Beyond the Stars: Exploring the Welfare Effects of Ratings in Differentiated Markets*

Noah Bohren Rustamdjan Hakimov Luís Santos-Pinto

October 17, 2025

Abstract

We study the causal impact of rating systems on consumer welfare under different forms of product differentiation. A Bayesian model and a large, pre-registered, online experiment show that in vertically differentiated markets, where consumers agree on quality rankings, exposure to ratings increases welfare by improving match rates: ratings steer consumers toward the high-quality product. In contrast, in horizontally differentiated markets, where preferences vary across individuals, ratings fail to improve welfare. To address this, we show that alternative mechanisms—filtered ratings and algorithmic recommendations—causally restore consumer welfare in horizontally differentiated markets. Our findings underscore the role of rating design in its effectiveness and highlight how market structure and information aggregation shape consumer outcomes.

JEL-codes: *D83, C90*

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

*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 universally valued, a single aggregated rating may conflate idiosyncratic taste with objective 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). The conventional 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).

We examine how the welfare effects of online rating systems depend on the nature of product differentiation. Specifically, we study the limitations of standard rating systems in markets where consumer preferences vary. We develop a stylized Bayesian decision-theoretic model in which risk-neutral consumers form expectations about product quality by combining prior beliefs with observed average ratings and the number of reviews for two products before choosing between them and an outside option of not buying. The model shows that in vertically differentiated markets, where one product is objectively superior, ratings increase welfare through two channels: (i) improving match rates by guiding consumers toward the higher-quality product and (ii) increasing 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.

We then conduct a series of preregistered experiments in a controlled field setting to test the model's predictions and alternative rating designs used in practice. We simulate 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 setup mirrors a natural decision environment, as participants regularly select among tasks for monetary compensation. The tasks are celebrity quizzes,[1] designed to reflect either vertical or horizontal product differentiation. In the vertical condition,

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

one quiz is objectively easier and thus dominates in expected payoff. In the horizontal condition, the optimal quiz depends on the participant's age, generating preference heterogeneity. We designed the quizzes based on a pretest: while age is irrelevant in the vertically differentiated market, in the horizontally differentiated market participants below 30 and above 50 years old exhibit opposing preferences.[2]

A key advantage of the experimental design over field data is the ability to enforce clear and observable product differentiation, enabling unambiguous identification of the welfare effects of ratings across market structures. In naturally occurring settings, products often differ along multiple, unobserved dimensions, complicating causal inference. Our design also eliminates pricing as a confound: quiz prices are fixed, ruling out strategic pricing by sellers and its endogenous influence on both ratings and consumer selection (Carnehl et al., 2022). Moreover, the experiment allows us to control selection into both purchase and rating decisions, an important source of bias in observational studies. While lab experiments often face external-validity concerns due to artificial products or weak incentives to engage with ratings, our approach mitigates these issues. Participants are drawn from a pool of workers accustomed to paid task selection, and ratings are tied to real monetary outcomes. This setting approximates consumer behavior while preserving the advantages of experimental control.

In the Baseline treatments, we establish benchmarks for each type of product differentiation—vertical and horizontal—by observing participants' task choices without access to ratings. In the Rating treatments, participants enter the market sequentially and observe the average rating (on a 1–5 star scale) and the number of ratings submitted by earlier participants. We implement 14 independent sequences of 30 participants for each Rating treatment. Each sequence constitutes a separate market with its own path of rating development, mirroring real-world platforms where individual ratings influence future choices. Multiple independent markets also allow us to average across sequences, mitigating stochastic variation in early ratings and isolating the causal impact of the rating system.

Comparing outcomes between Rating and Baseline, we find that ratings significantly increase earnings in vertically differentiated markets but have no effect in horizontally differentiated ones. The gains in vertical markets are driven primarily by higher match rates. However, contrary to the model's prediction, there is no evidence that ratings increase the likelihood of purchasing a task.

To address the failure of standard ratings in horizontally differentiated markets, we introduce three additional treatments. In the Filtering treatment, ratings are segmented by consumer demographics, following practices on platforms such as Booking.com and TripAdvisor. 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. Our model predicts the effectiveness of this treatment when filtering splits the horizontal market into vertical submarkets.

[2]To ensure comparability across all treatments, we restricted participation to individuals below 30 and above 50 years old, maintaining a 50% share of each group in every treatment.

In the Freezing treatment, ratings are withheld until at least five reviews have accumulated for a task, mirroring practices on platforms such as the Apple App Store and Kickstarter. 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.

Consistent with the model's prediction, the Filtering treatment significantly increases earnings in horizontal 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 arises from 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 but reduces dispersion across sequences (path dependence) by limiting herding in horizontally differentiated markets.

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. This near-universal take-up suggests that classic selection biases in consumer reviews (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: 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, contrary to our prediction. 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 horizontal 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 in the presence of consumer heterogeneity. Benkert and Schmutzler (2024) develop a theoretical framework to assess the informativeness and value of ratings under heterogeneous preferences, identifying conditions under which different consumer types follow recommendations and how optimal strategies differ for consumers and platforms. Lafky and Ng (2024) demonstrate in an online experiment that raters heavily favor their own preferences when leaving a rating, and consumers' interpretations of ratings are largely aligned with the preferences of the raters. When consumers are informed of the raters' preferences, their sensitivity to ratings depends on how similar their preferences are to those of the raters. We contribute to this literature by considering rating designs that can overcome the complexities of accommodating biases and learning in horizontally differentiated markets.

We also extend the literature on the limitations of ratings by showing that their informativeness depends on the nature of product differentiation. Prior work has identified several biases that limit the effectiveness of rating systems. Several studies argue that online reviews do not always reflect true product quality due to selection biases, leading to disproportionately extreme ratings—with an over-representation of five-star and one-star reviews—creating a misleading signal for future consumers (Hu et al., 2006, 2017, 2009). Mayzlin et al. (2014); Luca and Zervas (2016); He et al. (2022) provide evidence that firms strategically manipulate ratings by posting fake reviews, further distorting informativeness. 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 use prices to harvest good ratings from early consumers (Johnen and Ng, 2024). Che and Hörner (2018) and Bénabou and Vellodi (2024) study how recommender systems influence social learning, demonstrating that algorithmic interventions can improve efficiency by mitigating biases in consumer ratings. Vellodi (2018) explore how rating systems can create barriers to entry by reinforcing incumbents' advantages. We contribute to this by showing that even unbiased ratings can fail in horizontally differentiated markets, and show the ways to redesign them to recover welfare gains of rating systems.

This study also contributes to the literature on social learning and reputation system biases. Consumers revise beliefs based on observed reviews, yet selection effects and early ratings may distort long-run outcomes (Acemoglu et al., 2022; Ifrach et al., 2019; Salganik et al., 2006). In particular, herding behavior and initially biased ratings can generate path dependence, rendering market dynamics sensitive to early realizations. Our parallel-market

experimental design demonstrates that early participants exert disproportionate influence on subsequent choices in horizontally differentiated markets. Although the Freezing policy attenuates this path dependence by delaying rating visibility, it does not lead to measurable welfare improvements.

Finally, this paper contributes to the design of alternative rating mechanisms that correct for the limitations of traditional ratings (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 horizontally differentiated 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. Appendix A contains additional tables and figures, Appendix B the proofs of the theory results, and Appendix C the experimental instructions.

2 Theoretical Model

This section presents a straightforward Bayesian decision-theoretic framework to investigate how product ratings influence welfare in vertically and horizontally differentiated markets.

2.1 Vertically Differentiated Market

Risk-neutral subjects choose between a safe outside option or one of two quiz-based tasks: an Easy quiz (E) or a Hard quiz (H), with qualities q_E and q_H where $0 < q_H < q_E < 1$. Each subject has skill $\theta \in [0, 1]$ (with distribution functions $F(\theta)$, $f(\theta)$). The probability of success on quiz k is

$$\Pr(S \mid q_k, \theta) \;=\; \theta^{\frac{1}{q_k}-1},$$

so higher q_k or θ increases success.

Each quiz k has ratings $R_1^k, \ldots, R_{n_k}^k$ drawn from $N(q_k, \sigma^2)$, and the sample mean rating is

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

Bayesian updating from prior beliefs $N(\mu, \tau^2)$ with $q_H < \mu < q_E$, yields 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 the posterior variance

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

A subject's expected payoff from taking quiz k after observing mean rating $\bar{r}_k$ is

$$a \ + \ b \int \theta^{\frac{1}{q_k}-1} g(q_k|\bar{r}_k) dq_k,$$

where $a > 0$, $b > 0$, and $g(q_k|\bar{r}_k)$ is the posterior density $N(E(q_k|\bar{r}_k), V(q_k|\bar{r}_k))$. The outside option pays $z \in (a, a+b)$. Hence, a subject with skill θ prefers quiz k over the outside option after observing mean rating $\bar{r}_k$ when

$$a + b \int \theta^{\frac{1}{q_k}-1} g(q_k|\bar{r}_k) dq_k > z.$$

Finally, a subject with skill θ prefers quiz E to H after observing mean ratings $\bar{r}_E$ and $\bar{r}_H$ when

$$\int \theta^{\frac{1}{q_E}-1} g(q_E|\bar{r}_E) dq_E > \int \theta^{\frac{1}{q_H}-1} g(q_H|\bar{r}_H) dq_H.$$

Ratings thus improve welfare by guiding skilled buyers to purchase (rather than opt out) and by revealing which quiz is higher quality.

Proposition 1: *In a vertically differentiated market, average earnings are higher with ratings than without, i.e.*

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

H1: *Average earnings in vertical markets with ratings exceed those in vertical markets without ratings.*

2.2 Horizontally Differentiated 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 (old) subjects each comprise half the population. To introduce horizontal differentiation we assume $0 < q_{YO} < q_{YY} < 1$ and $0 < q_{OY} < q_{OO} < 1$, 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

$$\Pr(S \mid q_{kj}, \theta) \ = \ \theta^{\frac{1}{q_{kj}}-1}.$$

We assume the two quizzes share the same average quality and the same overall average success rate, ensuring no vertical differentiation:

$$\frac{q_{YO} + q_{YY}}{2} = \frac{q_{OY} + q_{OO}}{2}. \tag{2}$$

A subject of type j observes ratings $R_1^{kj}, \ldots, R_{n_{kj}}^{kj} \sim N(q_{kj}, \sigma^2)$ and infers the posterior mean

$$E(q_{kj} \mid \bar{r}_{kj}) \;=\; \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n_{kj}}}\, \bar{r}_{kj} \;+\; \frac{\frac{\sigma^2}{n_{kj}}}{\frac{\sigma^2}{n_{kj}} + \tau^2}\, \mu, \tag{3}$$

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 taking quiz k by a subject of type j who observes mean rating $\bar{r}_{kj}$ is obtained in an analogous way as in the vertically differentiated market.

Proposition 2: *In a horizontally differentiated market, average earnings remain the same whether or not ratings are present, i.e.*

$$E\big[I_H(\text{Ratings})\big] \;=\; E\big[I_H(\text{Baseline})\big].$$

H2: *Average earnings in horizontal markets with ratings do not differ from those without ratings.*

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 horizontally differentiated market, average earnings with filtering exceed those without, i.e.*

$$E\big[I_H(\text{Filtering})\big] \;>\; E\big[I_H(\text{Baseline})\big].$$

H3: *Average earnings in horizontal markets with ratings plus filtering are significantly higher than in markets without filtering.*

3 Experimental Design

This section presents the experimental protocol, which received ethical approval and was pre-registered prior to data collection.[3] 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 select between two primary tasks (quizzes), representing products in either vertically or horizontally differentiated markets, and a fallback task, representing the choice of "not buying". Upon completing a primary task, participants evaluate it on a 1-to-5-star scale. The experimental design varies both the type of product differentiation (vertical or horizontal) and the availability of ratings (no ratings vs. ratings). Additionally, we introduce variations in rating and recommendation systems for horizontal markets, testing filtering, freezing, and algorithmic recommendations.

The primary objectives of the study are: (1) to evaluate the causal effects of rating systems on consumer welfare in vertically and horizontally differentiated markets,[4] (2) to compare the effectiveness of alternative rating and recommendation systems in horizontal 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.

[3]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).

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

Each participant received an initial endowment of £2.50. The primary tasks were celebrity quizzes, with participants selecting and completing one of two available quizzes. Taking a quiz required a £1.70 fee, deducted from the initial endowment.[5]

Each quiz consisted of ten multiple-choice questions about celebrities, 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.[6]

Alternatively, participants could select a fallback task, which involved counting zeros in matrices for 110 seconds. This option was free, allowing participants to retain their full initial endowment of £2.5 but without the opportunity to earn additional income. The inclusion of this outside option aligns with the theoretical model and enables 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. Before the buying stage, we collected demographic information, including age, gender, as well as a measure of participants' confidence in their quiz-taking ability. Specifically, we 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?"* This non-incentivized measure serves as a proxy for participants' confidence, which in turn helps assess their perceived likelihood of benefiting from purchasing a quiz. After the rating stage, all participants completed a non-incentivized measure of risk preferences from 0 to 10, following Falk et al. (2018). These measures allow us to analyze the determinants of purchasing behavior.

3.3 Treatment Variation 1: Type of Product Differentiation in a Market

The first key dimension of treatment variation is the type of product differentiation in a market. In some treatments, we establish a market with vertically differentiated quizzes,

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

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

while in the other treatments, we create a market with horizontally differentiated quizzes. The next two subsections detail the differences between the two types of markets.

