



Testing the impact of algorithmic rankings on consumer choice

Authors: Steven Human (BIT) and Adam Jones (BIT)

Research Paper Number 2024/006



February 2024

Main report

Contents

Testing the impact of algorithmic rankings on consumer choice	1
Acknowledgements	2
Main report	4
Contents	4
1. Executive summary	6
1.1 Introduction	6
1.2 Methodology	6
1.3 Key Findings	7
1.3.1 Primary Analysis: Impact of algorithmic designs on consumer product choice	7
1.3.2 Secondary Analysis: Commercially-focused algorithms lead consumers to overspend, compared to random rankings or consumer-focused algorithms	8
1.3.3 Secondary Analysis: Ranking algorithms support market matching of supply and demand, which improves economic efficiency.	8
1.3.5 Feature effects on primary and secondary outcomes	9
1.3.6 Exploratory Outcomes: Sentiment	9
1.3.7 Exploratory Outcomes: Segmentation	10
1.4 Conclusions	10
2. Introduction	11
2.1 Background research	11
3. Methodology	15
3.1 Research aims and Overall approach	15
3.1.1 Research aims	15
3.1.2 Overall approach	15
3.2 Sampling criteria and Recruitment	15
3.2.1 Sampling criteria and Recruitment	16
3.3 Simulation and Algorithm designs	16
3.3.1 Product Database	17
3.3.2 Platform Design	17
3.3.3 Algorithm Designs	18
3.3.4 Additional feature designs	19
3.4 Experiment Design and Trial Arms	20
3.4.1 Experiment Design	20
3.4.2 Trial arms	22
3.5 Ethical considerations	23
3.6 Data collection	24
3.7 Analysis:	24
3.8 Limitations	25
4. Findings	27
4.1 Primary analysis: impact of ranking algorithms on consumer product choice	27
4.1.1 Proportion of respondents selecting the product with the top score	27

4.1.2 Proximity to top-scoring product	28
4.2 Secondary Analysis: financial impact of ranking algorithms	30
4.4 Feature Effects on Primary and Secondary Outcomes	32
4.5 Exploratory Analysis	34
4.5.1 Impacts of algorithmic ranking on search times	34
4.5.2 Sentiment analysis	35
4.6 Additional findings	39
4.6.1 Sub group analysis	39
4.6.2 Segmentation analysis	41
5. Conclusion	43
Appendices	46
Appendix 1: Product Database Development	46
Appendix 2: Platform Design	50
Appendix 3: Algorithm Design	52
Appendix 4: Additional feature designs	55
Appendix 5: Analysis Plan	56
Appendix 6: Statistical Clustering Methodology	59
Appendix 7: Analysis of participants who did not select their top-scoring product	59
Appendix 8: Sample composition and Preferences	61
Appendix 9: Sample Preferences	62
Appendix 10: Algorithm Diagnostics and Additional Summary Statistics	63
Appendix 11: Additional analysis of proximity score variable	64
Appendix 12: Sub group analysis	64

Acknowledgements

The authors of the Behavioural Insights Team, Steven Human and Adam Jones, acknowledge the support by Dr. Janna Miletzki and Bruno Galizzi of the Department for Science, Innovation and Technology to add to the body of research around business practices that can help consumer wellbeing.



© Crown copyright 2024

This publication is licensed under the terms of the Open Government Licence v3.0 except where otherwise stated. To view this licence, visit nationalarchives.gov.uk/doc/open-government-licence/version/3 or write to the Information Policy Team, The National Archives, Kew, London TW9 4DU, or email: psi@nationalarchives.gsi.gov.uk.

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned.

Any enquiries regarding this publication should be sent to us at: alt.formats@dsit.gov.uk

1. Executive summary

1.1 Introduction

Digital markets and technologies have profoundly changed the consumer experience, allowing consumers to make a choice from a range of products (and services) on e-commerce platforms that are wider than ever before. Online retailers have adopted ranking algorithms among other practices to help users navigate vast catalogues of products. The purpose of these algorithms is to sort product information for relevance and quality, for example by personalising the user experience by suggesting products that are relevant to their interests, needs, and preferences. Alongside the benefits that these practices can bring, consumers may also be subject to new and amplified risks, not least due to the scale of relevant digital markets, which shall be explored in this research.

Building on research into digital consumer issues by DCMS¹ and the CMA,² the main aim of this novel experimental research was to understand and quantify positive and negative impacts of three types of algorithmic design on consumer choice and the economy in a simulated e-commerce environment .

The findings of this research may provide valuable insights for online platforms and policymakers in understanding the potential impact of these algorithms on consumer choice. The findings should ultimately benefit consumers by helping enhance fair and open competition in digital markets where market power is concentrated in a small number of large tech firms.

1.2 Methodology

We designed an eight arm randomised controlled trial (RCT) with a total sample size of 8,009 respondents (n ~ 1,000 per arm) to test the impact of algorithmic practices on consumer behaviour in a simulated e-commerce environment. We focused on the earphones/headphones market (see Appendix 1), generating a list of 432 products simulating a normal distribution in line with the range we observed in the e-commerce environment.

When entering the experiment, we captured participants' preferences about earphones/headphones, as well as their willingness to pay. We used this information to present the products in different orders, simulating the effects of various algorithms including a random ranking:

¹ Digital consumer issues research 2023: <https://www.gov.uk/government/publications/digital-consumer-issues-research>

² CMA 2021: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/954331/Algorithm_++.pdf

- **Random ranking:** This presented products completely at random and served as a control.
- **Consumer-focused algorithm:** Aim of maximising consumer utility by optimising value and products matching the consumers' preferences.
- **Commercially-focused algorithm:** Aimed to maximise revenue for e-commerce websites by maximising total spend while still considering individual consumer preferences.
- **Income-based algorithm:** Aimed to maximise revenue for e-commerce websites by maximising total spend, based on household income which was used as a proxy for personal income and willingness to pay. This algorithm did not take into account consumer preferences.

Consumers were able to browse products at will, and were asked to choose the product they would purchase using the budget available to them. Products were assigned a score which was determined by the consumers' stated preferences for each participant. The product achieving the highest score for each participant is referred to as the top-scoring product.

We obtained consent³ on processing their non-personally identifiable data and responses for research purposes from all participants through the panel provider and ensured their anonymity and confidentiality. We followed ethical guidelines in conducting the study, including ensuring that participants had the right to withdraw from the study at any time and that their participation was voluntary. This study has a number of limitations in terms of external and internal validity which we have sought to mitigate. This is covered in detail in section 3.8.

1.3 Key Findings

1.3.1 Primary Analysis: Impact of algorithmic designs on consumer product choice

While both consumer- and commercially-focused algorithms support consumer choices more than random rankings, a consumer-focused algorithm does this more effectively than a commercially-focused one.

- 19% of participants who saw consumer-focused rankings selected the product that best satisfied all of their stated preferences⁴, which was statistically significantly higher than all other algorithms.⁵
- These findings are amplified when including a broader range of top-scoring products (products with high proximity scores, see below). We measured the closeness to the selected product by using a proximity score to the product that best satisfied their preferences, where lower scores indicate more distance from the top-scoring product.

³ Consent is covered in more detail in section 3.2.1 and 3.5.

⁴ This is based on the overall score assigned to a product by the consumer-focused algorithm, we use this score as the top score in all arms.

⁵ Those who saw a random algorithm selected the top-scoring product 2.6% of the time, those in the income-based ranking 1.5% of the time, and those in the commercially-focused algorithm 6.6% of the time.

- The consumer and commercially-focused algorithms achieved significantly higher scores of -0.80 and -1.12 respectively, whereas the income-based algorithm and the random ranking resulted in products that scored lowest with -4.82 and -3.42 respectively. This indicated that consumers were selecting products that were further from the top-scoring product when these algorithms were used

1.3.2 Secondary Analysis: Commercially-focused algorithms lead consumers to overspend, compared to random rankings or consumer-focused algorithms

Small variations in algorithm design can have a significant impact on consumers' financial wellbeing, whilst random rankings may be economically inefficient in the headphone market.

- The study analysed the financial impact of different algorithm designs, where we considered overspend compared to their top-scoring product to be not in the best interest to the consumer and underspend to be economically inefficient in the headphone market.
- The random ranking resulted in a considerable underspend (spending less than the value of the top-scoring product) averaging £15.60. The other algorithms resulted in small to large amounts of overspend, the highest being in the income-based algorithm with an average overspend of £17.17; then £7.96 in the commercially-focused algorithm and £3.41 in the consumer-focused algorithm.
- Using the difference in spend between the consumer-focused and commercially-focused rankings (£4.55) in this experiment and the total value of the UK headphone market of £490 million ([Statista Headphone Category Revenue Data](#)), we estimate the hypothetical annual overspend in the UK headphone market could be ~£46m when applying the commercially-focused algorithm.⁶⁷ Furthermore, we estimate the hypothetical annual overspend of the income-based algorithm could be ~£141m more compared to the consumer-focused algorithm based on this experiment.

1.3.3 Secondary Analysis: Ranking algorithms support market matching of supply and demand, which improves economic efficiency.

Random rankings may lead to economic inefficiency in the headphone market through mismatched supply and demand.

- While the underspend resulting from the random ranking is not necessarily harmful to a consumer, consumers may be less satisfied and there can be implications for the economy if the consumer is settling for an inferior good due to search frictions.
- Of those who saw the random ranking and did not select the top-scoring product, we found that 70% were willing to switch to the more expensive top-scoring product they were shown. In comparison, only 45% of those who saw the consumer-focused ranking would switch to a more expensive product.

⁶ For this calculation, we are using the difference in spend between the consumer-focused and commercially-focused rankings (£4.55) and the total value of the UK headphone market of £490 million ([Statista Headphone Category Revenue Data](#)).

⁷ Total overspend = ((commercially-focused overspend - consumer-focused overspend) * (Participants in consumer-focused arms)) / (Market value in consumer arms) * (Total UK market value)

- Using the average underspend for those that weren't satisfied and the total UK headphone market value of £490 million, we estimate the hypothetical annual value of economic loss to the headphone market by presenting a random ranking instead of a consumer-focused ranking could be ~£159 million⁸ based on this experiment.

1.3.5 Feature effects on primary and secondary outcomes

Features such as transparency messaging, sponsored products, platform recommendations, and sorting functionalities do not have a significant impact on primary and secondary outcome measures in this experiment; that is: respondents select products of similar characteristics and prices when they interact with these features and when they don't. This suggests that the order of the ranking has a stronger effect on consumer choice when compared with other features.

- In addition, we find that participants were significantly less likely to select a 'sponsored' product than one with no feature or a platform-recommended product. 30% of people selected a sponsored product, vs 35% and 38% selecting a platform recommended product or 'no feature' product in the arm without these features⁹.

1.3.6 Exploratory Outcomes: Sentiment

- **The majority of consumers found our simulation realistic, a fact which supports the external validity of our results.** A majority (69%) found the website realistic, regardless of algorithm design, suggesting a high external validity of the simulation.
- **Ranking algorithms impact sentiment even if the consumers are unaware of the ranking mechanism.** Those including consumer preferences resulted in a more positive sentiment, such as easier to use and less frustrating.
- **Sentiments to the commercially-focused and income-based algorithms become more negative once consumers are explicitly aware of the ranking mechanism.** Upon revealing the ranking algorithms, users responded most positively to the consumer-focused algorithm, perceiving it as fair, ethical and trustworthy. The income-based algorithm made users feel the most uncomfortable.
- **Consumers struggle to identify underlying algorithms:** only 40%-50% of users said they were able to discern the algorithm from the rankings, even after explaining the ranking mechanism.

1.3.7 Exploratory Outcomes: Segmentation

Using a clustering analysis we found three groups of consumers who are statistically different. Among them, we found that the elderly, those shopping online less frequently, live in rural areas and have the lowest confidence in their digital skills are more vulnerable to the impact algorithms have on them than the other groups.

- These vulnerable online consumers experienced the most financial harm in the form of overspend and had the lowest percentage of selecting the top-scoring product.

⁸ Total loss from random ranking = ((average underspend in random arm) * (Participants in consumer-focused arms))/(Market value in consumer arms) * (Total market value)

⁹ Sponsored and platform recommendation messaging was placed on products listed in positions 1, 2 and 5 in the search results. When making comparisons to the standard arm with no features, we compared the percentage of people selecting products at position 1,2, or 5.

1.4 Conclusions

In conclusion, this study highlights the significant influence of algorithm choice on consumer choice, with the consumer-focused algorithm outperforming the others with respect to the main outcomes measures and the commercially-focused algorithm performing better than both the random ranking and the income-based algorithm.

- **The design of algorithms applied to ranking and recommendations has a significant impact on consumer choice**, supporting digital consumers to find products meeting their preferences.
 - **Digital consumers can find disadvantageous outcomes** when these designs do not take consumer preferences into account, including overspending or settling for lower quality products.
 - We estimate a commercially-focused ranking can lead to a **hypothetical annual overspend of ~£46m** when compared to a consumer-focused algorithm in the UK headphone market. Furthermore, the income-based algorithm can lead to a **hypothetical annual overspend of ~£141m** compared to the consumer-focused algorithm.
 - **Vulnerable digital consumers**, such as the elderly and those shopping online less frequently, **can get more significant detrimental outcomes** when exposed to these algorithmic practices applied to ranking and recommendations.
 - The study also revealed that **random rankings can lead to a large underspend** by consumers, potentially resulting in reduced satisfaction and implications for the economy. **This can lead to a hypothetical economic loss to the headphone market of approximately ~£159 million.**
 - **These results emphasise the importance of algorithmic designs that align with consumer preferences and promote economic efficiency.** By considering these findings and optimising algorithm designs, policymakers, regulators, and online platform retailers can enhance consumer well-being, foster trust, and create a fair and efficient online marketplace.
-

2. Introduction

The exponential growth of e-commerce has led online platform retailers to adopt ranking and recommendation algorithms to help users navigate vast catalogues of products. The purpose of these algorithms is to personalise the user experience by suggesting products that are relevant to their interests, needs, and preferences. However, concerns have been raised about the potential impact of these algorithms on consumer choice, consumer experience and their overall economic efficacy. In this research we take an algorithm to mean a sequence of instructions to perform a computation or solve a problem.

2.1 Background research

2.1.1 What constitutes fair and unfair settings?

Our review of the literature has found several aspects that contribute to the fairness, or lack thereof, of algorithms. The CMA (2021)¹⁰ defines an unfair ranking algorithm as one that modifies rankings or other design features to influence what a consumer sees for the platform's own commercial advantage, while ultimately degrading or misrepresenting the offerings to a consumer. They break this down further into two ways in which platforms may use unfair ranking and design. Firstly by manipulating rankings to favour certain options, either because of certain commercial relationships, or by preferentially showing their own products. An example of this could be an online platform preferentially promoting their own electronics on their platform compared to others. Platforms may also use unfair design practices to exploit behavioural biases for commercial gain. An example of this could be using misleading scarcity messages, which exploit consumers' tendencies toward loss aversion.

