants respond not only to the relative valence of ratings, but also to the relative volume of rating information.

Model (3) tests a reduced-form implication of precision-sensitive updating by allowing the rating gap to matter more when the same option also has a larger review-count advantage. The interaction term is positive and statistically significant ($\hat{\beta} = 0.011$, $p < 0.01$), indicating that a one-point rating advantage shifts choices more strongly toward Blue when Blue is also the more-reviewed option. This pattern is consistent with participants placing greater weight on the more strongly supported rating signal, as the Bayesian updating formulas in equations (1) and (2) imply.

Finally, model (4) additionally controls for the total number of reviews across both quizzes and interacts it with the rating gap. While the interaction between the rating gap and volume gap remains positive and statistically significant, neither the total number of reviews nor its interaction with the rating gap is statistically significant. This suggests that participants respond primarily to the difference between the number of reviews of the two quizzes rather than to the number of reviews available overall. As a robustness check, Table A5 in the Supplemental Appendix re-estimates all four specifications using a probit model. The resulting average marginal effects are very similar to the linear-probability estimates, and the pattern of statistical significance is preserved, except that the final interaction becomes significant at the $p < 0.05$ threshold.

In Filtering, we consider two alternative patterns of choice: (i) selecting the task with the higher rating within one's age group, and (ii) selecting the task with the higher rating pooled across both age groups. Among purchasers who observed at least one age-specific rating for each task and for whom the two age-specific ratings differed, 66.2% followed the age-specific ratings, earning £4.09 on average; the earnings of those who did not average £2.73. Among the non-followers whose pooled ratings also differed, 58% selected the task with the higher pooled rating and 42% chose the alternative that was inferior under both criteria; the earnings of these two groups do not differ at the 5% level (£2.53 vs. £3.03, $p = 0.07$). These results indicate that filtering is effective when participants use it, while deviations from age-specific signals are associated with markedly lower earnings.

4.4 Ratings and Herding Behavior

Ratings not only guide individual decisions but also shape collective dynamics by triggering herding behavior—that is, systematic convergence in choices based on publicly visible signals. While this can help coordinate participants on better options, it may also amplify early noise and lead to inefficient lock-in. In digital marketplaces, for instance, a few unfavorable early reviews can suppress demand for high-quality new products, while unwarranted early enthusiasm can entrench inferior ones. Such dynamics are especially

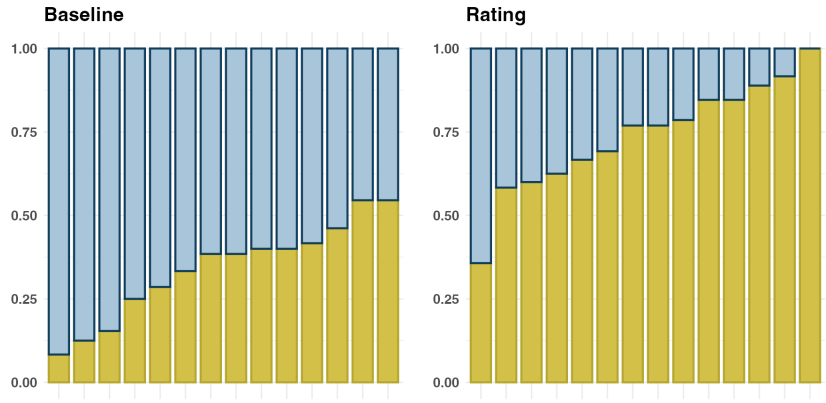
consequential for newcomers, whose success may depend on how a few early participants rate their offering.

In our experiment, we explore these dynamics by comparing markets with and without ratings. As we have seen, in homogeneous-taste markets, where one option is objectively superior, ratings consistently steer participants toward the better Yellow task (*easy* quiz) and the inferior Blue task (*hard* quiz) is gradually abandoned. In this case, herding facilitates efficient convergence, as quality is shared across participants and ratings aggregate meaningful information. We next examine whether these outcomes remain stable and consistent across independent homogeneous-taste market sequences, contrasting them with the fragmented dynamics seen in heterogeneous-taste market sequences.