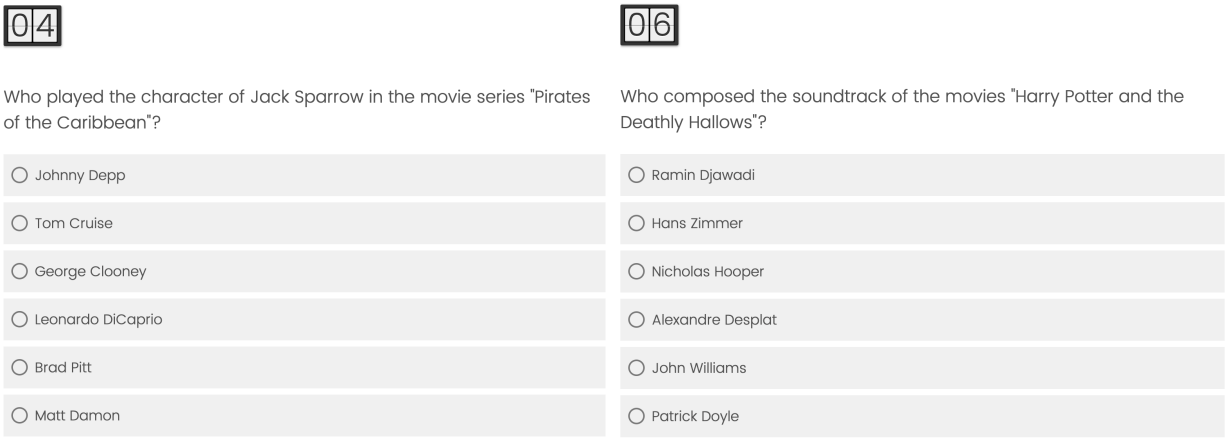
3.3.1 Vertically Differentiated Market

Two products are considered vertically differentiated when they vary in quality, and almost all consumers agree on the quality ranking. Thus, if both products were offered at the same price, all consumers would prefer the higher-quality option.

To establish a market with vertically differentiated products, we designed two celebrity 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 celebrity 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 vertical market.

Figure 1: Example of selected questions for vertical market



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

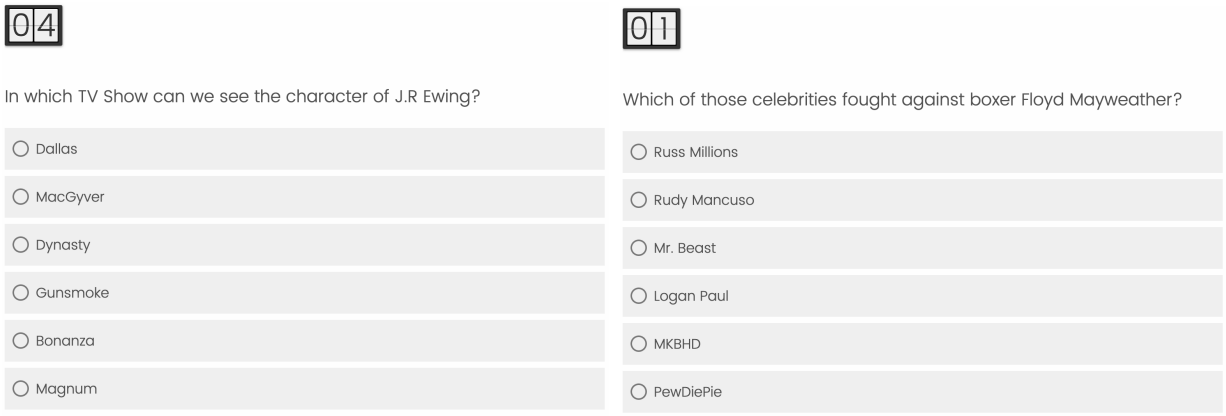
To confirm that the *easy* and *hard* quizzes represent a vertically differentiated market, we compared pretest participants' performance across both quizzes. The *easy* quiz had an average correct response rate of 69%, yielding expected earnings of £3.90, while the *hard* quiz had a correct response 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 vertical product differentiation.

3.3.2 Horizontally Differentiated Market

Products are classified as horizontally differentiated when, at the same price, preferences vary across the population, with different groups favoring distinct options. Thus, preferences depend on individual consumers' idiosyncratic tastes.

To create a horizontally differentiated market in the experiment, we designed two generation-specific celebrity 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 vertical market, with ten questions selected for each quiz.[7] Figure 2 provides an example of selected questions.

Figure 2: Example of selected questions for the horizontal market



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

To confirm that the generation-specific quizzes capture horizontal differentiation, 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.[8]

[7]For the vertical 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 ($p = 0.81$).

[8]For the horizontal 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 ($p = 0.46$).

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 effectively capture the intended horizontal product differentiation.

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 vertical and horizontal markets to assess the causal effects of ratings 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 200 participants (100 below 30 years old and 100 above 50 years old) for each market type (vertical and horizontal). 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 the average rating and the number of ratings submitted by prior participants for each quiz. Depending on the experimental condition, quizzes were either vertically or horizontally differentiated. This design reflects rating aggregation and display practices used by major 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 ratings on earnings in both vertically and horizontally differentiated markets.

The Rating treatments further enable an individual-level analysis of purchase behavior, which we use to explore two additional questions. First, we assess how participants respond to the information conveyed by ratings—specifically, whether their choices are systematically influenced by both the average rating and the number of reviews. Second, we examine whether these choices are consistent with Bayesian rationality, that is, whether participants integrate ratings and review counts in a way that reflects rational belief updating.

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 treatment allows 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 for each market type (vertical and horizontal), with 30 participants per market making sequential choices. 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 Additional Treatments: Enhanced Rating and Algorithm Recommendation Systems

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

3.5.1 Filtering

Ratings alone may not sufficiently improve welfare in horizontally differentiated markets. To address this, platforms such as Booking.com allow customers to filter ratings by categories like traveler type (e.g., family or business). This enhanced rating system leverages the structure of horizontally differentiated markets, which consist of subgroups with distinct preferences. By filtering ratings by subgroups, platforms can increase the informativeness of ratings relevant to each subgroup.

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 a horizontal market. Figure 4a summarizes the information provided at the buying stage.

3.5.2 Freezing

In horizontally differentiated 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.[9]

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

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

reviews. This delay is intended to reduce the impact of noisy early feedback and provide participants with more stable, representative signals.[10]

3.5.3 Algorithm

Many online platforms selling horizontally differentiated products—such as Netflix, Spotify, and Amazon—rely on algorithmic recommendation systems. These systems, often 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 horizontal 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.[11] 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."[12]

We recruited 200 participants for this treatment, evenly split between those under 30 and those over 50 years old. 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.

[10] 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).

[11] 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 9 in the Appendix.

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

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*

Below you can find a table summarizing the choices you can make.

✨ Based on previous performance of participants similar to you, we suggest you buy **Quiz YELLOW**

	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*

3.6 Implementation details

For the main experiment, we recruited a total of 2,279 participants either below 30 years old or above 50 years old, randomly assigned across the four main and three additional treatments. Note that we refer to the four main treatments as Baseline and Rating treatments in vertical and horizontal markets, respectively. We refer to Filtering, Freezing, and Algorithms as additional treatments, as they are only implemented in the horizontal market to overcome potential inefficiencies of the standard rating systems. Each participant could take part in the experiment only once. 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				Additional treatments		
Age Group	**Vertical Baseline**	**Vertical Rating**	**Horizontal Baseline**	**Horizontal Rating**	**Horizontal Filtering**	**Horizontal Freezing**	**Horizontal Algorithm**
Older than 50 y.o.	99 *(58.5)*	210 *(60.3)*	99 *(59.3)*	210 *(60.5)*	210 *(58.2)*	210 *(57.2)*	100 *(58.7)*
Younger than 30 y.o.	100 *(25.3)*	210 *(24.9)*	100 *(24.6)*	210 *(24.8)*	210 *(24.7)*	210 *(25.3)*	101 *(24.7)*
All	199	420	199	420	420	420	201

Average age in parentheses

4 Results

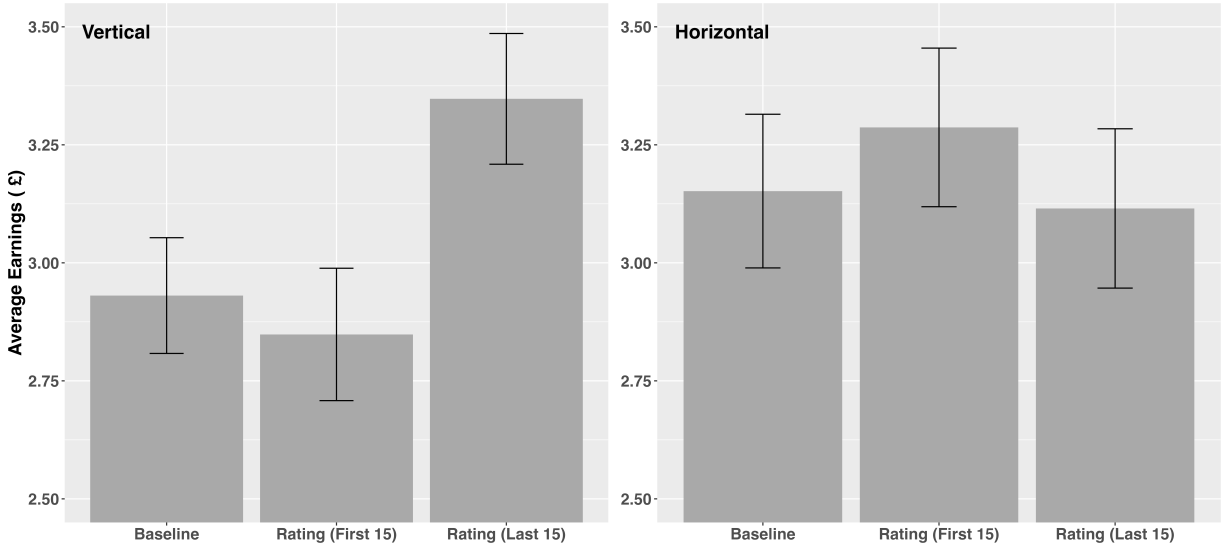
This section presents the experimental results. We begin by analyzing earnings in the two primary treatments, focusing on the effects of product differentiation and the availability of ratings. We then examine the impact of the additional treatments, followed by an analysis of the determinants of individual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

4.1 Earnings: Main Treatments

We first evaluate the welfare implications of two key experimental dimensions: (i) the type of product differentiation (vertical vs. horizontal) and (ii) the presence or absence of 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 ratings, we compare earnings in the Baseline treatment with those of the final 15 participants in the Rating treatment.[13] The results show that ratings significantly improve earnings in vertically differentiated markets, but have no measurable effect in horizontally differentiated markets.

Figure 5: Average earnings in main treatments



Error bars are 95% confidence intervals

[13]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 (%)	
	(Vertical)	(Horizontal)	(Vertical)	(Horizontal)	(Vertical)	(Horizontal)
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Constant	2.876***	3.020***	0.341***	0.563***	0.779***	0.804***
	(0.149)	(0.133)	(0.090)	(0.072)	(0.054)	(0.070)
Rating first 15	-0.066	0.137	0.186**	0.050	0.019	-0.025
	(0.113)	(0.122)	(0.084)	(0.053)	(0.041)	(0.040)
Rating last 15	0.433***	-0.044	0.417***	-0.058	0.039	0.023
	(0.099)	(0.103)	(0.060)	(0.054)	(0.046)	(0.046)
Observations	616	617	468	470	616	617
R^2	0.058	0.006	0.119	0.017	0.023	0.046

Results of OLS regressions with standard errors 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 present regression estimates of treatment effects on participant earnings. In vertically differentiated markets, the average earnings of the last 15 participants in each sequence of the Rating treatment are significantly higher than 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 after sufficient information has accumulated. The observed welfare gains correspond to a 14% increase in average earnings.[14] These findings support **Hypothesis H1** (*Ratings improve welfare in vertically differentiated markets*).

In horizontally differentiated markets, however, the average earnings of the last 15 participants in the Rating treatment are not significantly different from those in the Baseline, indicating that ratings fail to improve welfare in these settings. This is consistent with **Hypothesis H2** (*Ratings alone do not improve welfare in horizontally differentiated markets*).[15]

To further illustrate these dynamics, Figure 11 in the Appendix plots average earnings by each position in the sequence. In vertical markets, average earning of participants in the

[14] 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.

[15] We observe higher earnings in the first 15 rounds of the horizontal market relative to the vertical market. This is due to a bias preference of participants in favor of the Blue quiz, which happens to be the suboptimal option for all participants. In vertical markets, this bias results in a larger earnings loss, whereas in horizontal markets—the cost of the Blue preferences is lower as Blue is preferred by a half of the participants. This pattern also explains why Baseline earnings are higher in horizontal than in vertical markets. More details of the determinants of the quiz choice are presented in section 4.3.1.

last 15 participants consistently yield earnings above average earnings in Baseline, whereas in horizontal markets, only 5 out of these 15 exceed the Baseline earnings.

Two primary channels could drive the observed earnings improvements in the vertical market:

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

2. **Higher buying rate** – Ratings may encourage individuals who otherwise would not have purchased a quiz to make a purchase. If these new buyers earn more than they would in the outside option (the counting zeros task), overall earnings improves.[16]

Models (3) and (4) in Figure 3 analyze match rates, while models (5) and (6) examine buying rates across treatments. In the vertical market, we observe a substantial increase in match rates—from 32% in Baseline to 73% in the last 15 rounds of Rating.[17] By contrast, and contrary to theoretical predictions, the introduction of ratings did not significantly increase the rate at which participants chose to purchase a quiz. One possible explanation is that participants primarily used ratings to select between tasks rather than to assess whether any task was worth purchasing. We analyze this further in section 4.3.1.