Pitoura et al¹¹ extend the notion of fairness to a context with a lack of discrimination, meaning similar consumers should be treated similarly. Operationalising this requires a means of quantifying both input similarity (i.e. how do you quantify similarity of users?) and output similarity (i.e. what is similar treatment?). They also suggest 'disparate treatment' as a form of unfair differentiation based on protected characteristics. This could include things such as race or sex. In addition other demographics like a consumer's medical history, and income could be used to the same end.

In a similar vein, the extent to which an algorithm is 'unfair' can be measured by how much harm is caused by it. Essentially, by presenting different prices to different consumers, personalisation can affect vulnerable consumers, and result in unfair distributive effects (CMA Paper, 2018¹²). It is worth mentioning too that this personalisation, if used correctly, may have beneficial impacts.

¹⁰ [CMA \(2021\)](#)

¹¹ [Pitoura et al](#)

¹² [CMA Paper \(2018\)](#)

More recent work¹³ has argued that the fairness of ranking systems corresponds to how they allocate exposure of items based on their merit (i.e. relevance to the query). A model proposed by Pitoura et al sets out how fairness can be defined for ranking and recommendation algorithms:

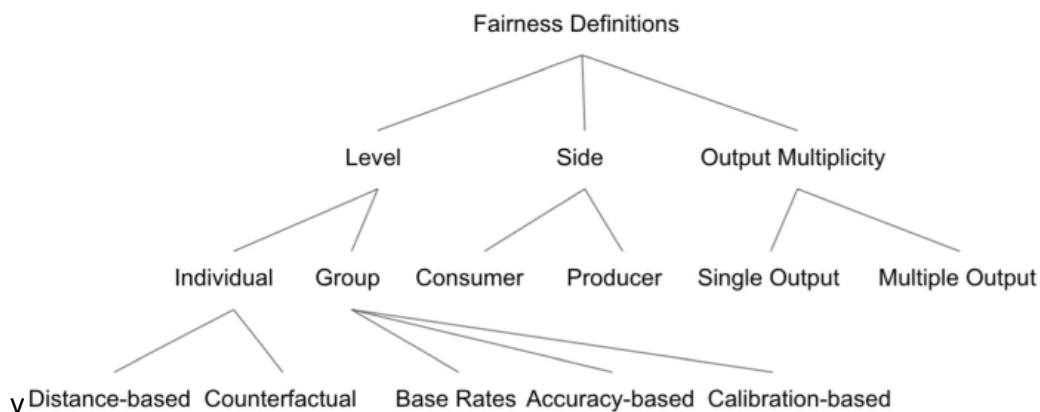


Figure 1: Pitoura et al fairness model for ranking and recommendation algorithms

- **Level.** Individual fairness definitions are based on the premise that similar entities should be treated similarly. Group fairness defines group entities based on the value of one or more protected attributes, ethnicity or sexual orientation for example, and asks that all groups are treated similarly.
- **Side.** Producer or item-side fairness focuses on the items that are being ranked or recommended (e.g. similar items or groups of items to appear in similar positions in a ranking). Consumer or user-sided fairness focuses on the users who receive or consume the items in a ranking (e.g. similar users, or groups of users, to receive similar rankings or recommendations). There may also be additional stakeholders involved, for example providers in a recommendation system (e.g. in an online craft marketplace, we may want to ensure market diversity and therefore treat the provider as a protected attribute).
- **Output multiplicity.** The authors distinguish between single output and multiple output fairness. Multiple output fairness strives for eventual or average consumer/producer fairness (e.g. consumers or producers are treated fairly in a series of rankings/recommendations as a whole, although they may be treated unfairly in one or more single ranking or recommendation in the series).

2.1.2 Personalisation in algorithmic practices

A key aspect of algorithmic design and outputs is personalisation. Personalisation of ranking and recommendation algorithms is possible through various avenues; businesses may use information that is observed, volunteered, inferred, or collected to present different products or pricing to different consumers or groups of consumers.

Personalisation may be beneficial to consumers or the market, for example targeted discounts might help new entrants to compete better in markets with high switching costs

¹³ [Other recent works](#)

which in turn increases competition, driving better innovation and pricing for consumers. However it can also lead to consumer harm, particularly when it is complex, lacking transparency or generally causes consumers to lose trust in online markets. One example of this is the practice of setting different prices for the same product, known as price discrimination.

There are three different types of price discrimination:

- **Perfect (first degree)** discrimination refers to when businesses can accurately determine what each customer will pay for a specific product or service and then sell it for that price. Client services industries often practise this kind of discrimination, as different prices are offered to different clients for the same service.
- **Second degree discrimination** refers to providing options which induce customers to self-select different effective prices (e.g. quantity discounts, coupons, buy-one-get-two offers).
- **Third degree discrimination** occurs when companies price products and services differently based on the unique demographics of subsets of their consumer base, such as students, military personnel, or older adults.

To date, there is limited empirical evidence of personalised advertised pricing in action, potentially due to potential consumer backlash. Businesses therefore may use other techniques to personalise prices that are harder for consumers to detect, such as loyalty programs or promotional offers. Personalisation may also be apparent in ranking of search results, where businesses use information such as the user's location, previous queries, and previous browsing/purchase behaviour to decide which results to display, and in what order. Consumers may therefore be manipulated into making a choice which is more profitable for the business, which they wouldn't have made under more neutral conditions. Personalisation can lead to additional consumer harm if businesses use (unwittingly or otherwise) categories that are correlated with consumer vulnerability and protected characteristics.

Another potential outcome of personalisation is 'price steering', which, although indirect, can yield similar results as personalised pricing. This practice involves presenting higher-priced products to consumers with a greater willingness to pay, thereby capitalising on their propensity to spend more. Personalised rankings and search results are already widespread in the e-commerce landscape. In 2018, the European Commission conducted a study that revealed 61 percent of the 160 e-commerce websites reviewed employed personalised ranking of search results.¹⁴ This type of practice can be harmful to consumers, as they could end up paying more for a product they don't actually need.

In conclusion, while personalised advertised pricing remains relatively limited in practice, businesses are increasingly employing subtler techniques such as price steering to cater to individual consumer preferences and willingness to pay. It is essential to understand the implications of these practices on consumer behaviour and the potential harm they may cause, especially to vulnerable consumer groups.

¹⁴ [Online market segmentation through personalised pricing/offers in the European Union](#)

2.1.3 Non-personalised additional features

There are also potential consumer harms that relate to (non-personalised) unfair ranking and design of online services. The position of the “buy” button on a shopping website, the colour of an information banner and a default payment method can all be examples of online choice architecture based on algorithms. Online choice architecture (OCA) can also involve preferencing others for commercial advantage, whereby firms may offer to pay more money to a platform in exchange for the platform giving more prominence to their options or otherwise distorting the ranking algorithm. These effects can also occur on mobile devices, where an increasing proportion of online shopping is taking place Wang et al (2015)¹⁵.

Online choice architecture

Businesses may also exploit behavioural biases (e.g. limited attention, loss aversion, inertia, or susceptibility to default options) to cause consumer harm, leading consumers to make purchasing decisions that they would not make under different OCA. These harmful user interface design choices are known as online choice architecture. Matur et al.¹⁶ have identified 7 categories of OCA (sneaking, urgency, misdirection, social proof, scarcity, obstruction and forced action). For example, scarcity messages can create a sense of urgency and lead to customers buying more, while spending less time to search.

Businesses can also manipulate competitive markets using exclusionary practices, whereby a dominant platform uses OCA that favour its own products and services (i.e. self preferencing), and/or exclude competitors. Examples of this include placing ‘sponsored’ or platform recommended products more prominently and higher up on search result pages, both of which were tested in this study. This could lead to harm to both the economy and the consumer, whereby the consumer purchases a product they are less satisfied with, and where the economy is less efficient, with potentially inferior goods being promoted using OCA.

Transparency

Transparency has been found to play a significant role. Veltri et al¹⁷ performed 3 different studies in 4 different countries, examining the impact of increased transparency in the presentation of online search information, details of contractual entities and the implications for consumer protection and user reviews and ratings would affect consumers’ choices with regard consumer ratings, to on restaurant bookings, hotel bookings, and cell-phone purchases. They found that increasing transparency increased the probability of participants selecting a product. Conversely, Ghose et al (2014)¹⁸ found that providing more information during the decision making process may decrease product selection due to information overload. This study however was conducted in the context of search engine results.

¹⁵ [Wang et al](#)

¹⁶ [Matur et al](#)

¹⁷ [Veltri et al](#)

¹⁸ [Ghose et al \(2014\)](#)

3. Methodology

In this section, we describe the methodology used in our study. We present our research aims and approach first; we then detail our product selection and pricing design, followed by the design of our algorithms, and the design of our RCT. We then discuss our analytical approach, ethical considerations, and the limitations and additional considerations of the study.

3.1 Research aims and Overall approach

3.1.1 Research aims

We aimed to address the following research questions:

1. *What is the impact of 1) a consumer-focused ranking algorithm, 2) a commercially-focused ranking algorithm, & 3) a commercially-focused ranking algorithm using personal income data on consumer choice in a simulated e-commerce environment?*
2. *What is the impact of additional 1) saliency of platform recommended products (iWeb's choice) , 2) saliency of sponsored products, & 3) transparency features in ranking algorithms¹⁹ on consumer choice in a simulated e-commerce environment?*

3.1.2 Overall approach

To answer our research questions on the impact of different algorithms and features on online consumer behaviour, we needed to design a simulated e-commerce environment that mimicked real-world online shopping experiences. We also needed to design bespoke algorithms with various underlying objectives and reproduce features commonly seen on real e-commerce sites. This simulated environment and the underlying algorithms and features were then embedded into an experimental survey with a randomised controlled trial (RCT) design, which allowed us to measure the impact of each algorithm on consumer behaviour. This approach enabled us to mimic real world shopping experiences and gain insights into how different algorithms affect the decisions consumers make when shopping online.

3.2 Sampling criteria and Recruitment

Our recruitment criteria was designed to include consumers who would be impacted by algorithmic practices (those who use e-commerce shopping platforms) and those to whom the experiment would be valid (users of headphones/earphones - please see an explanation of why we focus on this market below under 3.3.1 Product Database).

¹⁹ These business practices were chosen based on prevalence, estimated severity, evidence gap (especially in terms of quantifying severity), and perceived opportunity to influence live HMG policy. Within the design, DSIT would like to understand how factors such as personalisation, saliency of specific options (top choice, sponsored), or transparency influence the effectiveness of rankings or recommendations.

3.2.1 Sampling criteria and Recruitment

All respondents were required to meet the following criteria:

- Live in the UK and are older than 18
- Use online platforms
- Owner of headphones/earphones²⁰

All **8,009** respondents were recruited through a panel aggregator²¹. Respondents registered on panel supplier²² websites connected to the panel aggregator network were invited to participate through the supplier's portal or through a notification from the supplier directly to the respondent. Suppliers attain consent through the sign-up process from all participants. Respondents were given a link to the experiment which was hosted on BIT's proprietary online experiment platform Predictiv.

The aggregator corresponds with panel providers (market research organisations) to source potential participants, who are individuals that have signed up to participate in online surveys. When participants agree to participate in research through this aggregator, they typically provide basic demographic information that is required across research projects. This is not considered granular enough to identify an individual. This basic demographic data will be held by the aggregator for 30 days for the purposes of participants taking part in other research projects. After this, this data will be destroyed and participants will have to respond to the questions again. If participants who have clicked on the link meet BIT's screening criteria, then they will be able to access the Predictiv survey.

Participants were paid ~70p²³ for completing the experiment. The amount participants are paid for participation was at or above the average incentive across the panel aggregators payments, according to the target sample and time required to complete the experiment. We monitored balance throughout data collection and achieved UK online representativeness in this sample.

3.3 Simulation and Algorithm designs

This section of the paper focuses on the development of the simulation used in the study. The simulation was designed to create a realistic e-commerce environment where participants could browse and purchase products which required four key building blocks, a product database, a functional e-commerce platform, product search algorithms, and additional features in various treatment arms. A brief overview of each of these will be given here, and full details can be found in Appendix 1-4.

²⁰ The product selection rationale is outlined in detail in section 3.3.1

²¹ A panel aggregator is a company that gathers data from multiple online survey panels and combines it into one large database for use by market researchers. By combining data from multiple panels, panel aggregators can offer researchers a larger pool of potential respondents and a more diverse sample.

²² These survey panels are groups of people who have agreed to participate in online surveys for a reward, such as cash or gift cards.

²³ While we provide a recommended incentive per participant we are unable to determine the exact payment after the aggregator and panel providers distribute incentives.

3.3.1 Product Database

A pivotal aspect of our experiment design was the creation of a product database for use in the simulated e-commerce environment, looking at the headphones/earphones market (see Appendix 1 for more details on this choice). We decided to work with the headphones/earphones market as:

- Preferences are limited and easy to elicit (design and analysis simplification)
 - Almost everyone has ear-phones (drives high incidence rate²⁴ in survey)
- Few attribute levels (design and validity simplification)
- Most attributes are technical rather than subjective (noise cancellation = better earphones)
- Limited price range allowing us to keep the experiment design more simple
- Prices are low enough that consumers may not be as deliberate in purchasing as they would be with a phone meaning algorithms may be more impactful.

To create a product database, we generated a database of 432 pairs of headphones with physical attributes, pricing, images, and names. The prices were developed to mirror real-world headphone prices. We then used six product attributes: the earphone/headphone type (over ear, in ear wired, ear bud), noise cancellation functionality, sound quality, connectivity (wired, bluetooth), for bluetooth products battery life, and for wired products wire material quality. Product prices were also higher or lower depending on some of the product attributes. Noise cancellation, sound quality, battery life and wire quality were all factored into the prices of the products. The key factors in choosing the headphone market were a) that the market is one most consumers are a part of, b) we see close to perfect competition in this market, as the market has a low barrier to entry and many providers, and c) the products have definable, rankable attributes when compared to a market like clothing (see Appendix 1 for further details).