To assess the consistency of these dynamics, we examine the 14 independent rating sequences in homogeneous-taste markets. Each sequence evolved separately, generating distinct rating trajectories and subsequent choice patterns. As a benchmark, we use synthetic sequences built from participants in the homogeneous-taste Baseline treatment, who did not observe ratings. Specifically, we construct 14 synthetic baseline sequences by randomly drawing 15 task choices from the homogeneous-taste Baseline pool for each sequence, matching the number of late-sequence observations used for the Rating treatment, and compute task-choice shares within each group. Draws are made independently and with replacement for each synthetic sequence, so a Baseline observation can appear in more than one sequence. For graphical presentation, these synthetic baseline groups are ordered by the final share choosing the Yellow task.

Figure 8 reports the proportions choosing the Yellow task (*easy* quiz) and Blue task (*hard* quiz) across sequences in the Baseline and Rating treatments. In the synthetic baseline sequences (left panel), Blue often dominates, consistent with its greater appeal when ratings are absent. In contrast, in the rating sequences (right panel), Yellow tends to dominate: in 13 of the 14 sequences, it captures at least 50% of final task selections. This aggregate convergence suggests that ratings generate a robust shift toward the higher-quality option in homogeneous-taste markets.

Figure 8: Proportions of Yellow task (*easy* quiz) and Blue task (*hard* quiz) selections by sequence and treatment in homogeneous-taste markets



Baseline sequences are constructed by independently drawing 15 task choices, with replacement, from the corresponding Baseline pool to populate each synthetic sequence. Rating sequences report the choices of the last 15 participants for each sequence. Dark yellow indicates the share of participants who selected the Yellow task and light blue the share of participants who selected the Blue task.

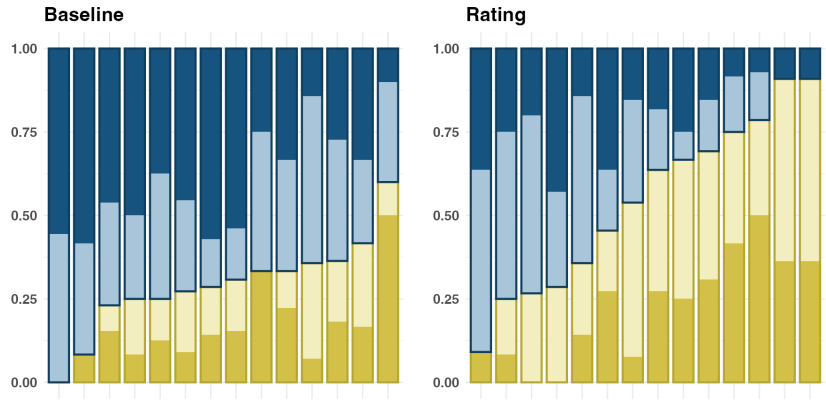
We then turn to heterogeneous-taste markets, where task appeal depends on individuals' age group, and no option is universally superior. Figure 9 displays the proportions of Yellow task (*30-* quiz) and Blue task (*50+* quiz) selections by young and old participants by sequence across Baseline and Rating treatments in heterogeneous-taste markets. In the synthetic baseline groups, constructed by reshuffling Baseline participants into artificial sequences, task choices remain relatively stable, with Blue typically favored. By contrast, sequences in the Rating treatment exhibit substantial heterogeneity: while some groups converge almost entirely on Blue, others converge on Yellow.

This divergence reflects herding in the absence of a common quality standard. Ratings increase the likelihood that one task becomes dominant within a given group, leading to endogenous coordination even when underlying product quality is symmetric. However, unlike in homogeneous-taste markets, such convergence does not improve welfare. Instead, it may result in arbitrary outcomes shaped by the preferences of early participants rather than the intrinsic match between consumers and products.