In the horizontal market, we find no significant change in either match rates or buying rates. These results reinforce the conclusion that the effectiveness of ratings hinges on the structure of product differentiation.

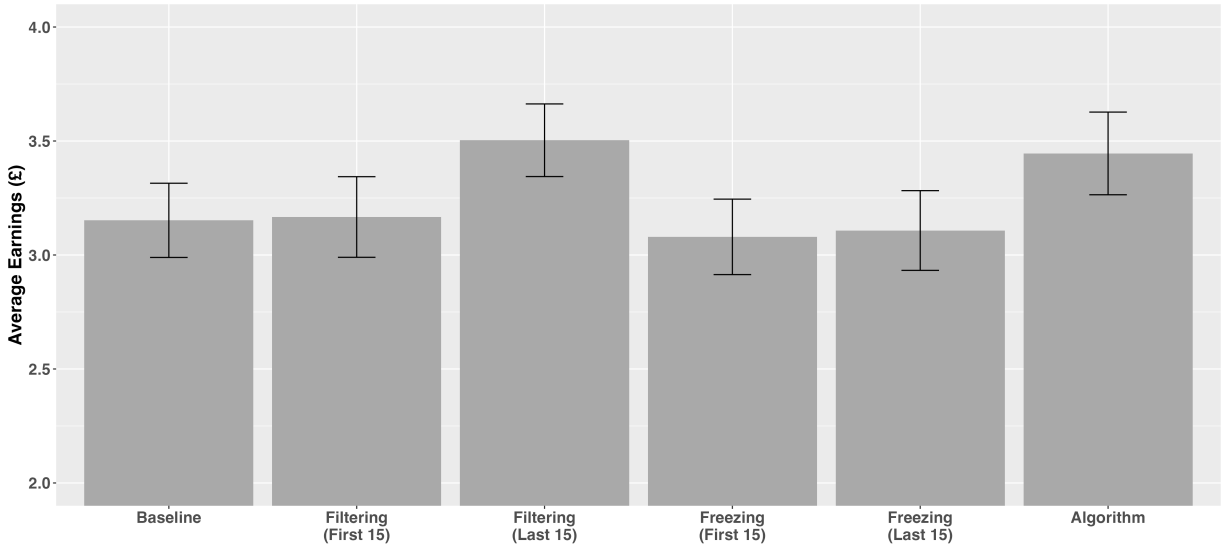
4.2 Earnings: Additional Treatments

We now present the earnings results for the additional treatments in the horizontal market. Figure 6 shows the average earnings across these treatments, while the Table presented in Table 4 reports the regression analyses for earnings, match rate, and buy rate relative to Baseline. Below, we discuss each treatment individually.

[16]Conversely, ratings could discourage low-type individuals from purchasing a quiz. However, as we will see later, ratings for the easy quiz were overwhelmingly positive, making this direction unlikely.

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

Figure 6: Average earnings in additional treatments



Error bars are 95% confidence intervals

Table 4: 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***
	(0.128)	(0.058)	(0.044)
Filtering first 15	0.021	0.004	-0.021
	(0.122)	(0.056)	(0.042)
Filtering last 15	0.347***	0.187***	0.065*
	(0.134)	(0.051)	(0.039)
Freezing first 15	-0.072	-0.040	-0.025
	(0.112)	(0.049)	(0.043)
Freezing last 15	-0.046	0.047	-0.043
	(0.138)	(0.054)	(0.040)
Algorithm	0.309**	0.429***	0.080**
	(0.125)	(0.046)	(0.040)
Observations	1,238	958	1,238
R^2	0.020	0.125	0.073

Results of OLS regressions with standard errors clustered at the individual level for 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 horizontal markets ($p = 0.01$).[18] These findings provide strong empirical support for **Hypothesis H3** *(Filtering improves welfare in horizontally differentiated markets).*

Filtering works by identifying subpopulations with similar preferences and displaying only the most relevant ratings, mimicking a vertically differentiated market. However, this approach 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 leverage the fact that average earnings in cases of task match and mismatch are statistically similar between vertical and horizontal markets.[19] 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 vertical rating treatment for participants in positions 8–15 ($p = 0.02$), reinforcing the welfare advantage of Filtering over Rating.[18] These findings suggest that filtering not only enhances the informativeness of ratings but also leads to greater perceived credibility, ultimately leading to greater 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 not surprising as we expected that delaying early ratings would not affect long-run outcomes, but rather help stabilize initial signals by reducing herding and lowering the variance of market dynamics driven by path dependence. We analyze these dynamics explicitly in the next section.

Algorithm

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

[18]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.

[19]In Baseline (where ratings have no influence), participant earnings are statistically similar between vertical and horizontal markets when matched ($p = 0.68$) and when mismatched ($p = 0.86$).

increase in the match rate, and significant increase in the buying rate. 91% of participants followed the algorithm's recommendation.[20]

Table 5 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, 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 reflects 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 4, algorithmic recommendations significantly increase the share of participants who purchase a quiz, particularly among those who, based on Baseline data, would have rationally abstained. 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).

Finally, we find that algorithmic recommendations do not lead to a significantly higher buying rate compared to the Filtering treatment. Note, however, that Filtering also increases the buying rate relative to Baseline, although the effect is only marginally significant. These findings suggest that both Filtering and Algorithm provide participants with additional information that encourages them to purchase the quiz. However, in the Algorithm treatment, this increase is often not rational, 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.

[20]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.

Table 5: Algorithm vs. filtering (last 15 participants)

Dependent Variable:	Earnings (£)			Match (%)	Buy (%)
	(All)	(Match)	(Mis-Match)	(All)	
Model:	(1)	(2)	(3)	(4)	(5)
Constant	3.388***	3.952***	2.426***	0.678***	0.910***
	(0.221)	(0.256)	(0.441)	(0.076)	(0.066)
Algorithm	-0.019	-0.403***	-0.380	0.250***	0.017
	(0.124)	(0.147)	(0.291)	(0.043)	(0.037)
Observations	409	267	73	340	409
R^2	0.015	0.060	0.103	0.104	0.040

Results of OLS regressions with standard errors clustered at the individual level for Algorithm treatment and at the independent sequence level for Filtering treatments. 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 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.

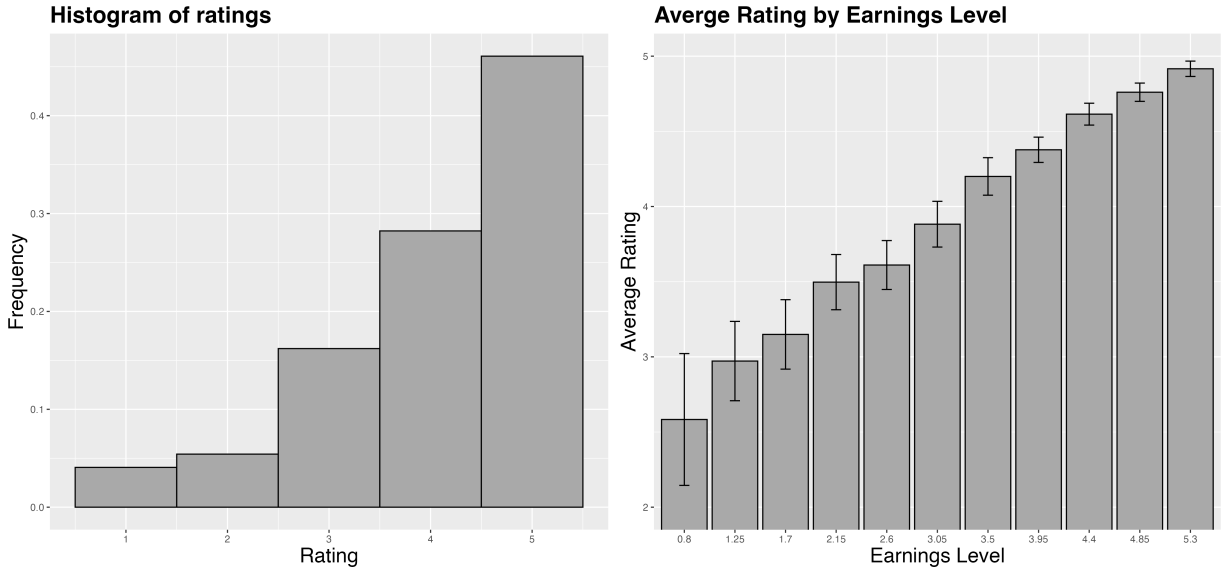
Rating Provision (Extensive Margin). Among participants who purchased a quiz, 97% provided a rating. This near-universal take-up contrasts with the lower and more selective participation observed in most digital platforms, where rating behavior is endogenous and potentially biased (Hu et al. (2009);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

Table 10 in Appendix A 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.[21] 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.

[21]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.

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.

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



Error bars are 95% confidence intervals

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 11 in Appendix A, reveal several robust patterns.

First, each additional pound earned increases the average rating by 0.5 points, indicating that participants evaluate quiz quality primarily through the lens of realized earnings. Second, conditional on earnings, participants matched to their age-specific quiz (i.e., under 30 or over 50) give systematically higher ratings in treatments with horizontal markets. This effect does not appear in vertical market treatments, suggesting that age-matched quizzes provide additional utility beyond monetary payoff. Third, more risk-averse participants give lower ratings in horizontal markets. Fourth, participants under 30 assign significantly lower ratings than those over 50, even after controlling for earnings and match status.

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 12 in Appendix A. 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 ratings—that is, in the Baseline treatment or among early participants in the other treatments—Blue was chosen 67.9% of the time, revealing a strong underlying preference for that quiz. 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 at least one review with 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 market structure, we find that compliance rates are similar in vertical and horizontal rating treatments. However, the consequences of non-compliance differ sharply. In horizontally differentiated markets, deviating from the higher-rated option does not entail a significant earnings loss ($p = 0.52$). By contrast, in vertically differentiated markets, non-compliance is highly detrimental: earnings differ substantially between followers and non-followers ($p = 0.00$). This asymmetry underscores that ratings only translate into welfare gains when they convey information about a universally superior option, whereas in horizontally differentiated settings, following or disregarding ratings has limited payoff consequences.

To formally assess how participants incorporate ratings, we estimate a series of regression models. We begin with a simplified "myopic" model in which participants base their decision solely on average ratings, ignoring the number of ratings. Model (1) of Table 6 confirms that an increase in the Blue quiz's average rating significantly increases the likelihood of choosing Blue, while a higher rating for Yellow reduces this probability. This

suggests that participants are more likely to follow ratings when the difference in average ratings between the two quizzes is more pronounced.

In Model (2), we include the difference in the number of ratings between the two quizzes. Conditional on average ratings, participants are more likely to choose the higher-rated quiz when it also has more reviews. This suggests that participants consider both the value and the number of ratings when deciding between quizzes.

Table 6: Effect of ratings on quiz choice

Dependent Variable:	Task Selected = Blue (%)	
Model:	(1)	(2)
Constant	0.513**	0.508**
	(0.188)	(0.172)
Average Blue	0.143***	0.131***
	(0.034)	(0.034)
Average Yellow	-0.152***	-0.137***
	(0.032)	(0.029)
Confidence (0-100)	-0.001	-0.001
	(0.001)	(0.001)
Risk aversion (0-10)	-0.006	-0.006
	(0.005)	(0.005)
Age group = Young	0.037	0.030
	(0.033)	(0.034)
Gender = Male	-0.004	-0.003
	(0.032)	(0.031)
Filtering	0.120*	0.108*
	(0.067)	(0.058)
Freezing	0.010	-0.001
	(0.069)	(0.063)
Rating Horizontal	0.038	0.023
	(0.044)	(0.041)
N Ratings (Blue - Yellow)		0.008**
		(0.004)
Observations	827	827
R^2	0.115	0.122
Adjusted R^2	0.105	0.111

Results of OLS regressions with standard errors clustered at the independent sequence level. Controls include gender, risk aversion, self-assessed confidence in the task, and individual treatment. Controls are not statistically significant. Significantly different from zero at 1% (***), 5% (**), 10% (*).

4.4 Ratings and Herding Behavior

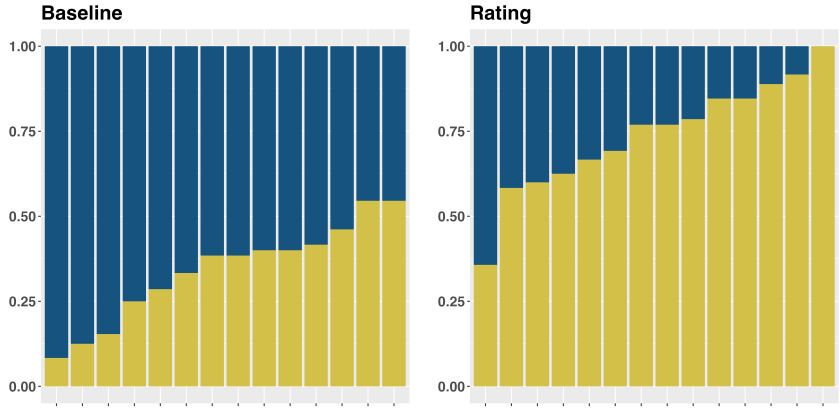
Ratings not only guide individual decisions but also shape collective dynamics by triggering herding behavior—that is, systematic convergence in choices based on publicly visible signals. While this can help coordinate participants on better options, it may also am-

plify 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. In vertically differentiated markets, where one option is objectively superior, ratings consistently steer participants toward the better Yellow task (*easy* quiz). Over time, this leads to increasing concentration in choices, with the inferior Blue task (*hard* quiz) gradually abandoned. In this case, herding facilitates efficient convergence, as quality is shared across participants and ratings aggregate meaningful information. In the sections that follow, we assess the stability and consistency of these outcomes across independent market sequences, and contrast them with the more fragmented dynamics observed in horizontally differentiated settings.