3.3.2 Platform Design

We developed an e-commerce platform named 'iWeb' that aimed to replicate traditional e-commerce platforms as realistically as possible. The platform included advertisements, a logo, a stationary search bar, and showed 15 products per page, along with their image, name, and price, as well as other attributes like noise-cancellation, sound quality, and battery life. Participants were given a budget based on what they said the maximum they would be willing to pay for headphones/earphones would be. We elicited this maximum budget directly and inflated it by 5% to allow for overspend. When selecting a product consumers were restricted to selecting only one product, with an error message appearing if they tried to buy a product that exceeded their budget. The platform was compatible with both mobile and desktop devices, with the only difference being the absence of advertisements on the mobile version. Full details of the platform design can be found in Appendix 2.

²⁴ The percentage of those who enter a survey that are eligible to complete the survey.

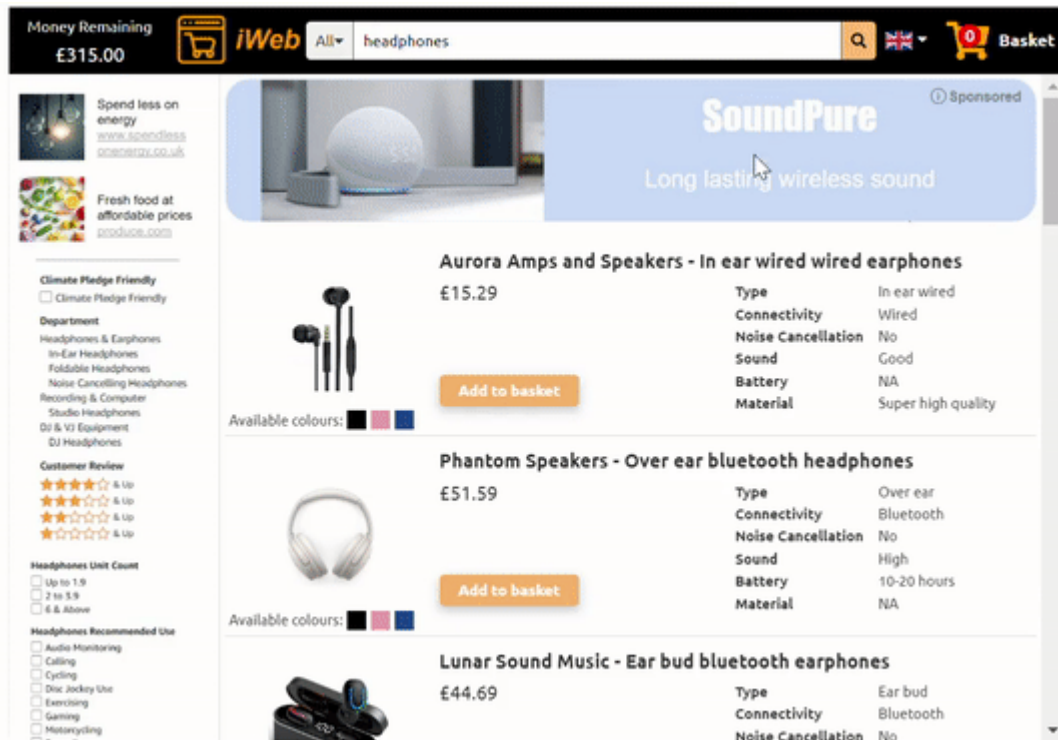


Figure 2: Overview of the e-commerce simulation

3.3.3 Algorithm Designs

We designed three algorithms to manipulate the display order of products in addition to a random control ranking. These algorithms made up the base of our 8 trial arms, and include:

- **Treatment 1: A completely random ranking.**
- **Treatment 2: A consumer-focused algorithm** that reflects the stated preferences and budgets we elicited from participants before the simulation. Preferences and budgets are described in section 3.4.1.
- **Treatment 4: A commercially-focused algorithm** that favours more expensive products, while still reflecting participants' stated preferences and taking budget into account.
- **Treatment 8: An income-based algorithm** that uses household income to personalise the products participants were shown, grouping products based on price points making assumptions on how much people would be willing to spend on headphones given participants' household income, with no budgetary restrictions and without taking into account their stated preferences.

Full details of how the algorithms were developed can be found in Appendix 3

3.3.4 Additional feature designs

In addition to the algorithms developed, we implemented four additional features to the e-commerce platform on specific arms:

- **Transparency Messaging (Treatment 3)**

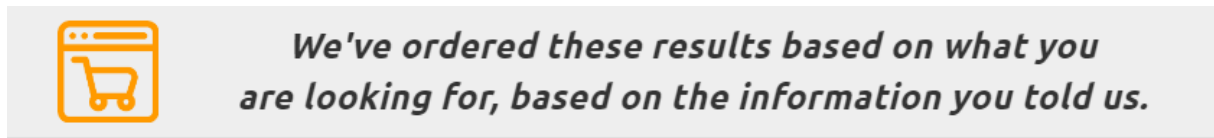



Figure 3: Transparency message shown at the top of the simulation

- **Sponsored Messaging (Treatment 5) - shown on products at position 1,2 and 5 in the ranking**

Sponsored **Echoes Guitars & Equipment - Over ear bluetooth headphones**



£59.49


Type	Over ear
Connectivity	Bluetooth
Noise Cancellation	No
Sound	Good
Battery	More than 20 hours
Material	NA

Available colours: Add to basket

Figure 4: Example of 'sponsored' messaging

- **'iWebs' Choice Messaging: Platform Recommendation (Treatment 6) - shown on products at position 1,2 and 5 in the ranking**

iWeb's choice **Vortex Music - Over ear wired headphones**



£51.19

Type	Over ear
Connectivity	Wired
Noise Cancellation	No
Sound	Super High
Battery	NA
Material	Super high quality

Available colours: Add to basket

Figure 5: Example of 'iWebs' choice messaging

- **Sort Function (Treatment 7)**

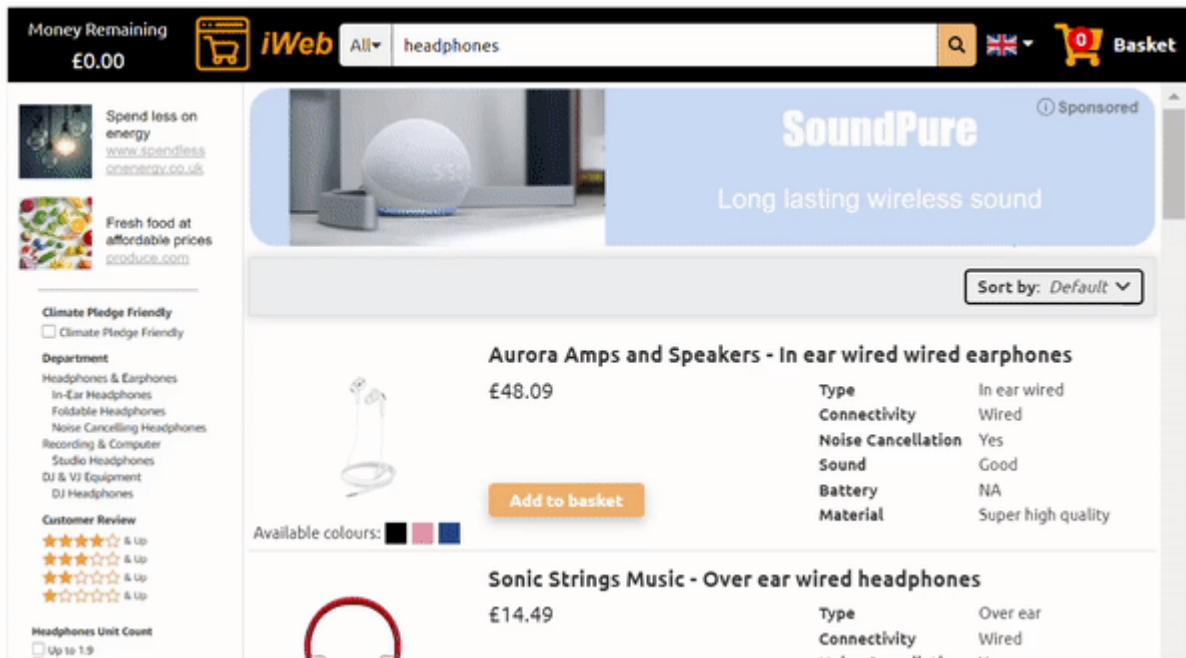


Figure 6: Overview the sort function

A fuller description of these features can be found in appendix 4, and an overview of the trial arms in section 3.4.2.

3.4 Experiment Design and Trial Arms

3.4.1 Experiment Design

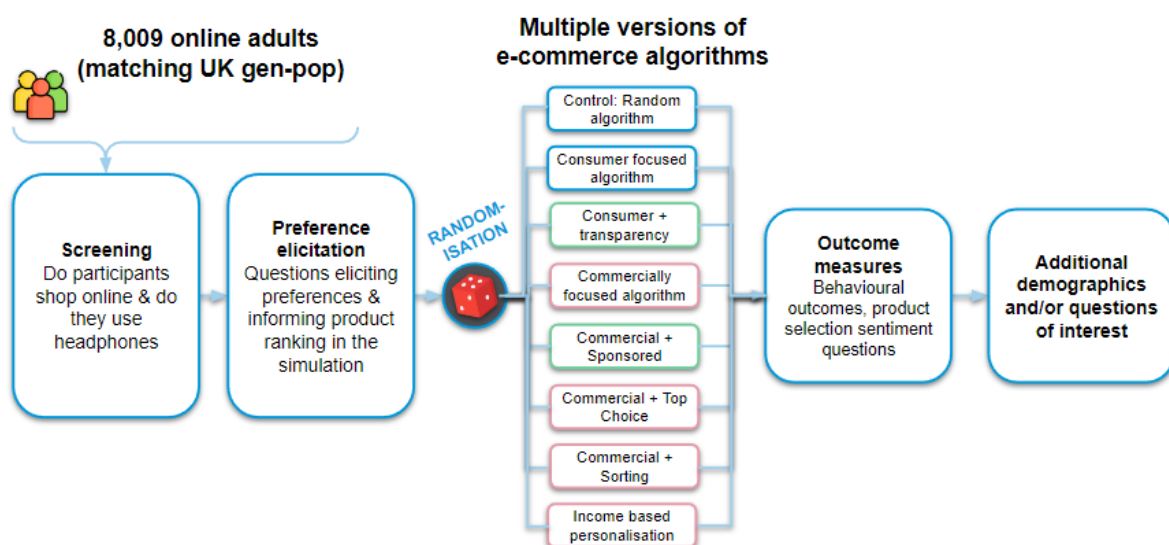


Figure 7: Experimental design flow

Once participants passed the screening criteria, attention checks²⁵ and reached the first page of the Predictiv platform, they were given an introduction to the experiment. We then elicited headphone/earphone preferences from participants and their willingness to pay for these products. These variables were used to customise the simulation and present their budget. Table 1 outlines the questions we asked to elicit preferences from participants. Once we elicited these responses, participants entered the simulated e-commerce environment. Participants were randomly assigned to one of 8 arms within the simulated e-commerce environment. Here they were free to browse products and select the product they feel is best for them within the budget they had available. A summary table of participants' preferences can be found in Appendix 9.

Table 1: Preference and budget elicitation questions and answer options

Preference elicited	Question	Answer Options
Headphone type preference	When listening to music/podcasts using headphones, which of the following headphone types do you prefer?	<ul style="list-style-type: none"> - Over-the-ear headphones - In-ear wired earphones - Earpods style earphones or bluetooth earphones (such as Airpods)
Wireless or wired	Which of the following do you prefer?	<ul style="list-style-type: none"> - Wireless (bluetooth) over ear headphones - Wired over ear headphones
Noise cancellation	Which of the following do you prefer?	<ul style="list-style-type: none"> - Noise cancelling headphones/earphones - Non-noise cancelling headphones/earphones
Ideal budget	Imagine you had lost or broken your current headphones/earphones, and wanted to purchase a new pair. What's the approximate amount you'd be willing to pay?	<ul style="list-style-type: none"> - Numerical free text response
Maximum budget	Realistically, what is the maximum amount you'd be willing to pay for the right headphones/earphones?	<ul style="list-style-type: none"> - Numerical free text response
Ranking of attributes	Please rank the following product attributes in terms of how important they are to you when purchasing headphones/earphones.	<ul style="list-style-type: none"> - Product is within budget - Product is an over the ear headphone/in-ear wired earphone/earpods style earphone - Whether the product is wired/wireless (bluetooth) - Having noise cancellation technology

²⁵ Screening based on sampling criteria (headphone.earphone users and e-commerce users), we use attention checks inattentive participants to improve data quality

- Battery life
- The sound quality of the product
- The material quality of the cable/wire for wired earphones/headphones

Once participants left the simulation, they were asked about their sentiments towards the simulation, the algorithms and in arms 3-7 they were asked about their sentiments towards the additional features which were layered over.

3.4.2 Trial arms

Table 2: Trial arm descriptions and rationale

Trial arm	Rationale
1. Random [Control] ranking - Items ranked randomly	Products are shown in a completely random order. Whilst this may not represent how online sites typically present results, it should provide a helpful baseline to compare trial arms to.
2. Consumer-focused algorithm	Create what we consider an ideal, dynamic online environment which best represents consumers preferences. Faithfully reflective of individual preferences and ordered according to the “best” result first Note: we can’t factor in search criteria as participants won’t be typing in search terms.
3. Consumer-focused algorithm with transparency messaging	This arm uses the same underlying algorithm as the “Consumer-focused” arm, and the aim of this arm is to test how transparency of the ranking and recommendation mechanics impact consumer choice. In doing this we added a banner stating <i>‘We’ve ordered these results based on what you were looking for, based on the information you told us’</i> .
4. Commercially-focused algorithm	This arm uses the algorithm which emulates commercial maximisation approaches that online platform retailers could potentially employ. The products are presented considering consumer preferences but are not optimised for value for the consumer. This arm was designed to test the impact an algorithm purposely designed to increase commercial profit for an online platform may affect consumers’ choices. Designs are covered in detail in section 3.3.3, the broad mechanism behind this algorithm was to promote products that were more expensive than the top-scoring product, while still being broadly in line with a consumer’s preferences.