The visual impression of divergence can be quantified. The standard deviation of the Yellow-task share among the last 15 participants across the 14 heterogeneous-taste Rating sequences is 0.26. If choices were unaffected by the rating history, as in the Baseline treatment, this dispersion would be far smaller: drawing groups of the same size at random from Baseline choices yields an average standard deviation of 0.13, and none of 3,000 such draws reaches the observed value ($p < 0.001$). The corresponding dispersion across homogeneous-taste Rating sequences is 0.17 ($p = 0.10$ against the same benchmark), reflecting that these markets also converge, but almost always in the same direction. Heterogeneous-taste markets with ratings are thus over-dispersed relative to

markets without ratings: ratings make each market converge, but the direction of convergence varies across replications of identical environments.

Figure 9: Proportions of Yellow task (*30-* quiz) and Blue task (*50+* quiz) selections by young and old participants by sequence in heterogeneous-taste markets for main treatments



Baseline sequences are constructed by independently drawing 15 task choices, with replacement, from the corresponding Baseline pool to populate each synthetic sequence. Rating sequences report the choices of the last 15 participants for each sequence. Dark yellow indicates the share of young participants who selected the Yellow task, light yellow the share of old who selected the Yellow task, light blue the share of young who selected the Blue task, and dark blue the share of old who selected the Blue task.

To test whether early participants' age shapes the task chosen by later participants, we analyze the relationship between the proportion of young participants among the first 15 in a sequence and the Yellow task choices of the last 15 participants. Recall that the Yellow task is tailored to participants below 30 years old. Table 8 reports results separately for the Rating, Freezing, and Filtering treatments in heterogeneous tastes markets. In Rating (Model 1), we observe a positive association: sequences with more young participants early on are more likely to converge on the Yellow task. This suggests that rating dynamics amplify the preferences of early movers, leading later participants to disproportionately select options aligned with those early preferences. Both the Freezing and Filtering treatments (Models 2-3) reduce the influence of early composition on later task choices. This is consistent with our design: in Freezing, early ratings are hidden until at least five observations are available, reducing the leverage of small early samples; in Filtering, participants effectively operate in a homogeneous taste submarkets where ratings convey primarily quality information rather than preference heterogeneity.

Table 8: Individual task choice of last 15 participants

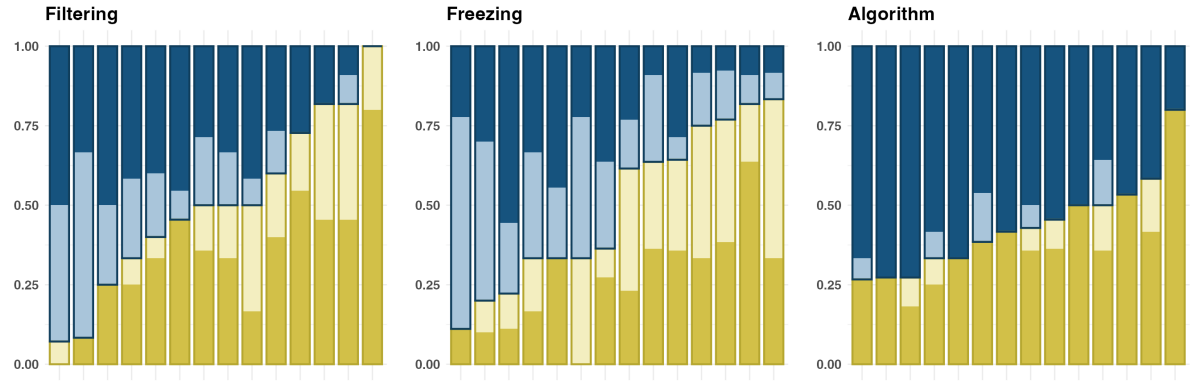
Dependent Variable:	Task Selected = Yellow (%)		
	Rating	Freezing	Filtering
Model:	(1)	(2)	(3)
Constant	-1.491**	-0.099	0.461
	(0.646)	(0.769)	(0.707)
Young (First 15) %	2.603**	0.071	-0.904
	(1.115)	(1.345)	(1.261)
Confidence (0-100)	0.001	0.000	0.001
	(0.004)	(0.004)	(0.004)
Risk aversion (0-10)	0.032	0.022	-0.031
	(0.046)	(0.044)	(0.041)
Gender = Male	0.154	0.120	0.149
	(0.209)	(0.226)	(0.209)
Observations	166	155	172
Squared Correlation	0.044	0.003	0.011
Pseudo R^2	0.032	0.002	0.008
BIC	246.8	239.3	262.2

Results of Probit regressions where the dependent variable is an indicator for choosing the Yellow task. The key explanatory variable is the sequence-specific share of young participants among the first 15. Controls include individual confidence, risk preferences, and gender. Significantly different from zero at 1% (***), 5% (**), 10% (*).