To assess the consistency of these dynamics, we examine 14 independent rating sequences in vertical markets. Each sequence evolves separately, generating distinct rating trajectories and choice patterns. As a benchmark, we construct 14 synthetic baseline groups by randomly regrouping participants from the vertical baseline condition, who did not observe ratings. Figure 8 compares the proportion of Blue task selections in these two environments. In the baseline (left panel), Blue often dominates, consistent with its higher appeal when ratings are absent. In contrast, in the rating sequences (right panel), Yellow tends to dominate: in 13 out of 14 sequences, it captures at least 50% of final task selections. This aggregate convergence suggests that ratings support consistent herding toward the higher-quality option.

Figure 8: Average yellow task (*easy* quiz) to blue task (*hard* quiz) market share by sequence and treatment in vertical markets

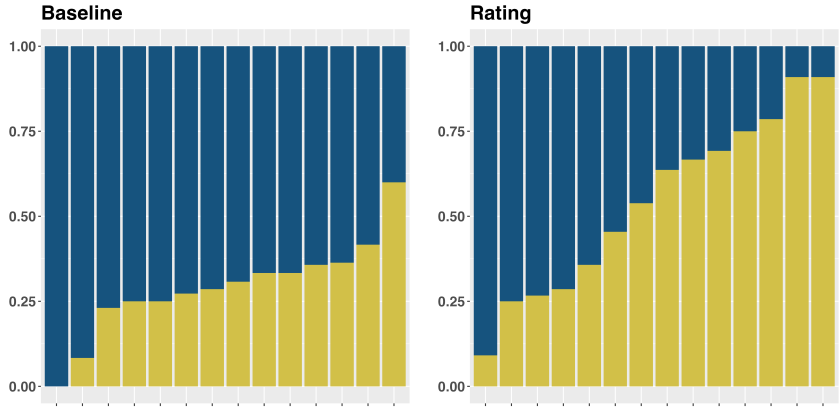


We then turn to horizontal markets, where task appeal depends on individual characteristics, and no option is universally superior. Figure 9 displays task market shares across

Baseline and Rating treatments for horizontal 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 vertically differentiated 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.

Figure 9: Average yellow to blue market share by sequence and treatment in horizontal markets



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 proportion of Yellow choices among the final 15 participants. Recall that the Yellow task is tailored to participants below 30 years old. Table 7 reports results separately for the Rating, Freezing, and Filtering treatments in horizontal 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.

Importantly, these regressions are conducted at the sequence level, with each observation corresponding to a single group. As a result, the analysis includes only 14 observations per treatment, which limits statistical power and precludes precise estimation of effect sizes. Nevertheless, the observed patterns are consistent with herding shaped by early group composition, underscoring the importance of initial conditions in horizontally differentiated environments. As a robustness exercise, Table 8 reports analogous regressions at

the individual level. The results remain qualitatively unchanged, though coefficients are estimated with greater precision.

Table 7: Effect of early composition on later task choice (sequence-level)

Dependent Variable:	Yellow Rate (Last 15)		
	Rating	Freezing	Filtering
Model:	(1)	(2)	(3)
Constant	-0.069	0.513	0.634
	(0.393)	(0.460)	(0.596)
Young (First 15) %	1.146	-0.032	-0.265
	(0.687)	(0.853)	(1.164)
Observations	14	14	14
R^2	0.170	0.0001	0.006
Adjusted R^2	0.101	-0.083	-0.077

Each observation corresponds to a group (14 per treatment). Significantly different from zero at 1% (***), 5% (**), 10% (*).

Table 8: Individual task choice of last 15 participants

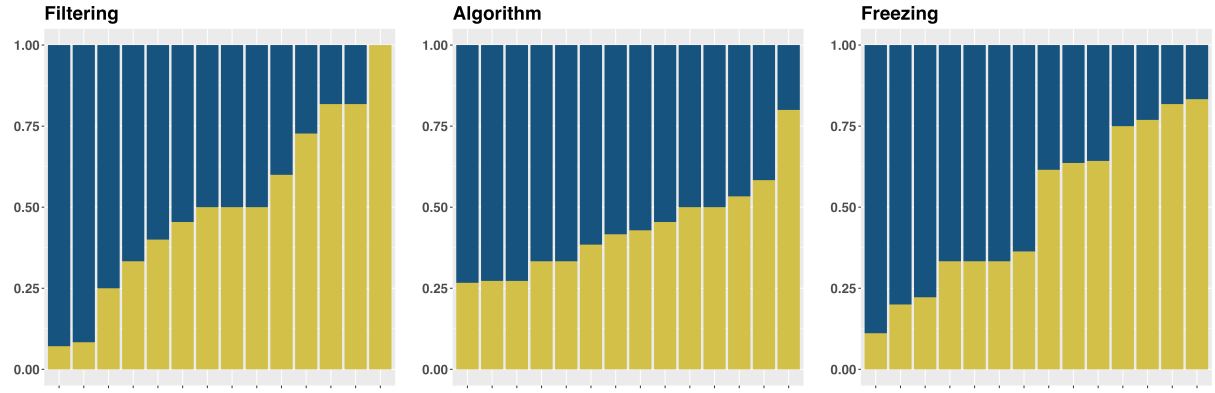
Dependent Variable:	Task = Yellow		
	Rating	Freezing	Filtering
Model:	(1)	(2)	(3)
Constant	-1.491**	-0.099	0.461
	(0.646)	(0.769)	(0.707)
Young (First 15) %	2.603**	0.071	-0.904
	(1.115)	(1.345)	(1.261)
Confidence (0-100)	0.001	0.000	0.001
	(0.004)	(0.004)	(0.004)
Risk aversion (0-10)	0.032	0.022	-0.031
	(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% (*).

Both the Freezing and Filtering treatments 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 vertical environment where ratings convey primarily quality information rather than horizontal preferences.

Figure 10 illustrates the resulting market shares of quizzes among the last 15 participants in each sequence. Filtering produces concentrated task selection patterns, similar to Ratings, while Algorithm leads to more balanced outcomes across tasks. By contrast, Freezing still displays considerable heterogeneity across groups.

Figure 10: Average yellow to blue market share by sequence and treatment in horizontal markets for additional treatments



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 13 in Appendix A 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.

5 Conclusion

This paper causally evaluates the impact of rating systems on consumer welfare across different market structures. We show that while ratings effectively guide consumer choices in vertically differentiated markets—where product quality is objectively ranked—they fail to enhance welfare in horizontally differentiated settings, where preferences differ across individuals. Aggregated ratings in such environments collapse heterogeneous experiences into a misleading signal, leading to suboptimal consumer-product matches.

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 horizontal 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.

Finally, leveraging 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. This highlights the fragility of rating systems in heterogeneous markets and the value of personalized or type-aware information.

Together, our findings demonstrate that the effectiveness of online rating systems depends critically on market structure and platform design. Future rating and recommendation mechanisms must move beyond uniform aggregation toward more context-aware and user-sensitive approaches to achieve welfare-enhancing outcomes.

Future research could explore additional mechanisms to improve rating informativeness, particularly in markets where preference heterogeneity is multi-dimensional. Multidimensional preferences can lower the applicability of filtering but would not affect algorithmic recommendations. However, algorithmic recommendations in cases of complex multidimensional preferences might raise concerns about transparency (Dargnies et al., 2024a) and the strategic use of algorithms by platforms to attract consumers to potentially more profitable options. Whether this could trigger a backlash from consumers remains an open question.

References

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

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

Roland Bénabou and Nikhil Vellodi. 2024. *(Pro-) Social Learning and Strategic Disclosure*. Technical Report. National Bureau of Economic Research.

Jean-Michel Benkert and Armin Schmutzler. 2024. A Theory of Recommendations. arXiv:2408.11362 [econ.GN] https://arxiv.org/abs/2408.11362

Gordon Burtch, Yili Hong, Ravi Bapna, and Vladas Griskevicius. 2018. Stimulating online reviews by combining financial incentives and social norms. *Management Science* 64, 5 (2018), 2065–2082.

Luis Cabral and Ali Hortacsu. 2010. The dynamics of seller reputation: Evidence from eBay. *The journal of industrial economics* 58, 1 (2010), 54–78.

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

Christoph Carnehl, Maximilian Schaefer, André Stenzel, and Kevin Ducbao Tran. 2022. Value for money and selection: How pricing affects airbnb ratings. *Innocenzo Gasparini Institute for Economic Research Working Paper Series* (2022).

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

Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* 43, 3 (2006), 345–354.

Marie-Pierre Dargnies, Rustamdjan Hakimov, and Dorothea Kübler. 2024a. Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science* (2024).

Marie-Pierre Dargnies, Rustamdjan Hakimov, and Dorothea Kübler. 2024b. Behavioral measures improve AI hiring: a field experiment. In *Proceedings of the 25th ACM Conference on Economics and Computation*. 831–832.

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

Florian Dendorfer and Regina Seibel. 2024. *The Cost of the Cold-Start Problem on Airbnb*. Technical Report.

Armin Falk, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. Global evidence on economic preferences. *The Quarterly Journal of Economics* 133, 4 (2018), 1645–1692.

Forbes. 2022a. Amazon Has Turned 'Rings of Power' Star Ratings Back On, Here's How Fans Are Scoring It. *Forbes* (2022). https://www.forbes.com/sites/paultassi/2022/09/10/amazon-has-turned-rings-of-power-star-ratings-back-on-heres-how-fans-are-scoring-it/ Accessed: January 12, 2025.

Forbes. 2022b. 'Rings of Power' Is Getting Review Bombed So Hard Amazon Suspended Reviews Entirely. *Forbes* (2022). https://www.forbes.com/sites/paultassi/2022/09/05/rings-of-power-is-getting-review-bombed-so-hard-amazon-suspended-reviews-entirely/ Accessed: January 12, 2025.

Andrey Fradkin and David Holtz. 2023. Do incentives to review help the market? Evidence from a field experiment on Airbnb. *Marketing Science* 42, 5 (2023), 853–865.

Sherry He, Brett Hollenbeck, and Davide Proserpio. 2022. The market for fake reviews. *Marketing Science* 41, 5 (2022), 896–921.

Nan Hu, Paul A Pavlou, and Jennifer 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*. 324–330.

Nan Hu, Paul A Pavlou, and Jie Zhang. 2017. On self-selection biases in online product reviews. *MIS quarterly* 41, 2 (2017), 449–475.

Nan Hu, Jie Zhang, and Paul A Pavlou. 2009. Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 10 (2009), 144–147.

Bar Ifrach, Costis Maglaras, Marco Scarsini, and Anna Zseleva. 2019. Bayesian social learning from consumer reviews. *Operations Research* 67, 5 (2019), 1209–1221.

Johannes Johnen and Robin Ng. 2024. *Harvesting ratings.* Technical Report. University of Bonn and University of Mannheim, Germany.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

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

Jonathan Lafky and Robin Ng. 2024. *Ratings with Heterogeneous Preferences.* Technical Report. University of Bonn and University of Mannheim, Germany.

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

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

Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management science* 62, 12 (2016), 3412–3427.

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

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

Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The value of reputation on eBay: A controlled experiment. *Experimental economics* 9, 2 (2006), 79–101.

Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311, 5762 (2006), 854–856.

Steven Tadelis. 2016. Reputation and feedback systems in online platform markets. *Annual review of economics* 8, 1 (2016), 321–340.

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

A Additional Tables and Figures

Table 9: Determinants of earnings for blue and yellow quizzes in baseline horizontal

Dependent Variable:	Earnings (£)	
	(Blue)	(Yellow)
Model:	(1)	(2)
Constant	3.637***	2.173***
	(0.345)	(0.631)
Age group = Young	-1.307***	1.640***
	(0.214)	(0.311)
Gender = Male	0.268	0.088
	(0.218)	(0.343)
Confidence (0-100)	-0.006	0.012
	(0.005)	(0.008)
Risk aversion (0-10)	0.111**	-0.002
	(0.044)	(0.066)
Observations	106	44
R^2	0.366	0.426
Adjusted R^2	0.341	0.367

Results of OLS regressions. Significantly different from zero at 1% (***), 5% (**), 10% (*).