<p>5. Commercially-focused algorithm with sponsored messaging</p>	<p>This arm uses the commercially-focused algorithm, and includes the addition of a sponsored flag on certain products (always the same positions) and seeks to understand the additional impact of these messages commonly used in online platform retailers. The arm seeks to determine if the additional feature layered on a commercially designed algorithm significantly impacts consumer choice and whether it leads to increased or decreased spend.</p>
<p>6. Commercially-focused algorithm with platform recommendation messaging</p>	<p>This arm is a variation of the “Commercial focus” arm and uses the same algorithm. This arm includes the addition of an “I-Web’s choice” flag on certain products (positions one, two and five) and seeks to understand the additional impact of these messages commonly used in online platforms. The arm seeks to determine if the additional feature layered on a commercially designed algorithm significantly impacts consumer choice and whether it leads to increased or decreased spend.</p>
<p>7. Commercially-focused algorithm with sorting option</p>	<p>The arm includes a sorting drop down list with the options to sort low to high, high to low or display the default products order. This arms seeks to determine if the additional feature layered on a commercially designed algorithm significantly impacts consumer choice and whether it leads to increased or decreased spend.</p>
<p>8. Income-based algorithm</p>	<p>This arm uses household income data from participants to present products based on pricing groups. The aim of this arm is to test the impact of “price steering” on consumer choice and determine its impact on consumers’ spend. There were no budgetary restrictions and it did not take into account participants’ stated preferences.</p>

3.5 Ethical considerations

The online shopping simulation experiment raises several ethical considerations that must be addressed to ensure the safety and wellbeing of participants. The research team made sure to consider ethical concerns such as data privacy and participant welfare. Participants gave their consent to the panel provider for their non-personally identifiable data and responses to be used for research purposes and information about the study and participant rights were provided at the beginning of the survey. Personal identifiable data was not collected. The participants were not deemed vulnerable, and they were free to withdraw from the study at any time. Additionally, the research team conducted a sentiment analysis in the pilot to check for any distress to participants during the pilot trial, and ensured that the experiment was conducted in compliance with relevant regulations.

3.6 Data collection

Data collection ran from 22 February 2023 to 21 March 2023. 8,009 participants completed the experiment, and the sample was nationally representative in terms of age, ethnicity, gender and location.

3.7 Analysis:

In this section we have outlined the main analysis we will be conducting in the study. Full details of the model specifications are provided in appendix 5. It should be noted that while we had eight trial arms, in the main analysis we present results for the three main algorithms in comparison to the random ranking, and examine the additional features separately.

Primary Outcome - % of participants selecting their top-scoring product:

This outcome looked at the simple outcome of consumers selecting a product that is top-scoring for them based on their stated preferences and budget. Here we report the proportion of individuals selecting the top-scoring product for them in each trial arm, taking the arm with the most desirable outcome as the reference category. We define the top-scoring product as the product that matched the preferences we elicited from consumers most closely while remaining affordable based on the budget they stated. We also assess the mean position of the selected products in each arm vs the mean position of the top-scoring products and assess the proportion of consumers selecting products from the first page in each type of arm (see Appendix 10).

Additionally, we report the proximity score to a participants' top available product based on their elicited preferences. The proximity will be shown as a negative value with a score of zero meaning they chose their top-scoring product. We compare a random ranking vs consumer-focused vs commercially-focused vs income-based algorithms, with the arm achieving the most desirable outcome as the reference category. In addition we will investigate proximity in a more broader sense looking at absolute distances to acknowledge that there are products which are more similar and less similar which may indicate less or more harm objectively (Appendix 11).

Secondary outcome - Financial Impact:

Here we report the financial impact in terms of deviation from the price of a participant's top-scoring product, and will report the mean overspend and underspend (£) in each arm. The mean spend will be shown as a positive value if they underspend²⁶, and a negative value if they overspend²⁷. We will measure this outcome by trial arm and compare random vs consumer-focused vs commercially-focused vs income-based rankings, with the random ranking as the reference category. While seeing overspend as potentially disadvantageous for consumers, we investigate underspend as potential economic loss to the headphone

²⁶ Underspend is shown as a positive value since it means that participants are saving a particular amount of money compared to the top-scoring product.

²⁷ Overspend is shown as a negative value since participants are essentially losing money when spending more on a product than the top-scoring product costs.

market whereby money was not spent in a market where a product could satisfy the demand. We consider underspend as economic loss to the headphone market only if consumers felt that the top-scoring product was preferential to the product they selected.

Features effects on Primary and Secondary outcomes:

We applied and tested different features to specific arms. We compared the transparency messaging arm to the arm with a consumer-focused design with no features, including a text that clarified how the algorithm was designed; and we compared the sponsored product, platform recommendation and sorting features arms to the commercially-focused arm with no additional features. We placed the sponsored and platform recommendation messages in positions 1, 2 and 5 in the ranking.

Exploratory Outcomes - Sentiment Measures:

As additional outcomes we will report sentiment measures when participants were asked about the e-commerce platform in general, the algorithms specific to their treatment arm, and the special features specific to their treatment arm. These measures will be presented as percentages based on the 4 point likert scales used (negative sentiment will be coded in reverse order), and include measures of trust, anxiety, ease of use, frustration, ethics, and discomfort, among others. We will compare these measures across all eight arms separately.

Additional Analysis - Multiple exploratory outcomes:

Here we report on specific preferences, such as noise cancellation technology and wired/wireless preference. In addition, we conclude the additional analysis with a statistical segmentation to group participants by demographics and identify targetable groups for policy or regulation. We explore how these segments are impacted by the algorithms using our primary and secondary outcomes.

3.8 Limitations

While the results of this study provide important insights into the impact of algorithms on consumer behaviour in e-commerce environments, there are some limitations to consider.

Hypothetical Bias

One of the main limitations is the fact that the study used online experiments rather than real-world transactions. Participants were not spending real money in an online shopping simulation, which may have influenced their behaviour and decision-making processes. Additionally, the online shopping simulation is not a perfect substitute for real-life shopping experiences, as participants are not spending real money and may not approach the simulation with the same level of attention or scrutiny that they would use in a real shopping situation.

Self-reported preferences

Another experimental limitation was using self-reported product preferences to determine a participant's top-scoring product. Firstly, participants may not have a clear understanding of their own preferences, or may not have been able to accurately communicate them. Secondly, participants may be influenced by factors outside of their true preferences, such as the ranking, i.e. the order in which options are presented, or social desirability bias. Additionally, the study may not have included all possible product attributes that could influence a participant's top-scoring product, leading to an incomplete understanding of their preferences. One important attribute here was the colour of the products, and to mitigate this we randomised the pictures shown to each participant, and added three standard colours to choose from to control for the impact of product colour on participants' choices in the analysis.

It is also possible that a participant's preferences may have changed while undergoing the simulation as analysis of post-experiment questions suggests (see Appendix 6 - between 30% and 43% of people who did not choose the top-scoring product said they had changed their mind on their preferences). It is also worth noting that 43% of participants, who had not chosen the top-scoring product, said they did not want to change their selected product to the top-scoring product after the experiment (see Appendix 6). This suggests that for these participants the chosen product is a 'better' product for them (and therefore this cannot be considered a harmful over- or underspend), revealing preferences that are different from their initially stated preferences. Removing these participants from the calculation of the outcome measures does not alter the results. Finally, the study was conducted in an online simulation rather than a real-world shopping experience, which may limit the generalisability of the findings. Despite these limitations, self-reported preferences can still provide useful insights into consumer behaviour and can be a valuable tool in product development and marketing.

Generalisability

Additionally, the study focused on a specific population of people who owned earphones and used online platform retailers, which limits the generalisability of the findings to other populations or product categories. In addition we decided not to take into account the possible effects of social influence, such as the impact of peer recommendations or reviews, which could have a significant impact on consumer behaviour.

External Validity

Another limitation of our study is that the product database designs, platform design and algorithm designs were developed in-house rather than using actual algorithmic functions or actual product databases. While we made every effort to ensure that our simulations were externally representative, there is always the possibility that our designs did not fully capture the complexities of real-world product databases. These limitations should be taken into consideration when interpreting the results of our study.

Treatment design variation

While not necessarily a limitation, we choose to mention the fact that one of the arms in the study includes a significant design change. Participants who saw the income-based ranking were not presented with the budget we elicited nor did they have a spending limit. The

design of this algorithmic ranking was focused on using demographic information rather than elicited information to rank products. The demographic information, household income in this case, was only collected during the experiment and therefore could not be used to define a budget or spending limit without significant guess work. We felt any limit on these aspects could be at odds with the intended design of the arm and therefore chose to remove them.

Consumer products and behaviour

Finally the products used in the simulation were fictitious, and therefore we did not account for the impact of branding on consumer behaviour. To mitigate this limitation, we created 10 neutral brand names and used 40 different product images that were matched to the product type. These brand names and product images were randomly assigned to each participant to ensure that the impact of branding was minimised. Additionally, we added different colour options to negate any gendering effects from the colour of products. However, it is important to note that this study was not designed to test the impact of branding on consumer behaviour, and further research may be necessary in this area.

4. Findings

In this section, we report the key findings from our experimental research. We conduct all analysis on those who selected a product in the simulation ($n = 7,882^{28}$).

4.1 Primary analysis: impact of ranking algorithms on consumer product choice

4.1.1 Proportion of respondents selecting the product with the top score

When analysing the impact of the ranking algorithms on consumer choice we found significant effects for the consumer-focused and commercially-focused algorithms ($p < 0.1$ in both cases). When measuring our primary outcome of selecting the top-scoring product available, the participants in the arms containing the consumer-focused algorithm had a much higher success rate of selecting the top-ranked product, with approximately one fifth (19.32%) participants selecting the top product (figure 8 below). This success rate was less than 3% (2.57%) for the random arm and even lower in the income-based arm. Less than 10% (6.55%) participants who saw the commercially-focused rankings selected the product with the highest score, but this was still significantly higher than in the group who saw a random ranking of products. Interestingly, there was no statistically significant difference between the random arm and the income-based arm on this outcome. These findings are amplified when looking not only at the top-scoring product but also at the top 5%, top 12.5% and top 20% of products based on proximity-scores to the top-scoring product. When looking at the top 20% of products based on proximity scores to the top-scoring product, 44% of people selected those in the consumer-focused algorithm, 28% in the commercially-focused

²⁸ Not all participants selected a product and this is the subset of those who did.

algorithm and 7% in the random ranking and 5% in the income-based algorithm (see Appendix 11 for more detail).

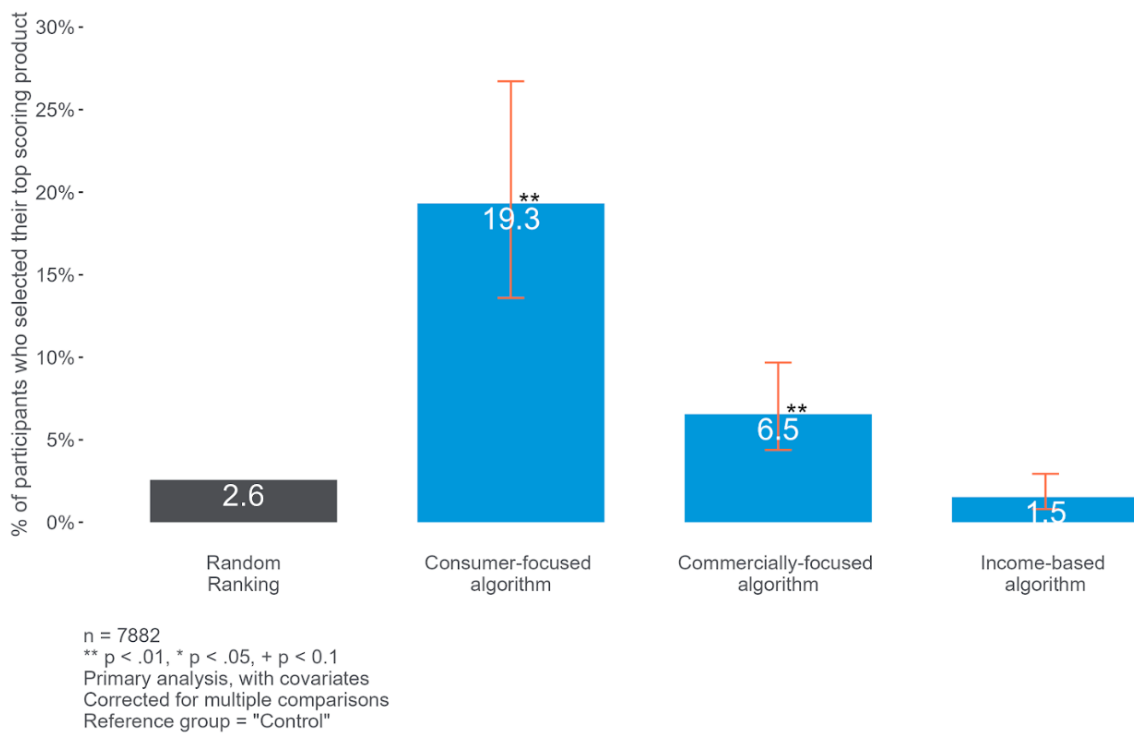


Figure 8: Treatment effect on consumer product choice by algorithm type

Overall, these findings demonstrate the significant impact of algorithm design on consumer behaviour and the importance of a consumer-focused approach when designing ranking algorithms. We show how small variations to the ranking design (whatever the motivation may be) can have large impacts, by presenting different product prices in the consumer-focused arm compared to the commercially-focused arm we see a 300% difference in our primary outcome of participants selecting the top-scoring product available to them between these two ranking designs. This finding indicates the importance of consumer consideration when designing ranking algorithms and shows how testing algorithms is imperative given the impact we show from small changes to the design.

4.1.2 Proximity to top-scoring product

We further examined how close the chosen product was to the top-rated product. This was achieved by determining a proximity score, which was calculated by subtracting the top-ranked product score from the selected product score. Negative scores are undesirable and a score of zero is ideal. The lowest score the consumer-focused algorithm can assign a product is 14.3²⁹. The random design resulted in an average proximity score of -3.4. We see similar significant results to the impact on consumer choice in the primary analysis. The consumer and commercially-focused algorithms achieved significantly higher proximity

²⁹ This is based on the linear design and the banking of attributes which are outlined in detail in sections 3.3.3 and 3.3.1 respectively.

scores to the random ranking design ($p < 0.01$) with average proximity scores of -1.1 and -0.8 respectively. Additionally, the consumer-focused arm does better than the commercially-focused arm in this outcome ($p < 0.01$). The income-based algorithm ranking leads to consumers choosing products that are further from the top-scoring product than any of the other rankings.

Appendix 10 outlines the median rankings of the top-scoring product in all algorithms, as well as the median rank of the product selected by respondents. The median position of the selected product was 10 for the random ranking, 5 for the consumer-focused ranking, 6 for the commercially-focused ranking and 10 for the income-based ranking. From this, it is clear that people did not just go for the first product they saw but took time to scroll down to a product of their choice.