Figure 10 displays the proportions of Yellow task (*30-* quiz) and Blue task (*50+* quiz) selections by young and old participants by sequence across policy treatments in heterogeneous-taste markets. Filtering and Algorithm steer young participants to the Yellow task and old participants to the Blue task, lowering the proportions of mismatches compared to the Baseline and Rating treatments. This is particularly true of the Algorithm treatment, which reduces the mismatch rate to nearly zero. By contrast, Filtering still displays considerable heterogeneity and mismatches across sequences. However, this difference does not drive welfare down.

This apparent paradox can be explained by the dynamics of informational revelation. In Freezing, later participants disproportionately select the task that first reaches the five-rating threshold, granting it an informational advantage. Regression results reported in Table A6 in the Supplemental Appendix confirm that when the Yellow task is the first to unfreeze, participants in the second half of the sequence are significantly more likely to select it. As a result, dominance in Freezing is shaped less by early demographics and more by the accident of which task unfreezes first. If both tasks were unfrozen simultaneously, we would expect considerably less variation in task selection across groups.

Figure 10: Proportions of Yellow task (*30-* quiz) and Blue task (*50+* quiz) selections by young and old participants by sequence in heterogeneous-taste markets for policy treatments



Filtering and Freezing sequences report the task choices of the last 15 participants for each sequence. Algorithm sequences are constructed by independently drawing 15 task choices, with replacement, from the Algorithm pool to populate each synthetic sequence. Dark yellow indicates the share of young participants who selected the Yellow task, light yellow the share of old who selected the Yellow task, light blue the share of young who selected the Blue task, and dark blue the share of old who selected the Blue task.

We conclude this section by summarizing its main insights: (1) in homogeneous-taste markets, ratings consistently induce herding toward the higher-quality option, which attracts at least half of final purchases in 13 of the 14 sequences; (2) in heterogeneous-taste markets, where individual preferences dominate, ratings amplify early idiosyncrasies and increase the variability of later choices, and the direction of convergence depends on the composition of early participants; and (3) the Freezing and Filtering treatments weaken the influence of early participants' tastes on subsequent decisions, but neither reduces the overall variability of outcomes across sequences—in Freezing, this variability instead reflects which task is the first to reach the display threshold.

5 Conclusion

This paper causally evaluates the impact of rating systems on consumer welfare across different structures of consumer tastes. We show that while ratings effectively guide consumer choices in homogeneous-taste markets—where product quality is objectively ranked—they fail to enhance welfare in heterogeneous-taste settings, where preferences differ across individuals. Aggregate ratings in heterogeneous-taste markets collapse heterogeneous experiences into an uninformative signal, leading to suboptimal consumer-product matches. Although average welfare is unchanged in these markets, ratings redistribute earnings across consumer types: the type whose preferred product is favored by the early rating history gains relative to the other type.

To address this limitation, we evaluate alternative mechanisms. Filtered ratings, segmented by observable characteristics—and algorithmic recommendations, derived from prior performance patterns—both significantly improve welfare in heterogeneous-taste markets. By contrast, delaying the availability of early ratings does not yield welfare gains, suggesting that inefficiencies arise primarily from preference heterogeneity, not early-stage noise. Consumer welfare gains are obtained mostly via improved match rates.

Finally, using a design with independent market replications, we examine the dynamics of sequential decision-making. We find that early participant characteristics can shape later choices, introducing path dependence even in the absence of informative early ratings. Market outcomes under ratings are therefore sensitive to initial conditions in heterogeneous-taste markets, which strengthens the case for type-specific information.