Figure 11: Average earnings in vertical and horizontal markets by sequence

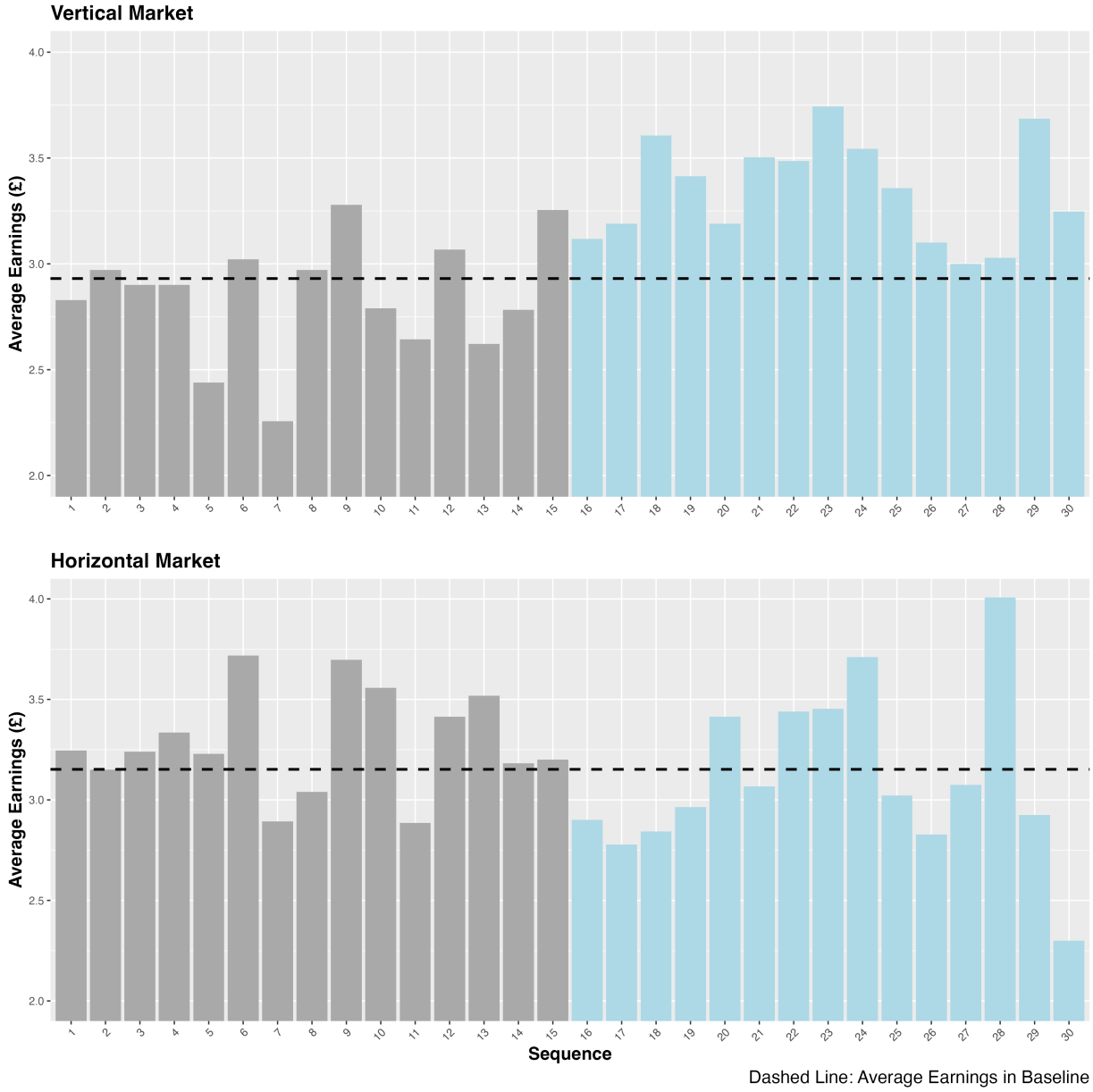


Table 10: Determinants of rating provision

Dependent Variable:	Rater (yes)	
Model:	(1)	(2)
Constant	1.296***	1.170*
	(0.347)	(0.632)
Earnings (£)	0.273***	0.270***
	(0.054)	(0.075)
Age group = Young	-0.160	-0.202
	(0.132)	(0.182)
Risk aversion (0-10)	-0.019	-0.063*
	(0.027)	(0.038)
Confidence (0-100)	-0.003	-0.003
	(0.003)	(0.004)
Gender = Male	0.547***	0.322
	(0.171)	(0.213)
Algorithm	0.439	
	(0.425)	
Baseline Horizontal	-0.342	
	(0.290)	
Filtering	-0.071	0.020
	(0.271)	(0.246)
Freezing	0.021	0.289
	(0.275)	(0.310)
Rating Horizontal	-0.163	-0.077
	(0.267)	(0.223)
Rating Vertical	-0.056	
	(0.272)	
Average rating of selected task		0.072
		(0.126)
Number of ratings of selected task		0.012
		(0.024)
Observations	1,746	1,020
Squared Correlation	0.040	0.031
Pseudo R^2	0.111	0.097
BIC	505.5	301.5

Results of Probit regressions where the dependent variable is a binary indicator equal to 1 if the participant submitted a rating for the quiz and 0 otherwise. Independent variables include participant earnings from the quiz, age group, risk aversion, confidence, gender, the average rating, and number of ratings of the selected quiz at the time of decision. The sample includes only participants who purchased a quiz. Significantly different from zero at 1% (***), 5% (**), 10% (*).

Table 11: Effects of earnings, matching, and risk preferences on rating generosity

Dependent Variable:	Rating (1-5)				
		(All)		(Horizontal)	(Vertical)
Model:	(1)	(2)	(3)	(4)	(5)
Constant	2.350***	2.463***	2.805***	2.710***	2.783***
	(0.061)	(0.094)	(0.114)	(0.133)	(0.215)
Earnings (£)	0.503***	0.503***	0.479***	0.478***	0.480***
	(0.015)	(0.016)	(0.020)	(0.024)	(0.057)
Match = Yes			0.168***	0.240***	-0.018
			(0.047)	(0.070)	(0.115)
Gender = Male			-0.014	0.000	-0.070
			(0.033)	(0.036)	(0.086)
Age group = Young			-0.259***	-0.247***	-0.286***
			(0.038)	(0.048)	(0.090)
Risk aversion (0-10)			-0.048***	-0.057***	-0.026
			(0.007)	(0.010)	(0.019)
Confidence (0-100)			0.001	0.001	0.001
			(0.001)	(0.001)	(0.002)
Horizontal Baseline		-0.070	-0.091		
		(0.108)	(0.106)		
Vertical Rating		0.032	-0.039		0.026
		(0.091)	(0.091)		(0.095)
Horizontal Rating		-0.228**	-0.263***	-0.175*	
		(0.096)	(0.094)	(0.096)	
Horizontal Filtering		-0.166*	-0.213**	-0.129	
		(0.085)	(0.085)	(0.087)	
Horizontal Freezing		-0.315***	-0.362***	-0.274***	
		(0.090)	(0.090)	(0.093)	
Horizontal Algorithm		0.144	0.029	0.086	
		(0.096)	(0.094)	(0.097)	
Observations	1,697	1,697	1,694	1,239	455
R^2	0.348	0.366	0.396	0.439	0.261
Adjusted R^2	0.348	0.363	0.392	0.434	0.249

Results of OLS regressions with standard errors clustered at the individual level for Baseline treatments and at the sequence level for Rating treatments. Models (1), (2), (3), and (5) use the "Vertical Baseline" as the reference category for treatment. Model (4) uses the "Horizontal Baseline" as the reference. Significantly different from zero at 1% (***), 5% (**), 10% (*).

Table 12: Effect of ratings and individual characteristics on selecting the outside option

Dependent Variable:	Outside Option (%)	
Model:	(1)	(2)
Constant	0.240***	0.464**
	(0.046)	(0.164)
At least 1 rating	0.000	
	(0.033)	
Max average rating		-0.032
		(0.034)
Min average rating		-0.021
		(0.015)
Total number of ratings		-0.002
		(0.002)
Confidence (0-100)	-0.002***	-0.002***
	(0.000)	(0.000)
Risk aversion (0-10)	0.025***	0.023***
	(0.004)	(0.005)
Age group = Young	-0.060***	-0.066**
	(0.015)	(0.027)
Gender = Male	0.084***	0.084**
	(0.022)	(0.039)
Observations	2,073	1,062
R^2	0.053	0.056
Adjusted R^2	0.048	0.047

Results of OLS regressions with standard errors clustered at the individual level for Baseline treatments and at the independent sequence level for Rating treatments. Includes individual treatment controls. Model (1) considers all observation except those from the Algorithm treatment. Model (2) is restricted to observations where at least 1 rating is made available for each task. Significantly different from zero at 1% (***), 5% (**), 10% (*).

Table 13: Effect of first unfrozen quiz on later task choice (freezing)

Dependent Variable:	Task=yellow
Model:	(1)
Constant	-0.561*
	(0.298)
First Unfrozen = Yellow	0.696***
	(0.193)
Confidence (0-100)	0.005
	(0.004)
Risk aversion (0-10)	-0.021
	(0.037)
Gender = Male	-0.073
	(0.183)
Observations	210
Squared Correlation	0.063
Pseudo R^2	0.046
BIC	293.8

Results of Probit regressions where the dependent variable is an indicator for choosing the Yellow task among the last 15 participants in each sequence. The key explanatory variable is an indicator for whether the Yellow task was the first to unfreeze (i.e., reach the five-rating threshold). Controls include self-assessed confidence, risk preferences, and gender. Standard errors are clustered at the sequence level. Significantly different from zero at 1% (***), 5% (**), 10% (*).

B Proofs

Proof of Proposition 1: We start by deriving the average payoff for a vertical market without ratings (baseline). In this situation, subjects have to choose between a safe outside option or purchasing one of two quizzes but do not have access to ratings. Accordingly, a subject with skill θ who chooses quiz k thinks his expected payoff is

$$E[I(\mu,\theta)] = a + b\theta^{\frac{1}{\mu}-1},$$

however, the true expected payoff of a consumer of skill θ who chooses quiz k is

$$E[I(q_k,\theta)] = a + b\theta^{\frac{1}{q_k}-1}.$$

A subject with skill θ prefers to buy quiz k instead of taking the outside option when

$$E[I(\mu,\theta)] > z,$$

or

$$a + b\theta^{\frac{1}{\mu}-1} > z,$$

or

$$\theta^{\frac{1}{\mu}-1} > \frac{z-a}{b},$$

or

$$\theta > \left(\frac{z-a}{b}\right)^{\frac{\mu}{1-\mu}} = \bar{\theta}.$$

The average payoff of subjects who select to take a quiz has to take into account that one quiz is easy and the other quiz is hard. Since we assume subjects assign the same prior mean quality to both quizzes, it is natural to assume that half of subjects choose the easy quiz and half choose the hard quiz. In this case the average payoff of subjects who take a quiz is:

$$\begin{aligned} E[I_V(\text{Baseline})|\theta > \bar{\theta}] &= \frac{1}{\Pr(\theta > \bar{\theta})}\left[\frac{1}{2}\int_{\bar{\theta}}^{1} E[I(q_E,\theta)]f(\theta)d\theta + \frac{1}{2}\int_{\bar{\theta}}^{1} E[I(q_H,\theta)]f(\theta)d\theta\right] \\ &= \frac{1}{1-F(\bar{\theta})}\left[\frac{1}{2}\int_{\bar{\theta}}^{1}(a + b\theta^{\frac{1}{q_E}-1})f(\theta)d\theta + \frac{1}{2}\int_{\bar{\theta}}^{1}(a + b\theta^{\frac{1}{q_H}-1})f(\theta)d\theta\right] \\ &= a + \frac{b}{2\left[1-F(\bar{\theta})\right]}\left[\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_E}-1}f(\theta)d\theta + \int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_H}-1}f(\theta)d\theta\right]. \end{aligned}$$

The overall average payoff for a vertical market without ratings (baseline) takes into account those who choose the outside option and those who take a quiz:

$$
\begin{aligned}
E[I_V(\text{Baseline})] &= \Pr(\theta < \bar{\theta}) z + \Pr(\theta > \bar{\theta}) E[I_V(\text{Baseline}) | \theta > \bar{\theta}] \\
&= F(\bar{\theta}) z + \left[1 - F(\bar{\theta})\right] a + \frac{b}{2} \left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_E} - 1} f(\theta) d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_H} - 1} f(\theta) d\theta \right] \\
&= a + (z - a) F(\bar{\theta}) + \frac{b}{2} \left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_E} - 1} f(\theta) d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_H} - 1} f(\theta) d\theta \right]. \qquad (4)
\end{aligned}
$$

Next, we derive the average payoff for a vertical market with ratings. We focus on a situation where, after many ratings have accumulated, the easy quiz is revealed. The weak law of large numbers implies

$$\bar{r}_E \xrightarrow{p} q_E \text{ and } \bar{r}_H \xrightarrow{p} q_H.$$

This together with equation (1) implies

$$E(q_E | \bar{r}_E) \xrightarrow{p} q_E \text{ and } E(q_H | \bar{r}_H) \xrightarrow{p} q_H.$$

Since $q_E > q_H$ subjects choose either the easy quiz or the outside option. What will be the percentages of subjects making each choice? A subject with skill θ taking the easy quiz prefers it to the outside option when

$$E[I(q_E, \theta)] = a + b\theta^{\frac{1}{q_E} - 1} > z,$$

or

$$\theta > \left(\frac{z - a}{b} \right)^{\frac{q_E}{1 - q_E}} = \bar{\theta}_{q_E}.$$

Accordingly, the average payoff of subjects who select to take the easy quiz is

$$E[I_V(\text{Ratings}) | \theta > \bar{\theta}_{q_E}] = a + \frac{b}{1 - F(\bar{\theta}_{q_E})} \int_{\bar{\theta}_E}^{1} \theta^{\frac{1}{q_E} - 1} f(\theta) d\theta.$$