Using the elicited preferences in our design we can ascertain how closely the product a consumer chooses are to those preferences and we successfully show ranking algorithms can have a significant impact on how likely consumers are to choose products that better align with these preferences.³⁰

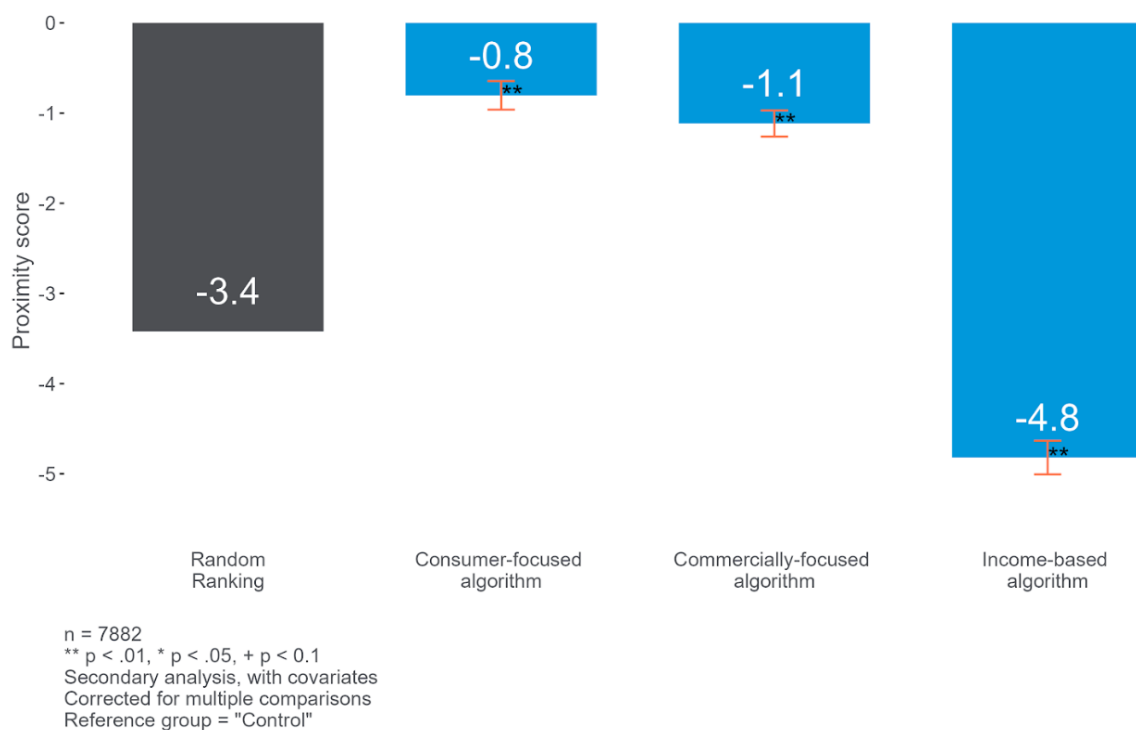


Figure 10: Treatment effects on proximity score by algorithm type

³⁰ We acknowledge and discuss the limitations of these elicited preferences in section 3.8 of this report.

4.2 Secondary Analysis: financial impact of ranking algorithms

We conducted an analysis of the financial implications of the various algorithmic designs. We consider a negative value (an overspend) as an outcome that is potentially disadvantageous for the consumer. This is a scenario where a consumer overspends on a product when a different product with the same attributes exists, presumably because they are unable to find a more suitable product due to the ranking. We also note two interpretations of underspend. First is that consumers found suitable products at a lower cost. A second interpretation of this underspend is consumers settling for inferior goods because they are unable to find a suitable product in the ranking they are presented with. While we don't consider this financially inefficient, this may still be harmful in a product satisfaction sense as well as economically inefficient in the headphone market if consumers would switch to more expensive goods with more suitable attributes which better align with their preferences.

We find significant treatment effects on this outcome for all of our algorithmic designs ($p < 0.01$). The random ranking leads to an average underspend of £15.60 for the consumer who saw this ranking and selected a product ($n = 933$). For those who saw the consumer-focused design, which attempted to present products in a way that consumers could meet their preferences at a price they could afford, the overspend per participant was £3.41 (a significant effect compared to the random arm $p < 0.01$). This overspend more than doubles in the commercially-focused arm at £7.96³¹. Finally the income-based ranking again more than doubles the overspend to consumers with an average overspend of £17.17³² for each participant who selected a product ($n = 911$). Using the difference in spend between the consumer-focused and commercially-focused rankings (£4.55) in this experiment and the total value of the UK headphone market of £490 million³³, we estimate the hypothetical annual overspend in the UK headphone market could be ~£46m.³⁴ Furthermore, we estimate the hypothetical annual overspend of the income-based algorithm could be ~£141m more compared to the consumer-focused algorithm based on this experiment.

As a robustness check after the simulation, we showed participants who did not select their top-scoring product their top-scoring product, and asked if they would switch products if they could go back. We then re-ran this analysis only with participants who said they would switch, and found no difference in terms of outcomes. Full details of this analysis can be found in Appendix 7.

³¹ Significantly larger effect than the consumer-focused arm $p < 0.01$

³² It should be noted that this arm did not have a budget limit by design, we discuss this limitation in detail in section 3.8

³³ Statista Headphone Category Revenue Data [\[link\]](#).

³⁴ Total overspend = ((commercially-focused overspend - consumer-focused overspend) * (Participants in consumer-focused arms)) / (Market value in consumer arms) * (Total UK market value)

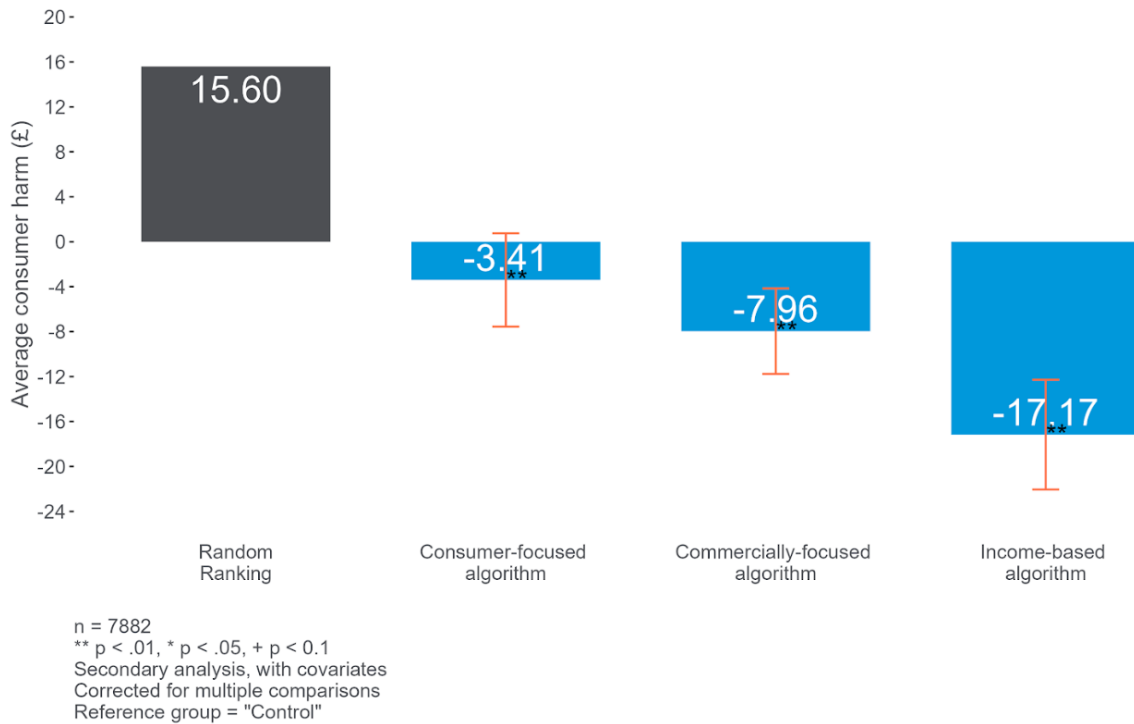


Figure 9: Treatment effect on average financial impact (£) by algorithm type

These findings demonstrate the significant financial impact of algorithm design on consumers. We show how a purely random ranking prevents online platforms from presenting consumers with products they would prefer to buy and may lead to consumers settling for inferior goods and cheaper prices. We estimated that this could lead to a hypothetical economic loss to the headphone market of approximately ~£159 million³⁵. This is supported by the fact that 66% of those who did not select the top-scoring product from a random ranking would switch to this product if they had the chance to reselect and 70% of those who underspent would spend more on the top-scoring product indicating they are willing to pay more to better satisfy their preferences. We also show that this misalignment of appropriate products with consumer demand can be reduced through algorithmic design. Only 48% of those who did not select the top-scoring product in the consumer-focused arms would switch to the top-scoring product (statistically lowest proportion of switching $P < 0.05$ of all arms). This suggests that even when consumers did not select the top-scoring product in a consumer-focused arm, this alternative was considered more suitable than when the same case occurs in the random ranking.

We demonstrate that minor modifications to algorithm design can exert a substantial effect on consumers. This is evidenced by the more than doubled overspend between our consumer-focused ranking and the commercially-focused ranking. The situation could worsen when ranking designs utilise personalised data like income. This financial impact may have important implications for the wider economy. Overspend in one sector may prevent spending in other sectors with unforeseen knock on effects. Moreover this overspend

³⁵ Total loss from random ranking = ((average underspend in random arm) * (Participants in consumer-focused arms))/(Market value in consumer arms) * (Total market value)

may have implications on consumers who take out credits and borrow money particularly with new “buy now pay later” credit models being introduced.³⁶ These findings demonstrate the impact algorithmic design could have on consumer choice and spend. While we have no data on the measures online platform retailers use to assess affordability it is clear that ranking algorithms may push consumers to overspend and this is something that should be considered in their design.

4.4 Feature Effects on Primary and Secondary Outcomes

Alongside algorithmic ranking, we assessed the impact of features like transparency messaging, sponsored products, platform recommendations, and sorting functionality on the outcome measures mentioned above. We hypothesised that sponsored products would deter product selection given their commercial motivation. We hypothesised the opposite for platform recommendations as this framing may elicit trust. We expected the transparency messaging to lead to significantly better outcomes.

- Sponsored products have a statistically significant financial impact on consumers, leading to overspend and also to higher proximity scores. This is because we expect people to click on and select sponsored products more often than on other products.
- We expected the opposite for platform recommendations.
- We expected the transparency messaging to lead to significantly more people selecting the correct product, significantly less overspend and a significantly higher proximity score.
- Finally, while we did not expect the sorting functionality to achieve better outcomes than preference-based algorithmic rankings (we used preferences in both the consumer and commercially-focused arms) we hypothesised that many participants would use the sorting functionality, which would neutralise the effect of the presented ranking. We explore this analysis in Section 4.6.1.

As explained above, we applied and tested different features to specific arms: we compared the transparency messaging arm to the arm with a consumer-focused design and no features, including a text that disclosed how the algorithm was designed; and we compared the sponsored product, platform recommendation and sorting features arms to the commercially-focused arm with no additional features. We placed the sponsored and platform recommendation messages in positions 1, 2 and 5 in the ranking. The findings are presented in Table 3 below.

Our overall conclusion is that the impact of algorithmic design on consumer choice is stronger than the additional features we tested. Surprisingly, we find few significant effects of the features on the outcome measures tested in this study. As shown in Table 4, our main finding is that participants were significantly less likely to select a ‘sponsored’ product than one with no feature or a platform recommended product. This outcome implies that platform recommendations do not enhance the likelihood of that product being chosen, hinting that these features exert minimal influence on consumer choice.

³⁶ Buy now pay later (BNPL) products are a form of short-term loan that are offered at the point of purchase, with either little or no fee/interest.

In the commercially-focused algorithm with no special features, 38% of participants selected a product in positions 1, 2, or 5 in the ranking, while 35% did so in the platform recommendation arm (not statistically significant), and 30% did so in the sponsored messaging arm (statistically significant $p < 0.05$). This indicates the recommendation is perceived as more appealing than the sponsorship messaging, but consumers do not select products based on it. Table 3 provides an overview of these results. The lack of effect in terms of outcomes again indicates the underlying algorithmic mechanism may be more important than the features layered on top of the rankings (assuming fixed positions of the feature as is the case in this experiment).

Maybe the most surprising finding around the features was the lack of impact of the sorting functionality, as only 13% of those who saw it used it. It is likely that consumers simply do not find the sorting function useful, but it is also possible that this behaviour was impacted by the fact we were dealing with a hypothetical scenario, and therefore participants did not want to spend the additional time using the sorting feature.

Table 3: Effects of features on outcomes

	Consumer-focused Algorithm		Commercially-focused Algorithm			
	No feature	Transparency messaging	No Feature	Sponsored products	Platform recommendations	Sorting functionality
Primary Outcome: Selecting top-scoring product	19%	20%	7%	7%	6%	7%
Secondary Outcome: Financial impact	-£3.63	-£2.62	-£7.79	-£7.71	-£7.50	-£8.95
Secondary Outcome: Proximity Score	0.78	0.80	1.04	1.10	1.10	1.19

Green and red shading indicate highest/lowest value in row and are significant at a 95% confidence level ($p < 0.05$)
 The comparison group for all statistical tests are conducted against the 'no feature' arms within the algorithm in question

Table 4: Proportion of respondents selecting position 1,2 or 5 by feature type

	Commercially-focused Algorithm		
	No feature	Sponsored products	Platform recommendations
% of participants selecting a product ranked 1st, 2nd, or 5th	38%	30%	35%

Green and red shading indicate highest/lowest value in row and are significant at a 95% confidence level ($p < 0.05$)
The comparison group for all statistical tests are conducted against the 'no feature' arms within the algorithm in question

4.5 Exploratory Analysis

The outcomes we have analysed to this point have focused on consumer choice and the potential impact on their financial well being and their satisfaction with their purchase decisions. In the following section we explore outcomes focused on how consumers' experience and sentiment may be impacted by ranking algorithms, the findings of which are summarised briefly here:

- **Participants spent less time when the algorithms took their preferences into consideration**
- **Sentiment toward the algorithm design was most positive for the consumer focused and commercially-focused algorithms (algorithms that included preferences in the ranking):** Platform sentiment was high for both of these algorithms, suggesting that participants were unable to discern the underlying algorithms. When we later explained the underlying algorithm, sentiment for the commercially-focused algorithm was more negative, suggesting that transparency surrounding algorithmic practices could help consumers make better purchasing decisions when we examine this finding in the context of our primary outcomes
- **Sentiment toward the sorting function was the most positive of all features:** This suggests participants appreciate the flexibility this provides more than transparent messaging around the algorithms.