The experimental design should be interpreted as a clean benchmark rather than a direct calibration of rating effects in field markets. We deliberately study two polar environments: one in which consumers share a common ranking of products, and one in which product fit is type-specific. This separation allows us to isolate the informational role of ratings, but it abstracts from markets in which common quality and idiosyncratic fit coexist. Similarly, participants receive only limited prior information about the two quizzes before observing ratings. In many field settings, consumers may also observe prices, brands, product descriptions, seller identity, or other signals, so the welfare gains from ratings in our experiment may overstate the magnitude of rating effects in environments with richer prior information. Rating provision is also unusually high in our setting: 97% of participants who completed a quiz submitted a rating. This reflects the fact that the rating prompt was immediate, costless, and embedded in the experimental task, and therefore our design is not suited to measuring rating-supply selection in settings where leaving feedback is costly or delayed. Finally, the heterogeneous-taste treatment uses age as a single, observable source of preference heterogeneity. This makes the mechanism transparent and allows filtered ratings to be implemented cleanly, but real-world preferences are often multidimensional. In such environments, simple demographic filters may be less effective, while algorithms that combine multiple behavioral and demographic signals may become more valuable, albeit at the cost of greater opacity.

Subject to these scope conditions, our findings show that the effectiveness of online rating systems depends on the structure of consumer tastes. Aggregation rules that work well when consumers agree on quality can fail when they do not, and market outcomes then depend on early rating history rather than on product characteristics. Simple design changes—displaying ratings by consumer type, or replacing them with recommendations based on observable characteristics—recover most of the welfare gains that aggregate ratings deliver in markets with homogeneous tastes.

Future research could explore additional mechanisms to improve rating informativeness, particularly in markets where preference heterogeneity is multi-dimensional. Multidimensional preferences may make simple filters less applicable and may make algorithms relatively more useful, provided that algorithms have sufficiently rich training data and can account for the outside option. However, algorithmic recommendations in cases of complex, multidimensional preferences might raise concerns about transparency (Dargnies et al., 2024a) and about the strategic use of algorithms by platforms to attract consumers to potentially more profitable options, especially in the absence of a strategic response (Hagenbach and Salas, 2024; Bó et al., 2024). Whether this could trigger a backlash from consumers remains an open question.

References

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2022). Learning from reviews: The selection effect and the speed of learning. *Econometrica 90*(6), 2857–2899.

Anderson, M. and J. Magruder (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal 122*(563), 957–989.

Bar-Isaac, H., G. Caruana, and V. Cuñat (2010). Information gathering and marketing. *Journal of Economics & Management Strategy 19*(2), 375–401.

Bar-Isaac, H., S. Tadelis, et al. (2008). Seller reputation. *Foundations and Trends® in Microeconomics 4*(4), 273–351.

Bénabou, R. and N. Vellodi (2024). (pro-) social learning and strategic disclosure. Technical report, National Bureau of Economic Research.

Benkert, J.-M. and A. Schmutzler (2024). A theory of recommendations.

Bó, I., L. Chen, and R. Hakimov (2024). Strategic responses to personalized pricing and demand for privacy: An experiment. *Games and Economic Behavior 148*, 487–516.

Bolton, G., B. Greiner, and A. Ockenfels (2013). Engineering trust: reciprocity in the production of reputation information. *Management Science 59*(2), 265–285.

Bolton, G. E., E. Katok, and A. Ockenfels (2004). How effective are electronic reputation mechanisms? an experimental investigation. *Management Science 50*(11), 1587–1602.

Burtch, G., Y. Hong, R. Bapna, and V. Griskevicius (2018). Stimulating online reviews by combining financial incentives and social norms. *Management Science 64*(5), 2065–2082.

Cabral, L. and A. Hortacsu (2010). The dynamics of seller reputation: Evidence from ebay. *The Journal of Industrial Economics 58*(1), 54–78.

Cabral, L. and L. Li (2015). A dollar for your thoughts: Feedback-conditional rebates on ebay. *Management Science 61*(9), 2052–2063.