The overall average payoff for a vertical market with ratings takes into account those who choose the outside option and those who take the easy quiz:

$$
\begin{aligned}
E[I_V(\text{Ratings})] &= \Pr(\theta < \bar{\theta}_{q_E}) z + \Pr(\theta > \bar{\theta}_{q_E}) E[I_V(\text{Ratings}) | \theta > \bar{\theta}_{q_E}] \\
&= F(\bar{\theta}_{q_E}) z + \left[1 - F(\bar{\theta}_{q_E})\right] a + b \int_{\bar{\theta}_E}^{1} \theta^{\frac{1}{q_E} - 1} f(\theta) d\theta \\
&= a + (z - a) F(\bar{\theta}_{q_E}) + b \int_{\bar{\theta}_E}^{1} \theta^{\frac{1}{q_E} - 1} f(\theta) d\theta. \qquad (5)
\end{aligned}
$$

Using equations (4) and (5), we have that

$$E[I_V(\text{Ratings})] > E[I_V(\text{Baseline})]$$

or

$$\Pr(\theta < \bar{\theta}_{q_E})z + \Pr(\theta > \bar{\theta}_{q_E})E[I_V(\text{Ratings})|\theta > \bar{\theta}_{q_E}] > \Pr(\theta < \bar{\theta})z + \Pr(\theta > \bar{\theta})E[I_V(\text{Baseline})|\theta > \bar{\theta}],$$

or

$$a + (z-a)F(\bar{\theta}_{q_E}) + b\int_{\bar{\theta}_E}^{1} \theta^{\frac{1}{q_E}-1} f(\theta)d\theta > a + (z-a)F(\bar{\theta}) + \frac{b}{2}\left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_E}-1} f(\theta)d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_H}-1} f(\theta)d\theta\right],$$

or

$$(z-a)F(\bar{\theta}_{q_E}) + b\int_{\bar{\theta}_E}^{1} \theta^{\frac{1}{q_E}-1} f(\theta)d\theta > (z-a)F(\bar{\theta}) + \frac{b}{2}\left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_E}-1} f(\theta)d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_H}-1} f(\theta)d\theta\right].$$

Note that $\mu < q_E$, $\bar{\theta} = \left(\frac{z-a}{b}\right)^{\frac{\mu}{1-\mu}}$, and $\bar{\theta}_{q_E} = \left(\frac{z-a}{b}\right)^{\frac{q_E}{1-q_E}}$ imply $\bar{\theta} > \bar{\theta}_{q_E}$. Hence, we can write

$$(z-a)F(\bar{\theta}_{q_E}) + b\left[\int_{\bar{\theta}_E}^{\bar{\theta}} \theta^{\frac{1}{q_E}-1} f(\theta)d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_E}-1} f(\theta)d\theta\right] > (z-a)F(\bar{\theta}) + \frac{b}{2}\left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_E}-1} f(\theta)d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_H}-1} f(\theta)d\theta\right],$$

or

$$(z-a)F(\bar{\theta}_{q_E}) + b\int_{\bar{\theta}_E}^{\bar{\theta}} \theta^{\frac{1}{q_E}-1} f(\theta)d\theta + \frac{b}{2}\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_E}-1} f(\theta)d\theta > (z-a)F(\bar{\theta}) + \frac{b}{2}\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_H}-1} f(\theta)d\theta,$$

or

$$b\int_{\bar{\theta}_E}^{\bar{\theta}} \theta^{\frac{1}{q_E}-1} f(\theta)d\theta + \frac{b}{2}\int_{\bar{\theta}}^{1} \left(\theta^{\frac{1}{q_E}-1} - \theta^{\frac{1}{q_H}-1}\right) f(\theta)d\theta > (z-a)\left[F(\bar{\theta}) - F(\bar{\theta}_{q_E})\right],$$

or

$$\int_{\bar{\theta}_E}^{\bar{\theta}} \left[b\theta^{\frac{1}{q_E}-1} - (z-a)\right] f(\theta)d\theta + \frac{b}{2}\int_{\bar{\theta}}^{1} \left(\theta^{\frac{1}{q_E}-1} - \theta^{\frac{1}{q_H}-1}\right) f(\theta)d\theta > 0,$$

or

$$\int_{\bar{\theta}_E}^{\bar{\theta}} \left(a + b\theta^{\frac{1}{q_E}-1} - z\right) f(\theta)d\theta + \frac{b}{2}\int_{\bar{\theta}}^{1} \left(\theta^{\frac{1}{q_E}-1} - \theta^{\frac{1}{q_H}-1}\right) f(\theta)d\theta > 0,$$

which holds since the first integral is positive given that the integrand is positive (within the integration range) and the second integral is also positive since $\theta \in [0,1]$ and $q_E > q_H$ implies the integrand is positive.

Proof of Proposition 2: We start by deriving the average payoff for a horizontal market without ratings (baseline). In this situation, subjects choose between a safe outside option or purchasing one of two quizzes but do not have access to ratings.

A subject of type j and skill θ who takes quiz k thinks his expected payoff is

$$E[I(\mu,\theta)] = a + b\theta^{\frac{1}{\mu}-1},$$

however, the true expected payoff of a subject of type j and skill θ who chooses quiz k is

$$E[I(q_{kj},\theta)] = a + b\theta^{\frac{1}{q_{kj}}-1}.$$

A subject of type j and skill θ prefers to take a quiz to the outside option when $E[I(\mu,\theta)] > z$, or

$$\theta > \left(\frac{z-a}{b}\right)^{\frac{\mu}{1-\mu}} = \bar{\theta}.$$

The average payoff of subjects who select to take a quiz take has to take into account that one quiz is for the young and the other quiz is for the old. Since we assume subjects assign the same prior mean quality to both quizzes it is natural to assume that half of the young choose the young quiz and the other half choose the old quiz. Similarly, half of the old choose the young quiz and the other half choose the old quiz. The average payoff of the young who select to take a quiz in horizontal baseline is

$$\begin{aligned}
E[I_{HY}(\text{Baseline})|\theta &> \bar{\theta}] = \frac{1}{\Pr(\theta>\bar{\theta})}\left[\frac{1}{2}\int_{\bar{\theta}}^{1} E[U(q_{YY},\theta)f(\theta)d\theta + \frac{1}{2}\int_{\bar{\theta}}^{1} E[U(q_{OY},\theta)f(\theta)d\theta\right] \\
&= \frac{1}{2\left[1-F(\bar{\theta})\right]}\left[\int_{\bar{\theta}}^{1}(a+b\theta^{\frac{1}{q_{YY}}-1})f(\theta)d\theta + \int_{\bar{\theta}}^{1}(a+b\theta^{\frac{1}{q_{OY}}-1})f(\theta)d\theta\right] \\
&= a + \frac{b}{2\left[1-F(\bar{\theta})\right]}\left[\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta + \int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OY}}-1}f(\theta)d\theta\right].
\end{aligned}$$

The overall average payoff of the young in horizontal baseline is

$$\begin{aligned}
E[I_{HY}(\text{Baseline})] &= \Pr(\theta<\bar{\theta})z + \Pr(\theta>\bar{\theta})E[I_{HY}(\text{Baseline})|\theta>\bar{\theta}] \\
&= F(\bar{\theta})z + \left[1-F(\bar{\theta})\right]a + \frac{b}{2}\left[\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta + \int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OY}}-1}f(\theta)d\theta\right]. \\
&= a + (z-a)F(\bar{\theta}) + \frac{b}{2}\left[\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta + \int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OY}}-1}f(\theta)d\theta\right].
\end{aligned}$$

The average payoff of the old who select to take a quiz in horizontal baseline is

$$\begin{aligned}
E[I_{HO}(\text{Baseline})|\theta &> \bar{\theta}] = \frac{1}{\Pr(\theta>\bar{\theta})}\left[\frac{1}{2}\int_{\bar{\theta}}^{1} E[U(q_{YO},\theta)f(\theta)d\theta + \frac{1}{2}\int_{\bar{\theta}}^{1} E[U(q_{OO},\theta)f(\theta)d\theta\right] \\
&= a + \frac{b}{2\left[1-F(\bar{\theta})\right]}\left[\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YO}}-1}f(\theta)d\theta + \int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OO}}-1}f(\theta)d\theta\right].
\end{aligned}$$

The overall average payoff of the old in horizontal baseline is

$$\begin{aligned} E[I_{HO}(\text{Baseline})] &= \Pr(\theta < \bar{\theta})z + \Pr(\theta > \bar{\theta})E[I_{HO}(\text{Baseline})|\theta > \bar{\theta}] \\ &= F(\bar{\theta})z + \left[1 - F(\bar{\theta})\right] a + \frac{b}{2}\left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{YO}}-1} f(\theta)d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{OO}}-1} f(\theta)d\theta\right] \\ &= a + (z-a)F(\bar{\theta}) + \frac{b}{2}\left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{YO}}-1} f(\theta)d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{OO}}-1} f(\theta)d\theta\right]. \end{aligned}$$

Hence, the overall average payoff of all subjects in horizontal baseline is

$$\begin{aligned} E[I_H(\text{Baseline})] &= \frac{1}{2}E[I_{HY}(\text{Baseline})] + \frac{1}{2}E[I_{HO}(\text{Baseline})] \\ &= \frac{1}{2}a + \frac{1}{2}(z-a)F(\bar{\theta}) + \frac{b}{4}\left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{YY}}-1} f(\theta)d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{OY}}-1} f(\theta)d\theta\right] \\ &\quad + \frac{1}{2}a + \frac{1}{2}(z-a)F(\bar{\theta}) + \frac{b}{4}\left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{YO}}-1} f(\theta)d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{OO}}-1} f(\theta)d\theta\right] \\ &= a + (z-a)F(\bar{\theta}) + \frac{b}{4}\left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{YY}}-1} f(\theta)d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{OY}}-1} f(\theta)d\theta\right] \\ &\quad + \frac{b}{4}\left[\int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{YO}}-1} f(\theta)d\theta + \int_{\bar{\theta}}^{1} \theta^{\frac{1}{q_{OO}}-1} f(\theta)d\theta\right]. \end{aligned} \tag{6}$$

Now we consider what happens in a horizontal market with ratings. A sequence of ratings for the young quiz by young subjects is a random sample drawn from $N(q_{YY}, \sigma^2)$. A sequence of ratings for the young quiz by old subjects is a random sample drawn from $N(q_{YO}, \sigma^2)$. Hence, a sequence of ratings for the young quiz by an equal number of young and old subjects is a random sample drawn from $N((q_{YY} + q_{YO})/2, \sigma^2/2)$. Similarly, a sequence of ratings for the old quiz by an equal number of young and old subjects is a random sample drawn from $N((q_{OY} + q_{OO})/2, \sigma^2/2)$. This and assumption (2) imply that the distribution of ratings of the young quiz is identical to that of the old quiz. Since the two distributions are identical, ratings are completely uninformative. Therefore, the average payoff in horizontal with ratings is the same as in horizontal without ratings.

Proof of Proposition 3: We start by deriving the average payoff for horizontal with ratings enhanced by filtering. We focus on a situation where, after many ratings accumulate, young and old know which quiz is best for the young and which is best for the old. A young subject who wishes to decide between the two quizzes uses the most informative signal. Hence, he compares the ratings young subjects attribute to the young and old quizzes ignoring the ratings of old subjects. The weak law of large numbers implies

$$\bar{r}_{YY} \xrightarrow{p} q_{YY} \text{ and } \bar{r}_{OY} \xrightarrow{p} q_{OY}.$$

This together with equation (3) implies

$$E(q_{YY}|\bar{r}_{YY}) \xrightarrow{p} q_{YY} \text{ and } E(q_{OY}|\bar{r}_{OY}) \xrightarrow{p} q_{OY}.$$

Since $q_{YY} > q_{OY}$ young subjects either choose the young quiz or the outside option. What will be the percentages of young subjects making each choice? Using the model, a young subject who knows which is the young quiz prefers it to the outside option when

$$E[I(q_{YY}, \theta)] = a + b\theta^{\frac{1}{q_{YY}}-1} > z,$$

or

$$\theta > \left(\frac{z-a}{b}\right)^{\frac{q_{YY}}{1-q_{YY}}} = \bar{\theta}_{YY}.$$

Similarly, an old subject who wishes to decide between the two quizzes uses the most informative signal. Hence, he compares the ratings old subjects attribute to the young and old quizzes ignoring the ratings of young subjects. The weak law of large numbers implies

$$\bar{r}_{YO} \xrightarrow{p} q_{YO} \text{ and } \bar{r}_{OO} \xrightarrow{p} q_{OO}.$$

This together with equation (3) implies

$$E(q_{YO}|\bar{r}_{YO}) \xrightarrow{p} q_{YO} \text{ and } E(q_{OO}|\bar{r}_{OO}) \xrightarrow{p} q_{OO}.$$

Since $q_{OO} > q_{YO}$ old subjects either choose the old quiz or the outside option. What will be the percentages of old subjects making each choice? Using the model, an old subject who knows which is the old quiz prefers it to the outside option when

$$E[I(q_{OO}, \theta)] = a + b\theta^{\frac{1}{q_{OO}}-1} > z,$$

or

$$\theta > \left(\frac{z-a}{b}\right)^{\frac{q_{OO}}{1-q_{OO}}} = \bar{\theta}_{OO}.$$

The average payoff of young subjects who buy the young quiz is

$$E[I_{HY}(\text{Filtering})|\theta > \bar{\theta}_{YY}] = a + \frac{b}{1-F(\bar{\theta}_{YY})}\int_{\bar{\theta}_{YY}}^{1} \theta^{\frac{1}{q_{YY}}-1} f(\theta)d\theta.$$