4.5.1 Impacts of algorithmic ranking on search times

To explore consumer experience, we analysed the time spent browsing and selecting a product in the simulated e-commerce environment. The average time spent browsing and selecting in the simulation for the experiment was just short of 2 minutes (1 minute 57 seconds). Participants exposed to the random ranking or the consumer-focused algorithm spent 10 seconds longer on average browsing and selecting a product compared to those who encountered the commercially-focused ranking and income-based ranking. We do not draw strong conclusions from these results due to their size and descriptive nature. We believe that the similarity in time spent in the treatments with random rankings is not due to a similar consumer experience but rather two conflicting experiences that lead to the same time spent in the simulation. We think the longer time in the simulation for those who saw the

random ranking is due to difficulty finding relevant products and products that are affordable for them, this is essentially a negative and frustrating experience which leads to longer time in the simulation. We believe the converse is true for those in the consumer-focused arm who are presented with multiple products that are relevant and therefore the time spent in the simulation is spent comparing and selecting a top-scoring product which could be a more positive experience that leads to longer time spent in the simulation. We believe engagement may also explain the shorter time spent in the commercially-focused ranking which presents relevant products but not as affordable and therefore the comparison process may not be as engaging. Those who saw the income-based ranking would have seen an assortment of products which appeared random but were priced specifically based on their income without any consideration for their preferences and we believe they select the first product that matches their main preferences irrespective of the price which leads to a shorter browsing time.

4.5.2 Sentiment analysis

Website realism and experience

Once the participants had completed the simulation, we elicited their sentiments. We asked a set of questions regarding their sentiment to the website, their sentiments towards the algorithms that ranked the products they saw and for those who saw a message or additional feature overlaid on to the simulation we asked the sentiment around this feature. The findings are presented in Table 5 below.

The first set of sentiment questions were regarding the experience with the simulated online environment, we wanted to assess the realism and functionality of the simulation as well as the overall experience using it. The overall sentiment towards the website was most positive among those who saw the consumer and commercially-focused rankings, while both the random and income-based rankings elicited statistically more negative sentiments.

This statistical difference was not driven by realism of the site (at least 67% of people thought the simulation was realistic), the perceived transparency of the site (at least 76% of people believed the simulation was transparent) or the trustworthiness of the site (73% of people believed the simulation to be trustworthy). All of these sentiments were statistically the same across algorithm designs which presents evidence that the simulation itself is not responsible for the difference in outcomes across treatments.

There were however statistical differences in how easy respondents felt the site was to use, how frustrating the site was to use and the relevance of the ranking presented across algorithm designs ($p < 0.05$). For ease of use, the random ranking and income-based algorithm led to significantly more negative sentiment than the other two designs (75%-77% vs 80%-82%). We believe this is due to the experience of having to search harder for products that match the consumers' preferences. This is mirrored in sentiments of frustration, more people found the simulation to be frustrating in the two arms that do not consider consumer preferences, 22%-24% thought the simulation was frustrating in the random and income-based designs vs 20%-21% in the arms that factors in consumer preferences. This is also clear in the final sentiment question which asks about relevance of products shown, the

random and income-based rankings have the lowest sentiment scoring and the income-based arm is statistically the lowest scoring 70% vs 74%-80% ($p < 0.05$) in the other rankings.

Table 5: Sentiment scores for consumer experience of simulated e-commerce environment

	Random ranking	Consumer-Focused algorithm	Commercially-focused algorithm	Income-based algorithm
Overall Website sentiment (mean)	74%	77%	76%	74%
... is realistic	67%	69%	69%	69%
... is easy to use	75%	82%	80%	77%
... is <u>not</u> frustrating to use	76%	80%	79%	78%
... offered products that reflected preferences	74%	80%	77%	70%
... was transparent	77%	78%	77%	76%
... seemed trustworthy	73%	73%	73%	73%

Green and red shading indicate highest/lowest value in row and are significant at a 95% confidence level ($p < 0.05$)
 All statistical tests are conducted with the random algorithm as the comparison group

These findings indicate that the consumer experience is significantly impacted by ranking algorithm design. Consumers found rankings which considered their preferences to be inherently easier to use and felt their preferences were reflected in the rankings. An interesting and important finding is that despite the differences in rankings presented, consumer perceptions surrounding transparency and trust were stable which implies consumers are unable to identify ranking types and are therefore powerless to reap the benefits or avoid overspending or underspending that can occur in these mechanisms.

Algorithm sentiment

Next we explained the underlying mechanism for the ranking they were shown and elicited sentiment from the participants through a set of survey questions.³⁷ The findings are presented in Table 6 below.

³⁷ For more information on specific survey questions please send your query to predictiv@bi.team

Our findings indicate consumer perceptions and sentiments would be significantly impacted by the underlying mechanisms of ranking if they knew about them. Overall the consumer-focused algorithm had the most positive sentiment with regards to the design. In terms of fairness, ethics, trustworthiness and appropriateness the consumer-focused ranking had the significantly most positive sentiment by 10% compared to all arms ($p < 0.05$). Conversely the income-based arm saw the most negative sentiment overall.

About half of those who saw the consumer and commercially-focused rankings believed the ranking mechanisms were obvious to see (54%) this drops to 45% and 41% in the random and income-based designs. Despite the random and income-based arms having conceptually simple mechanisms of ranking they are less obvious to consumers. This coupled with the fact that these two designs lead to less preferable outcomes than the other algorithm types presents a key aspect of algorithm design for policy makers.

We asked consumers how anxious and uncomfortable these designs made them feel. A relatively low proportion of consumers found the designs to cause anxiety (~20%) but this was significantly higher in the arms which did not reflect any consumer preferences ($p < 0.05$). Interestingly when asked about discomfort caused by knowledge of the underlying design the random arm causes the same level of discomfort as the consumer-focused arm (20%-22%) while the commercially-focused and income-based ranking cause significantly more (24%-28%) for consumers ($p < 0.05$).

Table 6: Sentiment scores for algorithm design and ranking

	Random ranking	Consumer-Focused algorithm	Commercially-focused algorithm	Income-based algorithm
Overall Algorithm sentiment (mean)	69%	79%	72%	65%
... is fair	61%	73%	62%	57%
... is ethical	54%	65%	52%	50%
... is trustworthy	58%	69%	59%	57%
... is appropriate	63%	78%	66%	60%
... obvious from the rankings	45%	54%	54%	41%
... makes them feel anxious	21%	18%	19%	23%

... makes them uncomfortable	22%	20%	24%	28%
------------------------------	-----	-----	-----	-----

Green and red shading indicate highest/lowest value in row and are significant at a 95% confidence level ($p < 0.05$)
 All statistical tests are conducted with the random algorithm as the comparison group

Consumer-focused rankings are most well regarded while an income-based ranking on income is the least well regarded. It is also clear that despite random ranking being arguably the fairest, appropriate and ethical approach it is not considered as such by consumers which adds more weight to the argument that some form of algorithm is better than none. That being said this form of ranking does not cause discomfort to the extent commercially-focused or income focused designs do. These findings clearly show that preference base rankings with no commercial agenda are seen as the most ethical and appropriate way to present results.

Feature and functionality sentiment

The final element of sentiment we investigated was the sentiment towards transparency messaging, sponsored items, platform recommendations and sorting functionality. These features are outlined in detail in section 3.3.4. The findings are presented in Table 7 below.

Overall sentiment for the sorting functionality feature was the most positive of all the features at 84%, this is significantly higher than all other features including transparency messaging overlaid on the consumer-focused algorithm. This is driven by significantly more positive sentiments for all aspects investigated including fairness, ethics, trustworthiness, appropriateness, obviousness, causing anxiety and causing discomfort ($p < 0.05$). Despite this strongly positive sentiment to the sorting functionality only 13% of respondents in the arm with this functionality used it. Consumers seem to want the functionality even though they don't use it nor does it necessarily present the rankings in a more helpful or beneficial order. It should be noted that transparency messaging was more positive in most aspects of sentiment than sponsored product messaging and platform recommendations. Transparency messaging was not statistically significantly different from the sentiments of sorting functionality in terms of causing anxiety or discomfort while the other two messaging types were.

The most notable aspect to these findings is the sentiment around messaging. Messaging in both sponsored and platform recommendations generate statistically identical results. This is a surprising finding as we had initially believed these two types of messages would be perceived in different if not opposite ways. Equally notable is the relatively low sentiment for both messages around fairness, ethics, trustworthiness and appropriateness. Only around 50% of consumers had positive sentiment for these aspects for both messages indicating that any form of advice regardless of potential motivation is viewed negatively. This is an incredibly valuable finding for online platform retailers who may be unwittingly negatively impacting consumer experience through this type of feature. In addition these messages cause significantly more anxiety and discomfort than transparency messaging.

Table 7: Sentiment scores for features and sorting functionality

	Transparency messaging	Sponsored products	Platform recommendations	Sorting functionality
Overall feature sentiment (mean)	79%	66%	66%	84%
... is fair	74%	50%	51%	78%
... is ethical	65%	45%	45%	68%
... is trustworthy	68%	49%	50%	74%
... is appropriate	76%	54%	57%	82%
... obvious from the rankings	51%	54%	51%	63%
... makes them feel anxious	16%	21%	19%	17%
... makes them uncomfortable	17%	25%	23%	19%

Green and red shading indicate highest/lowest value in row and are significant at a 95% confidence level ($p < 0.05$)
All statistical tests are conducted with the highest value in each row as the reference

Through these findings, we demonstrate that consumers have positive sentiments towards transparency and control over the rankings they are shown. Conversely having sponsored tags and recommendations from online platforms is not in high demand and leads to relatively negative sentiment. These findings may be most relevant to online platform retailers as they reflect consumer experiences..

4.6 Additional findings

4.6.1 Sub group analysis

In this section we explore the impact of algorithmic rankings on different subgroups of the data set, and full tables of the results can be found in Appendix 12. Firstly we analyse the main outcome measures for those who use the sorting function (13% of the arm with this feature) vs those who do not, this analysis is carried out within the treatment arm which includes the sorting function as an additional feature. We then look at differences in outcome measures across demographics as well as the comparison between those with mental health challenges and those without and we compare those with self-reported digital confidence and

those without. We conduct this analysis within each algorithmic ranking design for comparability.³⁸

While the primary outcome of selecting the top-scoring product is not statistically different between those who use the sorting function and those who do not (6% and 7% respectively) we see statistical differences on both secondary outcomes. Those who do not use the sorting function experience an overspend of £6.63 which is in line with the results for the commercially focused algorithm. Those who use the sorting function are split into two groups 1) the group who sort low to high and experience a significant underspend of £5.04 ($p < 0.05$) and 2) those who sort high to low and experience a significant overspend of £58.29 ($p < 0.05$). In terms of proximity score those who do not sort (allowing the algorithm to include preferences in the ranking) have a significantly higher proximity score to those that do (-0.96 and -2.71 respectively $p < 0.05$). Interestingly the proximity score for those who sort low to high is very similar to those who sort high to low (-2.60 and -2.85 respectively). In summary, based on this limited sample of 13% of participants of this trial arm (133 people), sorting on price as a feature does not seem to impact the proportion of consumers who select the top-scoring product but it does affect the proximity score and is more likely to lead to underspend or overspend.

For the primary outcome of selecting the top-scoring product we see a substantially better outcome among those aged under 25 in the consumer-focused ranking, 26% of these consumers selected their top-scoring product compared to only 18% and 19% for those aged 25 to 54 and those aged 55 and over respectively. In terms of financial impact, consumers aged under 55 overspend less in both the consumer-focused ranking and the commercially-focused ranking than those aged over 55. This indicates that older consumers may be more impacted by algorithmic rankings that have the potential to cause overspend.

Gender only seems to play some kind of role in determining the level of financial impact. Males who saw the commercially-focused and income-based rankings experienced substantially more overspend than females who saw these same rankings.

Ethnicity appears to have very little predictive power on the outcome measures we tested across algorithm designs. The only exception is proximity scores further away from the top-scoring product among non-white consumers but this is not statistically significant.

We asked participants if they had experienced mental health challenges in the past 12 months. Those who reported mental health challenges experienced less overspend and achieved better proximity scores. We acknowledge this is based on a single self-reported measure but this finding does indicate there is a potential difference in outcomes measures amongst those who have and those who believe they have not experienced mental health challenges in the past 12 months and this could be driven by a number of reasons. This could be a key area for deeper research.

³⁸ We **bold** values with what appear to be substantial differences within a demographic but these have not been statistically compared. We have chosen not to compare statistically to decrease the likelihood of spurious results coming from multiple comparisons.

We also assessed the outcomes among those who self reported high digital confidence and those who reported relatively low digital confidence.³⁹ We found that those with higher online confidence experienced substantially less favourable outcomes than those who self reported as lower confidence. This could indicate that those who are self reporting as confident are actually over confident and less thoughtful when shopping online. This could also be an interesting avenue for further research.

Finally we assessed the impact of the primary and secondary outcomes on those who reported being on universal credit vs those who did not. There was little notable difference between these groups in terms of the primary outcome of selecting their top-scoring product or the proximity score. This group does differ drastically in terms of the financial impact when seeing the income-based ranking. Those who report to be on universal credit actually slightly underspent (+£1.42) when seeing this ranking compared to the considerable overspend (-£21.45) of those who are not on universal credit.

4.6.2 Segmentation analysis

We identified three clear segments in our data using K-medoids statistical clustering. Full details of the method used can be found in Appendix 6. For the segmentation we only included those who selected a product (n = 7,882). Cluster “A” consists of 3,858 consumers making up 49% of our sample, Cluster “B” consists of 1,251 consumers making up 16% and Cluster “C” consists of 2,773 consumers, making up 35% of the sample. From the differences in demographics and outcomes we have termed these clusters A - Over confident online consumers, B - Savvy online consumers and C - Vulnerable online consumers. Table 8 presents key demographic characteristics and outcomes for these clusters.

Cluster A - Overconfident online consumers

Cluster A represents consumers who exhibit overconfidence in their digital abilities and are frequent online shoppers. This group demonstrates the least amount of anxiety when informed about algorithm design and are more comfortable with the ranking algorithms than other clusters. Cluster A has the highest level of trust in the algorithm they were presented with, perceiving it as ethical and appropriate compared to other groups. This cluster is older and more female than the others, and shop online most frequently. This group has the lowest proportion of degree-educated individuals. This group selects the top-scoring product a little more often than the savvy online consumer group but experiences lower proximity scores and more overspend.

Cluster B - Savvy online consumers

Cluster B represents consumers with moderate confidence in their digital abilities, between the overconfident and vulnerable groups. This group shops online less frequently than Cluster A but more than Cluster C. Cluster B self reports mental health challenges more frequently than the other groups and experiences the most anxiety and discomfort upon

³⁹ We used a set of 7 questions relating to risk, data usage, advertising, misinformation, online abuse and content sharing to develop a score, those above the median score are considered to be higher confidence and those who scored below the median are considered lower confidence.

learning they were shown a ranking based on an algorithm. They're less trusting of the algorithm and less likely to consider it ethical than the overconfident group. However, a similar proportion of savvy consumers consider the algorithms to be apparent as the overconfident consumers. While experiencing the least overspend of the three clusters, the savvy group chooses the top-scoring product almost as often as the overconfident group. Cluster B consumers are more likely to be non-white, live in densely populated areas, and are younger than the other two clusters.