Carnehl, C., M. Schaefer, A. Stenzel, and K. D. Tran (2022). Value for money and selection: How pricing affects airbnb ratings. *Innocenzo Gasparini Institute for Economic Research Working Paper Series*.

Charness, G., R. Oprea, and S. Yuksel (2021). How do people choose between biased information sources? evidence from a laboratory experiment. *Journal of the European Economic Association 19*, 1656–1691.

Che, Y.-K. and J. Hörner (2018). Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics 133*(2), 871–925.

Chevalier, J. A. and D. Mayzlin (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research 43*(3), 345–354.

Dargnies, M.-P., R. Hakimov, and D. Kübler (2024a). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*.

Dargnies, M.-P., R. Hakimov, and D. Kübler (2024b). Behavioral measures improve ai hiring: a field experiment. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pp. 831–832.

Dellarocas, C. and C. A. Wood (2008). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science 54*(3), 460–476.

Dendorfer, F. and R. Seibel (2024). The cost of the cold-start problem on airbnb. Technical report.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics 133*(4), 1645–1692.

Forbes (2022a). Amazon has turned 'rings of power' star ratings back on, here's how fans are scoring it. *Forbes*. Accessed: January 12, 2025.

Forbes (2022b). 'Rings of power' is getting review bombed so hard Amazon suspended reviews entirely. *Forbes*. Accessed: January 12, 2025.

Fradkin, A. and D. Holtz (2023). Do incentives to review help the market? evidence from a field experiment on airbnb. *Marketing Science 42*(5), 853–865.

Hagenbach, J. and A. Salas (2024). Strategic information disclosure to classification algorithms: An experiment.

He, S., B. Hollenbeck, and D. Proserpio (2022). The market for fake reviews. *Marketing Science 41*(5), 896–921.

Hu, N., P. A. Pavlou, and J. Zhang (2006). Can online reviews reveal a product's true quality? empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pp. 324–330.

Hu, N., P. A. Pavlou, and J. Zhang (2017). On self-selection biases in online product reviews. *MIS Quarterly 41*(2), 449–475.

Hu, N., J. Zhang, and P. A. Pavlou (2009). Overcoming the j-shaped distribution of product reviews. *Communications of the ACM 52*(10), 144–147.

Ifrach, B., C. Maglaras, M. Scarsini, and A. Zseleva (2019). Bayesian social learning from consumer reviews. *Operations Research 67*(5), 1209–1221.

Johnen, J. and R. Ng (2024). Harvesting ratings. Technical report, University of Bonn and University of Mannheim, Germany.

Klein, T. J., C. Lambertz, and K. O. Stahl (2016). Market transparency, adverse selection, and moral hazard. *Journal of Political Economy 124*(6), 1677–1713.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics 133*(1), 237–293.

Lafky, J. (2014). Why do people rate? theory and evidence on online ratings. *Games and Economic Behavior 87*, 554–570.

Lafky, J. and R. Ng (2024). Ratings with heterogeneous preferences. Technical report, University of Bonn and University of Mannheim, Germany.

Li, L., S. Tadelis, and X. Zhou (2020). Buying reputation as a signal of quality: Evidence from an online marketplace. *The RAND Journal of Economics 51*(4), 965–988.

Luca, M. (2016). Reviews, reputation, and revenue: The case of yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper* (12-016).

Luca, M. and G. Zervas (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science 62*(12), 3412–3427.

Mayzlin, D., Y. Dover, and J. Chevalier (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review 104*(8), 2421–2455.

Pustejovsky, J. E. and E. Tipton (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics 36*(4), 672–683.

Reimers, I. and J. Waldfogel (2021). Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings. *American Economic Review 111*(6), 1944–1971.

Resnick, P., R. Zeckhauser, J. Swanson, and K. Lockwood (2006). The value of reputation on ebay: A controlled experiment. *Experimental Economics 9*(2), 79–101.

Salganik, M. J., P. S. Dodds, and D. J. Watts (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science 311*(5762), 854–856.

Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics 8*(1), 321–340.

Vellodi, N. (2018). Ratings design and barriers to entry. *Available at SSRN 3267061*.