The overall average payoff of young subjects in horizontal with ratings enhanced by filtering is

$$\begin{aligned} E[I_{HY}(\text{Filtering})] &= \Pr(\theta < \bar{\theta}_{YY})z + \Pr(\theta > \bar{\theta}_{YY}) \left[a + \frac{b}{1 - F(\bar{\theta}_{YY})} \int_{\bar{\theta}_{YY}}^{1} \theta^{\frac{1}{q_{YY}} - 1} f(\theta) d\theta \right] \\ &= F(\bar{\theta}_{YY})z + \left[1 - F(\bar{\theta}_{YY})\right] a + b \int_{\bar{\theta}_{YY}}^{1} \theta^{\frac{1}{q_{YY}} - 1} f(\theta) d\theta \\ &= a + (z - a)F(\bar{\theta}_{YY}) + b \int_{\bar{\theta}_{YY}}^{1} \theta^{\frac{1}{q_{YY}} - 1} f(\theta) d\theta. \end{aligned}$$

The average payoff of old subjects who buy the old quiz is

$$E[I_{HO}(\text{Filtering}) | \theta > \bar{\theta}_{OO}] = a + \frac{b}{1 - F(\bar{\theta}_{OO})} \int_{\bar{\theta}_{OO}}^{1} \theta^{\frac{1}{q_{OO}} - 1} f(\theta) d\theta.$$

The overall average payoff of old subjects in horizontal with ratings enhanced by filtering is

$$\begin{aligned} E[I_{HO}(\text{Filtering})] &= \Pr(\theta < \bar{\theta}_{OO})z + \Pr(\theta > \bar{\theta}_{OO}) \left[a + \int_{\bar{\theta}_{OO}}^{1} \theta^{\frac{1}{q_{OO}} - 1} f(\theta) d\theta \right] \\ &= F(\bar{\theta}_{OO})z + \left[1 - F(\bar{\theta}_{OO})\right] a + b \int_{\bar{\theta}_{OO}}^{1} \theta^{\frac{1}{q_{OO}} - 1} f(\theta) d\theta \\ &= a + (z - a)F(\bar{\theta}_{OO}) + b \int_{\bar{\theta}_{OO}}^{1} \theta^{\frac{1}{q_{OO}} - 1} f(\theta) d\theta. \end{aligned}$$

Hence, the overall average payoff of all subjects in horizontal with ratings enhanced by filtering is

$$\begin{aligned} E[I_H(\text{Filtering})] &= \frac{1}{2} E[I_{HY}(\text{Filtering})] + \frac{1}{2} E[I_{HO}(\text{Filtering})] \\ &= \frac{1}{2} \left[a + (z - a)F(\bar{\theta}_{YY}) + b \int_{\bar{\theta}_{YY}}^{1} \theta^{\frac{1}{q_{YY}} - 1} f(\theta) d\theta \right] \\ &\quad + \frac{1}{2} \left[a + (z - a)F(\bar{\theta}_{OO}) + b \int_{\bar{\theta}_{OO}}^{1} \theta^{\frac{1}{q_{OO}} - 1} f(\theta) d\theta \right]. \end{aligned} \tag{7}$$

Using equations (6) and (7), we have that

$$E[I_{HY}(\text{Filtering})] > E[I_{HY}(\text{Baseline})]$$

or

$$a+\frac{1}{2}(z-a)\left[F(\bar{\theta}_{YY})+F(\bar{\theta}_{OO})\right]+\frac{b}{2}\int_{\bar{\theta}_{YY}}^{1}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta+\frac{b}{2}\int_{\bar{\theta}_{OO}}^{1}\theta^{\frac{1}{q_{OO}}-1}f(\theta)d\theta>$$
$$a+(z-a)F(\bar{\theta})+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OY}}-1}f(\theta)d\theta+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YO}}-1}f(\theta)d\theta+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OO}}-1}f(\theta)d\theta.$$

Note that $\mu<q_{YY}$, $\bar{\theta}=\left(\frac{z-a}{b}\right)^{\frac{\mu}{1-\mu}}$, and $\bar{\theta}_{YY}=\left(\frac{z-a}{b}\right)^{\frac{q_{YY}}{1-q_{YY}}}$ imply $\bar{\theta}>\bar{\theta}_{YY}$. Similarly, $\mu<q_{OO}$, $\bar{\theta}=\left(\frac{z-a}{b}\right)^{\frac{\mu}{1-\mu}}$, and $\bar{\theta}_{OO}=\left(\frac{z-a}{b}\right)^{\frac{q_{OO}}{1-q_{OO}}}$ imply $\bar{\theta}>\bar{\theta}_{OO}$

$$\frac{1}{2}(z-a)\left[F(\bar{\theta}_{YY})+F(\bar{\theta}_{OO})\right]+\frac{b}{2}\left[\int_{\bar{\theta}_{YY}}^{\bar{\theta}}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta+\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta\right]$$
$$+\frac{b}{2}\left[\int_{\bar{\theta}_{OO}}^{\bar{\theta}}\theta^{\frac{1}{q_{OO}}-1}f(\theta)d\theta+\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OO}}-1}f(\theta)d\theta\right]>(z-a)F(\bar{\theta})+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta$$
$$+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OY}}-1}f(\theta)d\theta+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YO}}-1}f(\theta)d\theta+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OO}}-1}f(\theta)d\theta,$$

or

$$\frac{1}{2}(z-a)\left[F(\bar{\theta}_{YY})+F(\bar{\theta}_{OO})\right]+\frac{b}{2}\int_{\bar{\theta}_{YY}}^{\bar{\theta}}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta+\frac{b}{2}\int_{\bar{\theta}_{OO}}^{\bar{\theta}}\theta^{\frac{1}{q_{OO}}-1}f(\theta)d\theta$$
$$+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OO}}-1}f(\theta)d\theta>(z-a)F(\bar{\theta})+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{OY}}-1}f(\theta)d\theta+\frac{b}{4}\int_{\bar{\theta}}^{1}\theta^{\frac{1}{q_{YO}}-1}f(\theta)d\theta,$$

or

$$\frac{b}{2}\int_{\bar{\theta}_{YY}}^{\bar{\theta}}\theta^{\frac{1}{q_{YY}}-1}f(\theta)d\theta+\frac{b}{2}\int_{\bar{\theta}_{OO}}^{\bar{\theta}}\theta^{\frac{1}{q_{OO}}-1}f(\theta)d\theta+\frac{b}{4}\int_{\bar{\theta}}^{1}\left(\theta^{\frac{1}{q_{YY}}-1}-\theta^{\frac{1}{q_{OY}}-1}\right)f(\theta)d\theta$$
$$+\frac{b}{4}\int_{\bar{\theta}}^{1}\left(\theta^{\frac{1}{q_{OO}}-1}-\theta^{\frac{1}{q_{YO}}-1}\right)f(\theta)d\theta>\frac{1}{2}(z-a)\left[F(\bar{\theta})-F(\bar{\theta}_{YY})\right]+\frac{1}{2}(z-a)\left[F(\bar{\theta})-F(\bar{\theta}_{OO})\right],$$

or

$$\frac{1}{2}\int_{\bar{\theta}_{YY}}^{\bar{\theta}}\left(a+b\theta^{\frac{1}{q_{YY}}-1}-z\right)f(\theta)d\theta+\frac{1}{2}\int_{\bar{\theta}_{OO}}^{\bar{\theta}}\left(a+b\theta^{\frac{1}{q_{OO}}-1}-z\right)f(\theta)d\theta$$
$$+\frac{b}{4}\int_{\bar{\theta}}^{1}\left(\theta^{\frac{1}{q_{YY}}-1}-\theta^{\frac{1}{q_{OY}}-1}\right)f(\theta)d\theta+\frac{b}{4}\int_{\bar{\theta}}^{1}\left(\theta^{\frac{1}{q_{OO}}-1}-\theta^{\frac{1}{q_{YO}}-1}\right)f(\theta)d\theta>0,$$

which holds given that all integrals are positive.

C Instructions (Qualtrics)