Cluster C - Vulnerable online consumers

This group of consumers shops online least frequently of the three clusters. They are also least confident in their digital abilities. This is the group where the lowest proportion of the consumers believe the algorithm is appropriate, trustworthy or ethical. It is also the group with the lowest proportion of consumers who believe the algorithm was obvious to see. This group selected the top-scoring product least often but had proximity scores higher than the median score across the experiment. This group are the oldest, the highest proportion of females and have the highest proportion of individuals living in rural areas. They are the group who experienced the most overspend.

Table 8: Key demographics by cluster

	Over confident online consumers	Savvy online consumers	Vulnerable online consumers
Proportion of cluster...			
... aged under 55	64%	81%	62%
... that are female	51%	41%	52%
... with a household income above £40k	48%	46%	49%
... are from Asian, Black and other ethnically diverse backgrounds	15%	23%	13%
... living in rural areas	21%	15%	23%
... that are degree educated	31%	36%	37%
... with higher digital confidence than the median	59%	47%	35%

... who report having experienced mental health issues	23%	30%	23%
... are on universal credit	19%	27%	15%
... who shop online rarely	8%	10%	13%

No statistical comparisons conducted

Table 9: Key outcome measures by cluster

	Over confident online consumers	Savvy online consumers	Vulnerable online consumers
Proportion of cluster...			
... selecting the top-scoring product	10%	9%	7%
... experiencing more overspend than the median	44%	40%	62%
... with higher proximity scores than the median	43%	40%	65%
... feel the algorithm makes them anxious	3%	83%	13%
... feel the algorithm makes them uncomfortable	4%	88%	20%
... feel the algorithm is appropriate	94%	75%	29%
... feel the algorithm is ethical	81%	70%	13%
... feel the algorithm is fair	93%	74%	21%
... feel the algorithm is obvious from the rankings	70%	71%	18%
... feel the algorithm is trustworthy	8%	10%	13%

No statistical comparisons conducted

5. Conclusion

The study examined the impact of different algorithms on consumer choice in an e-commerce environment, focusing on a hypothetical headphone market. Through a simulated e-commerce platform, three algorithms were compared to a random ranking: a consumer-focused algorithm, a commercially-focused algorithm, and an income-based algorithm. The findings provide valuable insights and have important implications for algorithm design, consumer choice, and market efficiency.

This report provides evidence that the design of algorithms has a significant influence on consumer choice, highlighting the importance of considering consumer preferences in algorithmic rankings.

First, we analysed how different algorithmic designs impact consumers' ability to find products that are aligned with their preferences. The consumer-focused algorithm outperformed the other algorithms with respect to participants selecting the top-scoring product, better supporting consumers to find the product in line with their declared preferences. In this analysis, the commercially-focused algorithm also performed better than both the random ranking and the income-based algorithms, which did not include consumers' preferences in their design.

The commercially-focused algorithm outperformed the random ranking in terms of the primary outcome, suggesting that commercial considerations play a role in product recommendations, thus affecting consumers and market efficiency. However, it fell short in supporting consumers to find a product in line with their preferences, when compared to the consumer-focused algorithm with respect to the main outcomes measures, reinforcing the need for a balanced approach that prioritises both consumer well-being and commercial performance. Furthermore, the study demonstrated that a random ranking may be as ineffective as an algorithm that disregards consumer preference.

The report also looked into the presentation of suitable products to consumers through proximity scores. Both the consumer-focused and commercially-focused algorithms outperformed the random ranking, indicating their effectiveness in offering relevant product choices. The consumer-focused algorithm exhibited superior performance compared to the commercially-focused algorithm in this aspect.

Secondly, the study delineates how small variations in algorithm designs had a significant financial impact for consumers, with potential effects on digital markets.

The income-based algorithm led to the highest overspend for digital consumers, followed by the commercially-focused algorithm and the consumer-focused algorithm. We estimate the hypothetical annual overspend of the commercially-focused ranking in the UK headphone market could be ~£46m compared to the consumer-focused algorithm based on this experiment. Furthermore, we estimate the hypothetical annual overspend of the income-based algorithm could be ~£141m compared to the consumer-focused algorithm. This highlights the need for careful algorithmic design to protect consumers' interests.

Additionally, the study revealed a potential economic impact of algorithms that do not consider consumer preferences when applied to rankings and recommendations. Random rankings led to a large underspend for consumers. This may not necessarily be financially

harmful for individuals, but it can result in broader economic implications and lower individual satisfaction if consumers settle for inferior goods due to difficulty in finding products that align with their preferences. Over 70% of those who under-spent within the random ranking indicated their willingness to switch to the more expensive top-scoring product they were presented with, representing a hypothetical economic loss to the headphone market of approximately ~£159m. These findings highlight the economic implications and emphasise the importance of algorithmic designs that take into account consumer preferences to avoid underspending on inferior goods by consumers.

The results emphasise the importance of algorithmic designs that align with consumer preferences and promote economic efficiency. By considering these findings and optimising algorithm designs, policymakers, regulators, and online platform retailers can enhance consumer well-being, foster trust, and create a fair and efficient online marketplace. While the study had limitations, such as the use of hypothetical products and the absence of real monetary transactions, it provides valuable insights into the impact of algorithm designs on consumer choice in e-commerce environments, highlighting the significance of considering consumer preferences and optimising algorithmic designs to maximise consumer satisfaction and market efficiency.

This is particularly relevant for vulnerable digital consumers, such as the elderly, those who shop online less frequently, live in the rural areas and have low confidence in their digital skills. Our clustering analysis showed that they are more susceptible to the impacts of algorithms in the online shopping space, experiencing the most financial harm in the form of overspending.

Moving forward, testing algorithm designs in real-world settings and collaborating with policymakers, regulators, and online platform retailers are crucial steps. Such efforts will help quantify the financial impact caused by current algorithm designs, refine policies, and improve the overall market for consumers and the economy.

In conclusion, the study highlights the importance of consumer-focused algorithmic rankings in enhancing consumer well-being and maximising economic efficiency. It emphasises the need for a balanced approach that considers both consumer preferences and commercial considerations. By striving for algorithmic transparency and optimising designs, policymakers, regulators, and online platform retailers can foster trust, increase consumer vigilance, and create a fair and efficient online marketplace. Further research in the field of algorithm design, as well as exploration of other aspects of e-commerce, presents promising opportunities for future investigation and collaboration among stakeholders.

Appendices

Appendix 1: Product Database Development

A pivotal aspect of our experiment design was the creation of a product database for use in the simulated e-commerce environment. We had to carefully choose the type of products to include, their attributes and attribute levels, the corresponding product images, and the product names, all while ensuring that the pricing model for the products was reflective of real-world pricing.

The selection of products involved several stages and a series of rationalisation steps. Initially, we considered phones, clothes, and white goods (e.g., dishwashers, washing machines) and evaluated their suitability for implementation in a simulated environment, the types of product attributes, consumer purchase frequency, and other critical factors. Ultimately, we concluded that none of these categories were appropriate and shifted our focus to additional electronic categories, specifically earphones/headphones, chargers, and kindles/tablets. The table below outlines the advantages and drawbacks of each product category.

Table 10: Product Category Benefits and Limitations

Category	Benefits	Limitations
Phones	<ul style="list-style-type: none"> – Preferences are limited and easy to elicit (design and analysis simplification) – Almost everyone has a smartphone (high IR in survey) – Fewer attribute levels (design and validity simplification) – Most attributes are technical rather than subjective (better camera = better phone) – Limited price range 	<ul style="list-style-type: none"> – Relatively infrequent purchase – Brand loyalty is a big preference – There are unobservables which may be linked to brand or preferences which we can't measure (implementation very difficult)
Clothing	<ul style="list-style-type: none"> – Common market that everyone uses or understands – Frequent purchase 	<ul style="list-style-type: none"> – Wide variety of products, styles, needs (design and analysis complications) – Attributes and preferences are so varied (algorithm complications) – Consumer issues hard to identify – Gendered products (design complications) – Price ladders are long and complicated – Harder to buy online – Has become commoditised (probably filter low to high)
Whitegoods	<ul style="list-style-type: none"> – Preferences are limited and easy to elicit (design and analysis simplification) – Fewer attribute levels (design and validity simplification) – Most attributes are technical rather than subjective 	<ul style="list-style-type: none"> – Low prevalence – Infrequent purchase – Retained for years – Huge variety – Renter vs owners

Earphones/ Headphones	<ul style="list-style-type: none"> - Preferences are limited and easy to elicit (design and analysis simplification) - Almost everyone has ear-phones (high IR in survey) - Fewer attribute levels (design and validity simplification) - Most attributes are technical rather than subjective (noise cancellation = better earphones) - Limited price range - Prices are low enough that consumers may not be as deliberate in purchasing as they would be with a phone 	<ul style="list-style-type: none"> - Brand loyalty may be somewhat of a preference and we are not testing this in the experiment
Chargers	<ul style="list-style-type: none"> - Preferences are limited and easy to elicit (design and analysis simplification) - Almost everyone uses phone chargers (high IR in survey) - Limited price range - Prices are low enough that consumers may not be as deliberate in purchasing as they would be with a phone 	<ul style="list-style-type: none"> - Very limited product variation (most consumers will pick the cheapest)
Kindles/ Tablets	<ul style="list-style-type: none"> - Preferences are limited and easy to elicit (design and analysis simplification) - Fewer attribute levels (design and validity simplification) - Most attributes are technical rather than subjective (noise cancellation = better earphones) - Limited price range 	<ul style="list-style-type: none"> - Limited consumer base (not everyone uses tablets) - Suffers from a similar problem to a phone in that it is an expensive, infrequent purchase - Brand loyalty may be a big preference (algorithm complications)

After considering various product categories, DSIT and BIT collaborated to choose the most appropriate product for our experiment. The research team decided to use earphones/headphones to test the impact of algorithms on consumer behaviour. The other product categories were not selected for specific reasons outlined below.

Phones/white goods were not considered suitable as they are purchased infrequently, and their high cost means that consumers make more deliberate purchasing decisions. The various purchasing options available for phones and white goods would require us to narrow our screening criteria or create a less relatable experiment setup.

Clothing is purchased frequently, but the large number of preferences and permutations makes it challenging to measure outcomes and determine preferences.

Chargers have limited product variation, making it difficult to differentiate between products, especially when not factoring in the brand.

Kindles/tablets suffer from similar pitfalls as phones, and their usage is not as widespread.

Headphones/earphones were selected as the category used in this experiment because

- Preferences are limited and easy to elicit (design and analysis simplification)
 - Almost everyone has ear-phones (high IR in survey)
- Fewer attribute levels (design and validity simplification)

- Most attributes are technical rather than subjective (noise cancellation = better earphones)
- Limited price range allowing us to keep the experiment design more simple
- Prices are low enough that consumers may not be as deliberate in purchasing as they would be with a phone meaning algorithms may be more impactful

Once the product category was selected we needed to decide which attributes we would use for the products, these would be used in the product design, the pricing and play a key part in the algorithm functionality. We selected a set of key attributes which allowed us to elicit preferences from consumers and create a wide database of products but not too many attributes that the experiment would become unwieldy. We used six product attributes 1) the earphone/headphone type (Over ear, in ear wired, ear bud), 2) noise cancellation functionality (Yes, No), 3) sound quality (Good, High, Ultra high), 4) connectivity (Wired, Bluetooth), for bluetooth products 5) battery life (Less than 10 hours, 10-20 hours, More than 20 hours) and for wired products 6) wire material quality (Good, High, Ultra high). Using these attributes we generated 216 products including all permutations of attribute combinations⁴⁰. There are 72 unique combinations of attributes within the 216 products. This is due to the fact that the levels of battery life are null in wired products and the levels of material quality are null in bluetooth products⁴¹.

All attributes except the headphone/earphone type are ranked from low to high (lowest value score 1, middle value scores 2, highest value scores 3) to emulate real world products with higher value features and functionality. We summed these scores for attributes to determine a product score. This defined 216 products normally distributed on product score (figure 11).

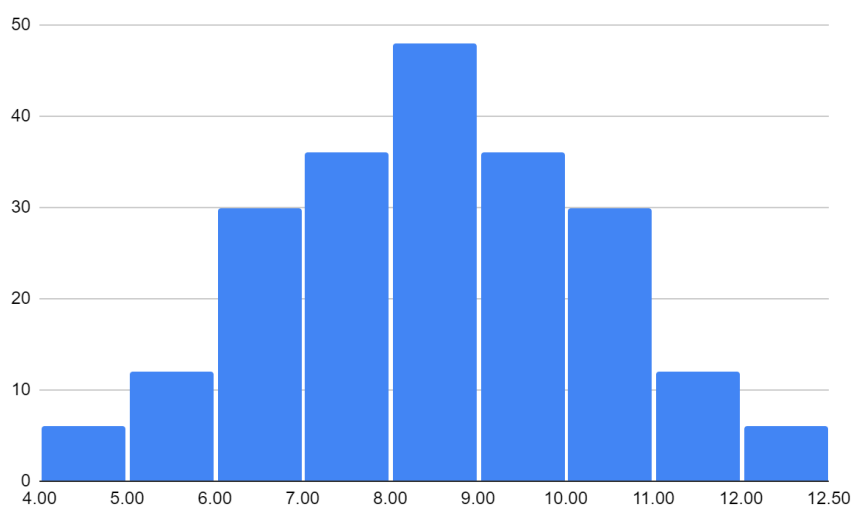


Figure 11: Histogram of Product attribute scores unadjusted

We developed a pricing range based on product pricing we observed in large online platforms and developed pricing brackets based on this range. We then assigned a price from these brackets based on the product score. Pricing brackets were a range and pricing

⁴⁰ This means in the realm of these attributes any combination of consumer preference can be met

⁴¹ Over ear headphones can be bluetooth or wired

was assigned at random within this range. Table 11 below shows how price brackets were associated with product scores.

Table 11: Product scores and associated pricing brackets

Score	Pricing range
<5	2.99-11.99
=5	12-24.99
=6	25-37.99
=7	38-50.99
=8	51-53.99
=9	54-100
=10	100-149.99
=11 or 12	150-399.99

This resulted in a set of 216 products with 72 unique attribute combinations with pricing that emulates the real world (figure 12).