Figure 12: Welcome

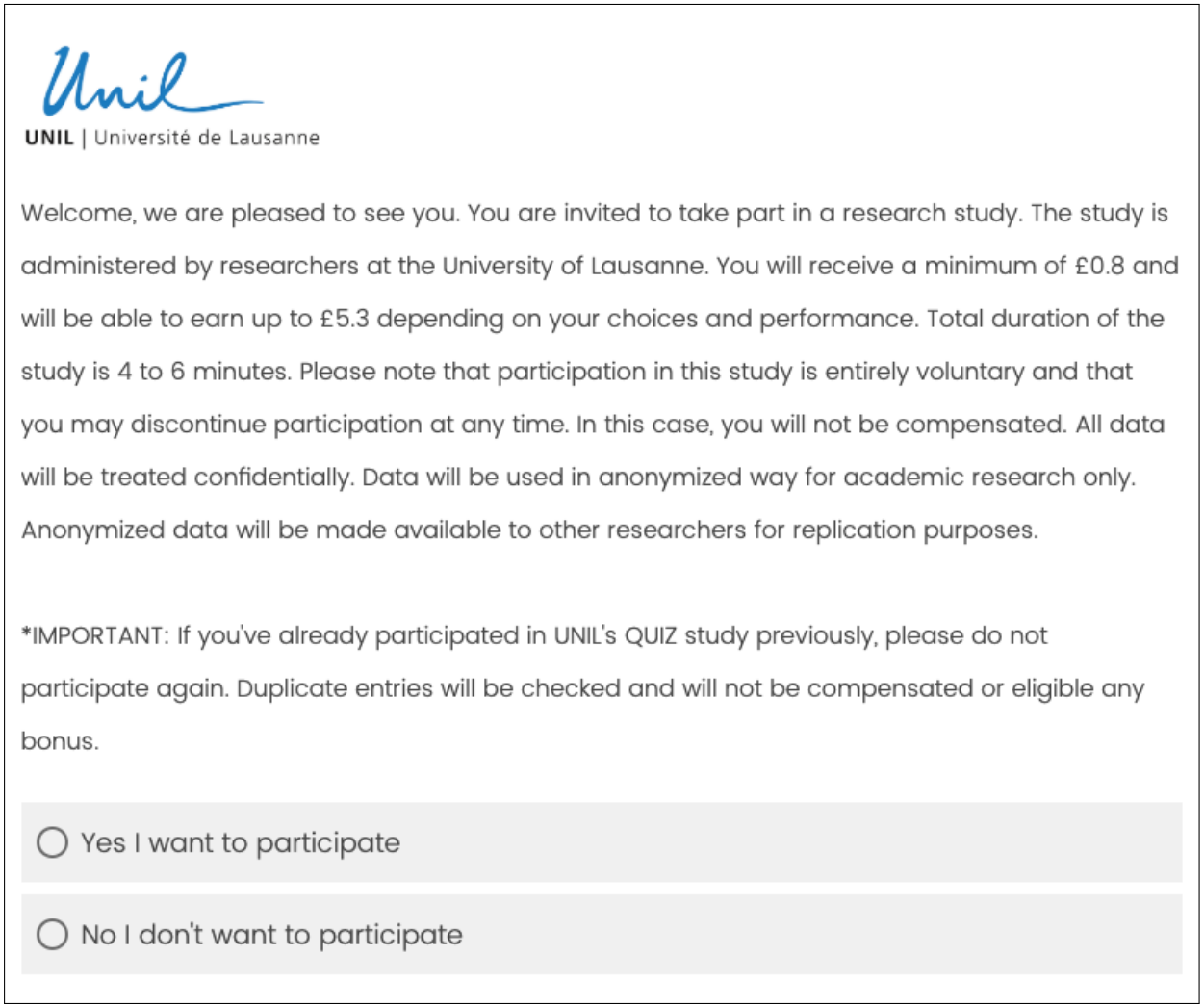


Figure 13: Self Reported Confidence

Figure 14: Main Instructions 1

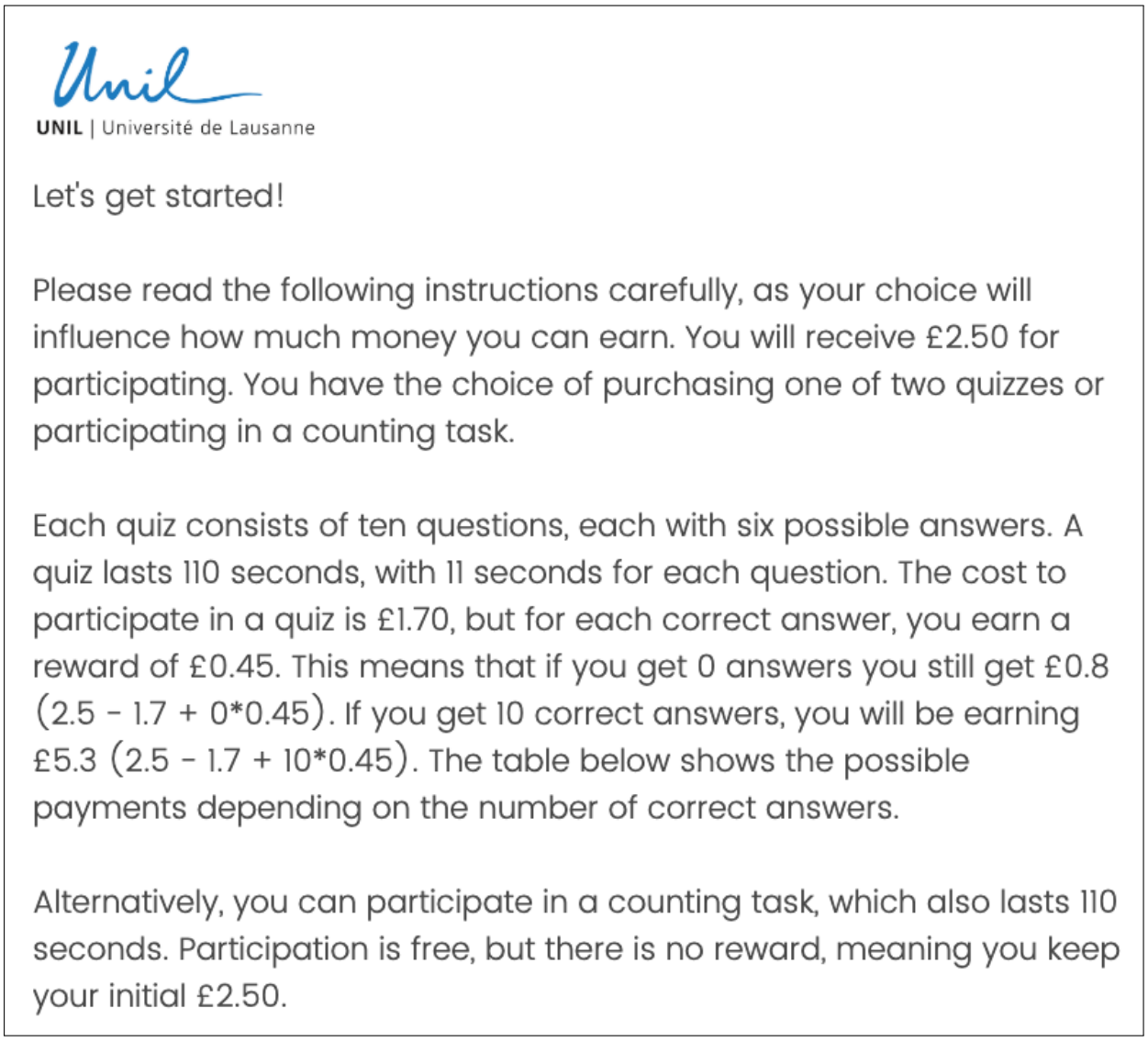


Figure 15: Main Instructions 2

Correct answers	£ reward
0	0.8
1	1.25
2	1.7
3	2.15
4	2.6
5	3.05
6	3.5
7	3.95
8	4.4
9	4.85
10	5.3

*Note: £1 ≈ $1.25

Figure 16: Choice Table

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	3.3 (6)	4.6 (16)	

Which Quiz/Task would you like to enter

- ○ Quiz YELLOW
- ○ Quiz BLUE
- ○ Counting Zeros

Figure 17: Instruction Quiz

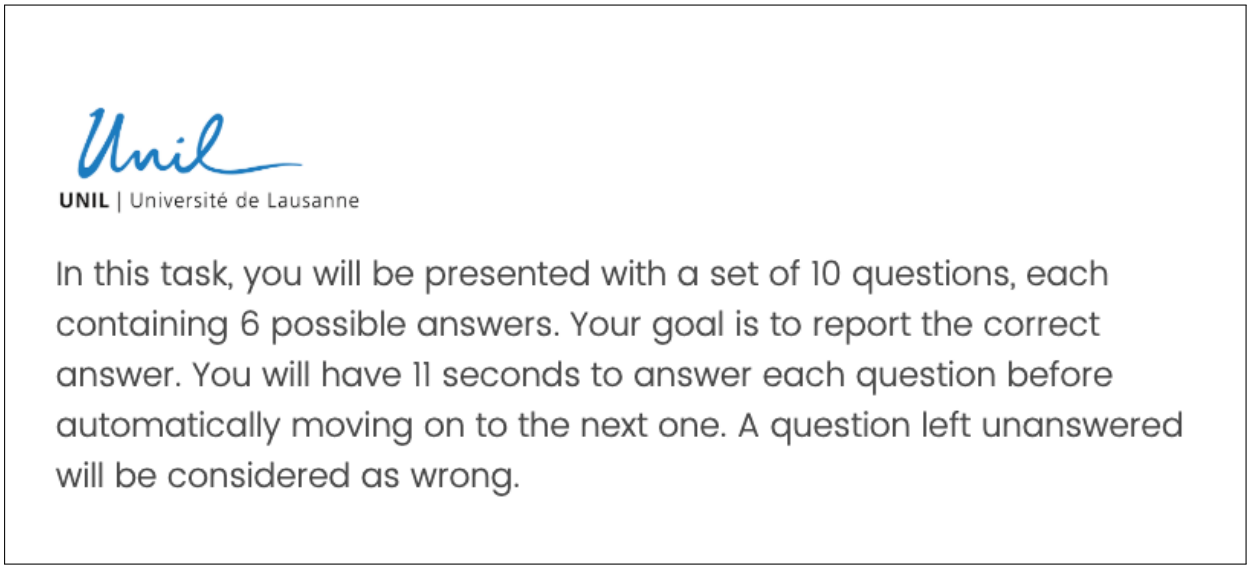


Figure 18: Instruction Counting Zeros

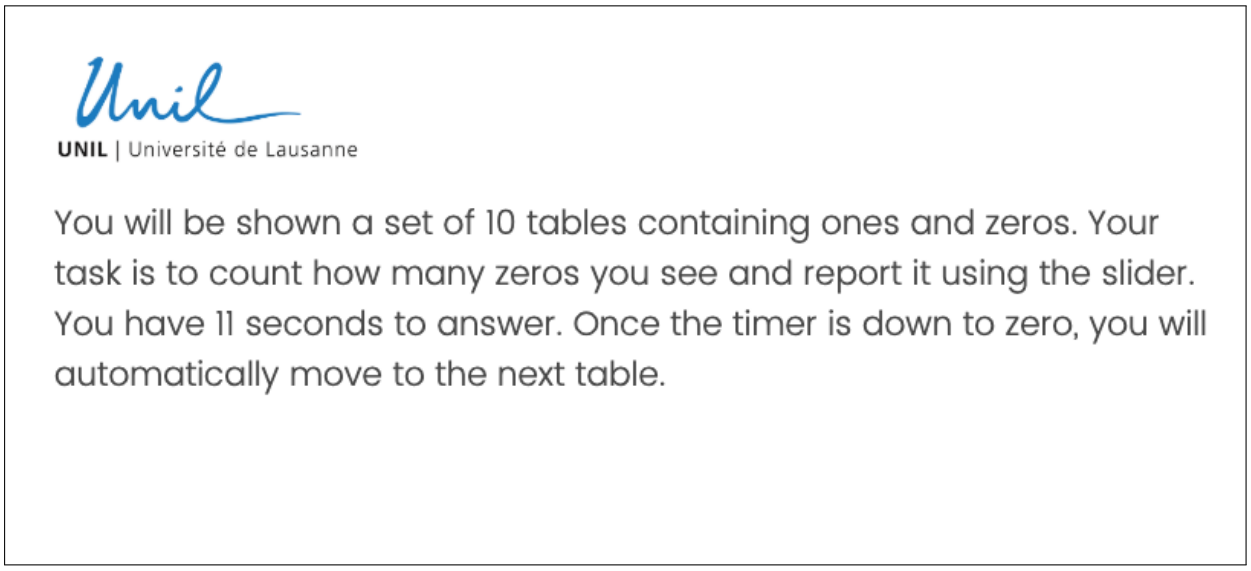


Figure 19: Example Counting Zeros

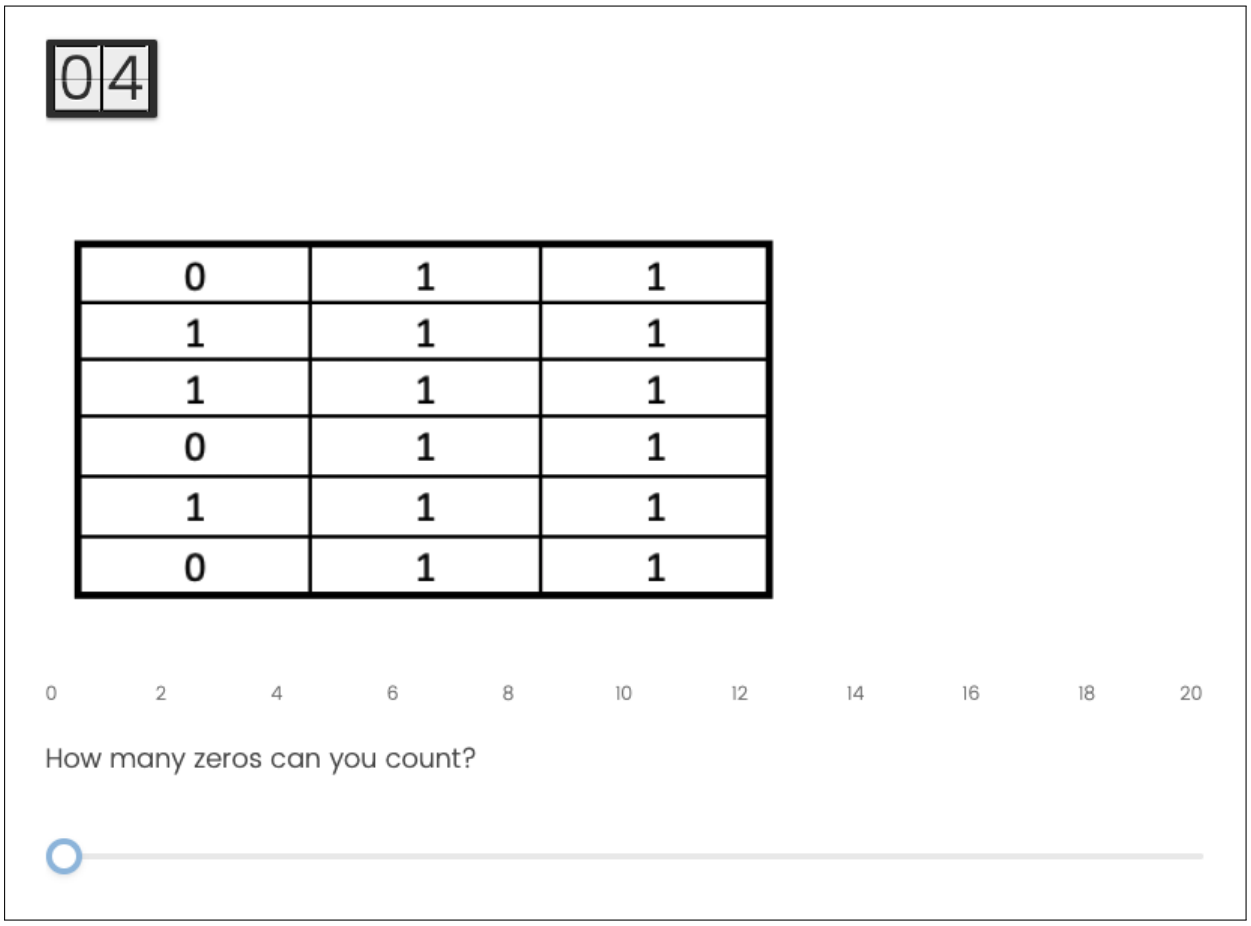


Figure 20: Rating Stage

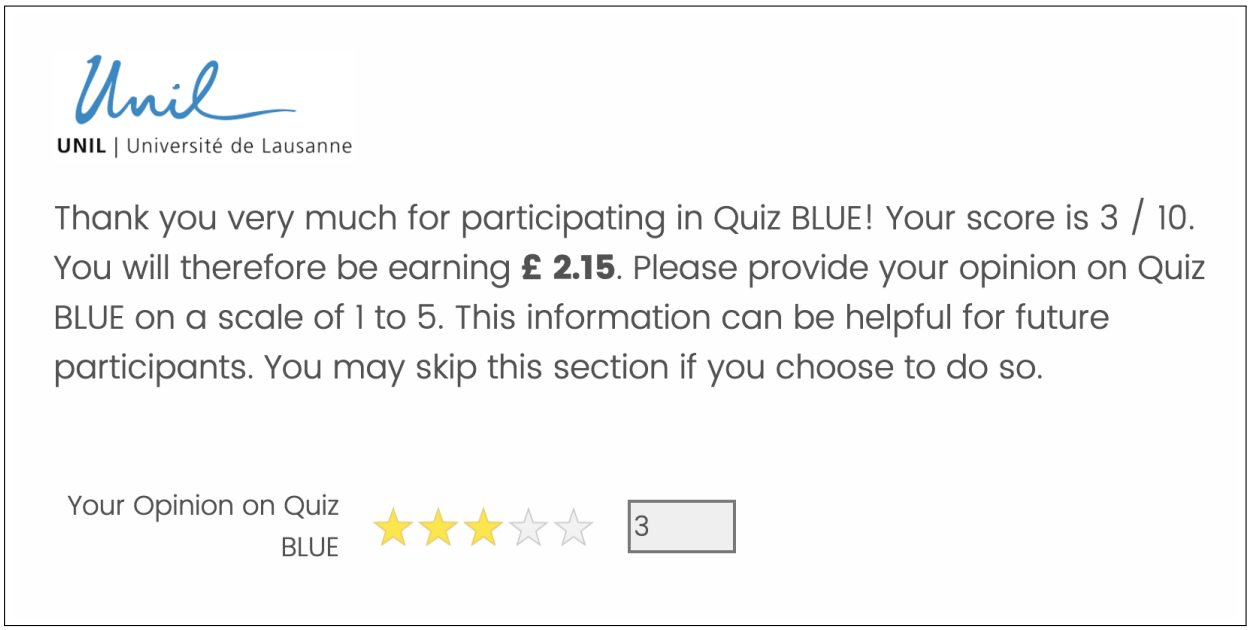


Figure 21: Risk Aversion