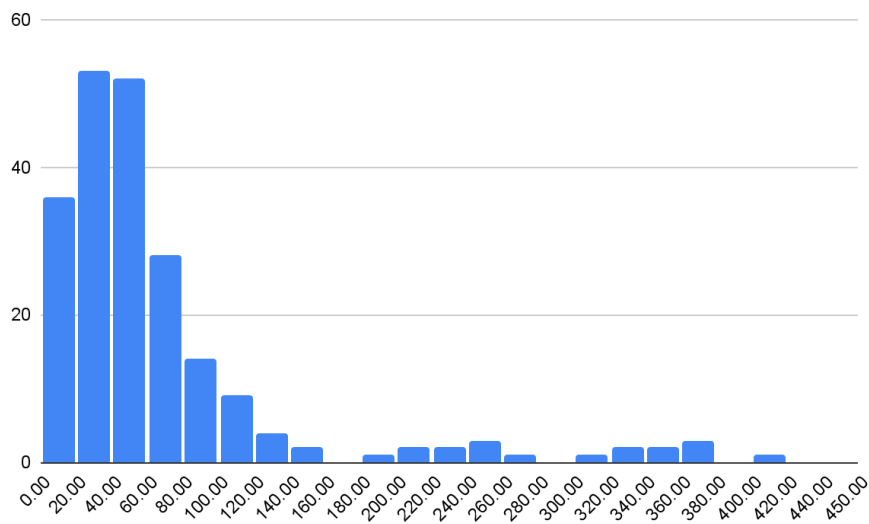


Figure 12: Histogram of product pricing in GBP

Finally we added a duplicate set of products to the database and all duplicated products included a 20% price premium. This was to represent pricing of branded products in a traditional market, products with the same attributes but higher pricing due to “brand premium”. We adjusted all prices to end with a £0.09 value to match what we observe online. Therefore, finally the product database consists of **432 products** with a price range of £3.49 to £480.69, Table 12 below outlines the composition by price of the product database.

Table 12: Database composition by price increment

	Cumulative	Line	Count in range
Under £10	6%	6%	24
Under £20	15%	10%	42
Under £30	26%	11%	47
Under £40	36%	10%	44
Under £50	51%	15%	65
Under £60	62%	11%	47
Under £70	69%	7%	30
Under £100	82%	13%	54
Under £130	88%	6%	25
Under £260	94%	7%	29
Under £490	100%	6%	25

Appendix 2: Platform Design

In this study, we developed an e-commerce platform that replicated traditional e-commerce platforms, which we named 'iWeb'. The aim of our e-commerce simulation was to create a platform that was as realistic as possible without jeopardising the integrity of the experiment. To achieve this, the site was designed to include advertisements along the side and a top banner to enhance its realism. Additionally, we created a logo for the site and added a stationary search bar at the top for ease of navigation. 15 products were shown per page, which was generally in line with what we saw online. We presented not only the product image and name but also the price of the product, as well as several other attributes that were elicited during the pre-experiment questions. These attributes included whether the product was noise-cancelling, the sound quality, the material quality (for wired products), the battery life (for wireless products), the headphone type (in-ear/over-ear), and whether the product was wireless. Note, the platform was compatible with both mobile and desktop devices, the only difference being that the mobile version did not include the advertisements on the side.

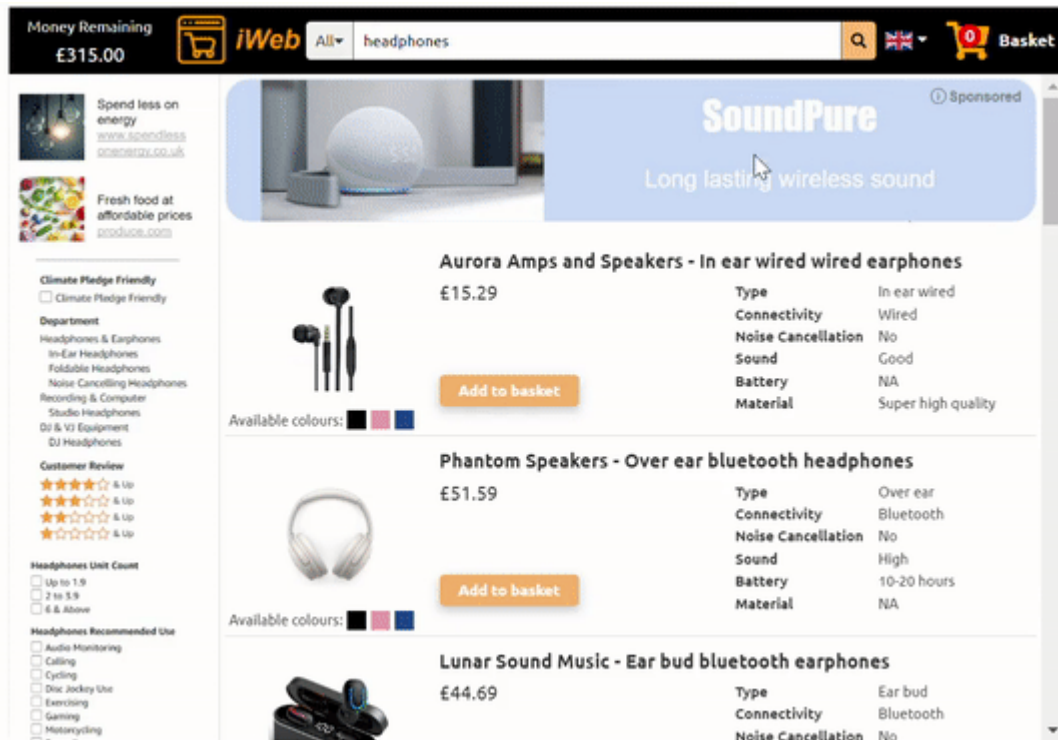


Figure 13: Overview of the e-commerce simulation

To give participants a sense of their budget, we displayed the total amount of money they had to spend in the top left corner of the screen. This amount would update in real-time as participants added products to their basket. The only exception to this was in the income-based arm of the study, where we removed the budget element as we based the algorithm around household income rather than elicited preferences.

In order to populate the platform with products, we sourced 10 brandless images for each of the following four categories: wired over-ear headphones, wireless over-ear headphones, wired in-ear earphones, and wireless in-ear earphones. These images were randomly assigned to products on each page load to ensure that the picture did not impact the results. We also included a note indicating there were multiple colours available (black, blue, pink) to attempt to mitigate the impact of the image colours. Similarly, we generated 40 random names for the headphones, which were also randomly assigned to products.

Participants were asked to select the product that was most in-line with their preferences. To do so, they were required to click an 'add to basket' button for their desired product. Participants were restricted to selecting only one product, and if they wished to change their selection, they were required to remove the first item from their basket. Once they were satisfied with their choice, they were asked to click the basket and proceed to the 'buy now' option.

If a participant tried to buy a product that exceeded their available budget, an error message would appear, informing them that they did not have sufficient funds to purchase that product. This feature was included to prevent participants from selecting products that they could not afford. Budget is calculated by eliciting the maximum a participant would be willing to pay for headphones/earphones and inflating this by 5% to allow for an overspend.

Overall, our e-commerce platform aimed to provide a realistic and comprehensive shopping experience for participants, while also ensuring the integrity of the experiment was maintained at all times. By including a range of product attributes and budget constraints, we were able to gather valuable data on participant preferences and purchasing behaviour.

Appendix 3: Algorithm Design

In this section, we will discuss the three algorithms that were designed for our experiment as well as the random control ranking. These algorithms were implemented in the simulated e-commerce environment and were used to manipulate the display order of products. The three algorithms are as follows: a consumer-focused algorithm, a commercially-focused algorithm, and an income-based algorithm compared to a random ranking. The consumer-focused algorithm and the commercially-focused algorithm were inspired by linear pricing models with a value normalisation term. The income-based algorithm, on the other hand, was an attempt to emulate what we believe online platforms may be able to do with a proxy for personal income. We will describe each algorithm in detail and explain how they were used in the experiment.

Random ranking: This was used as a control arm, and showed the 432 products completely randomly to participants. We did this as we wanted to be able to compare all of the algorithms to an impractical, but still completely random ranking.

Consumer-focused algorithm: The consumer-focused algorithm was designed to show products in an order that aimed to strictly reflect their stated preferences. The product with the highest score was shown first, the next best second, and so on. This algorithm was specified as follows:

$$\begin{aligned}
 Y_{pi} = & \\
 & \alpha_1 type_{pi} + \\
 & \alpha_2 inBudget_{pi} + \\
 & \alpha_3 noise_{pi} + \\
 & \alpha_4 sound_{pi} + \\
 & \alpha_5 connectivity_{pi} + \\
 & \alpha_6 material_{pi} + \\
 & \alpha_7 battery_{pi} + budgetDelta_{pi}
 \end{aligned}$$

Where:

- Y_{pi} is the resulting score assigned to each product in the algorithm.
- $\alpha_1 - \alpha_7$ are the model coefficients for the variables, and are linked to participant i 's relative ranking of the different product attributes. Participant i 's most important attribute was assigned a coefficient of 1.7, the second most important 1.6, and so on, with the least important attribute receiving a value of 1.1.

- $\alpha1type_{pi}$ is a variable given a value of 1 or 3, indicating whether product p matches participant i 's stated preference for either earpods, in ear wired headphones, or over-the-ear headphones.
- $\alpha2inBudget_{pi}$ is a value of 1 or 3, indicating whether product p is within participant i 's budget
- $\alpha3noise_{pi}$ is a variable given a value of 1 or 3, indicating whether product p matches participant i 's stated preference for noise cancellation technology
- $\beta4sound_{pi}$ is a value of 1,2 or 3 indicating the sound quality of product p .
- $\alpha5connectivity_{pi}$ is a variable given a value of 1 or 3, indicating whether product p matches participant i 's stated preference for either earpods, in ear wired headphones, or over-the-ear headphones.
- $\alpha6material_{pi}$ is a variable given a value of 1,2 or 3 indicating the material quality of product p , only relevant for wired headphones. A value of zero was assigned if participant i preferred wireless products.
- $\alpha7battery_{pi}$ is a variable given a value of 1,2 or 3 indicating the battery life of product p , only relevant for wired headphones. A value of zero was assigned if participant i preferred wired products.

The $budgetDelta_{pi}$ term was a term used to ensure participants were not shown too many products they could not afford to buy. If the participant was able to purchase product p , Meaning the price of the product was less than or equal to the total budget participants were given to spend, $budgetDelta_{pi}$ was normalised to a value between 1 and 3 using the following formula:

$$(productPrice_p - totalBudget_i)/(totalBudget_i) * (3 - 1) + 1$$

If the product was too expensive for participant i to purchase, the $budgetDelta_{pi}$ term was specified as:

$$(productPrice_p - totalBudget_i)/(totalBudget_i)$$

This was varied to ensure that products that couldn't be purchased were more strongly pushed to the bottom of the search results, ensuring participants weren't only shown products that were too expensive for them.

Commercially-focused algorithm: The commercially-focused algorithm was designed in a similar fashion to the consumer-focused algorithm, but had two additional mechanisms added, both with the aim of favouring more expensive products, and is specified as follows:

If product i was above participant i 's total available budget:

$$\begin{aligned}
Y_{pi} = & \alpha 1 type_{pi} + \\
& \alpha 2 inBudget_{pi} + \\
& \alpha 3 noise_{pi} + \\
& \alpha 4 sound_{pi} + \\
& \alpha 5 connectivity_{pi} + \\
& \alpha 6 material_{pi} + \\
& \alpha 7 battery_{pi} + identifierSet_{pi} + budgetDelta_{pi}/2
\end{aligned}$$

Or, if product i was within participant i 's total available budget:

$$\begin{aligned}
Y_{pi} = & \\
& \alpha 1 type_{pi} + \\
& \alpha 2 inBudget_{pi} + \\
& \alpha 3 noise_{pi} + \\
& \alpha 4 sound_{pi} + \\
& \alpha 5 connectivity_{pi} + \\
& \alpha 6 material_{pi} + \\
& \alpha 7 battery_{pi} + identifierSet_{pi} + budgetDelta_{pi}/4
\end{aligned}$$

Where:

- $identifierSet_{pi}$ was a variable given a value of 1 or 3. As stated in the product development section, a second set of 216 products was created as a duplicate of the first 216, but with a 20% price premium. A value of 3 indicates that product p is part of the second set of products, meaning these are favoured in the algorithm.
- $budgetDelta_{pi}$ is the same as in the consumer-focused algorithm, but is divided by 2 or 4, depending on whether product p was in participant i 's total available budget. Similar to $identifierSet_{pi}$, this division helps promote more expensive products in the algorithm, while not being so overpowering that it only shows products a participant cannot purchase.

Income-based algorithm: This algorithm aimed to use household income to personalise the products participants were shown, while maximising commercial profit.

The algorithm categorises the 432 products into 4 distinct categories based on their price points. Participants' Household Income (HHI) is then utilised to divide them into low, medium, and high income groups. Those participants in the low-income group are presented with product category 2 first, meaning the first 108 products shown are in this group, while participants in the medium-income group are shown category 3 first, and those in the high-income group are shown category 4 first, which were the most expensive products.

Notably, the algorithm did not incorporate any budgetary restrictions, meaning that participants were not limited to a specific budget based on their self-reported maximum budgets.

Appendix 4: Additional feature designs

In addition to the algorithms developed, we implemented four additional features to the e-commerce platform:

- **Transparency Messaging:** Trial arm 3 included a transparency message on the consumer-focused algorithm, informing participants that the algorithm was ordered based on their preferences. This message appeared at the top of the page in larger text to ensure clear visibility.

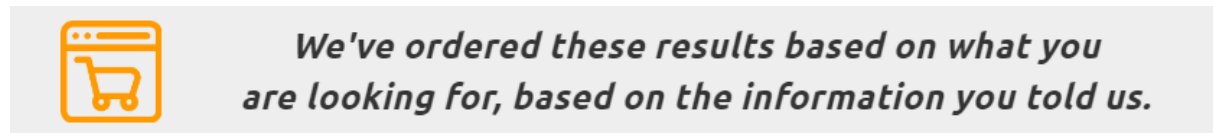


Figure 14: Transparency message shown at the top of the simulation

- **Sponsored Messaging:** In trial arm 5, a 'sponsored' message was added on top of the commercially-focused algorithm, replicating the type of 'sponsored' messaging observed on real e-commerce platforms. This message was added to products listed at position 1, 2, and 5 of the first page. We chose positions 1,2 and 5 as this reflected similar positions to the online platforms we looked at initially. Companies add this type of messaging to promote their products, indicate a commercial partnership, and increase their sales.

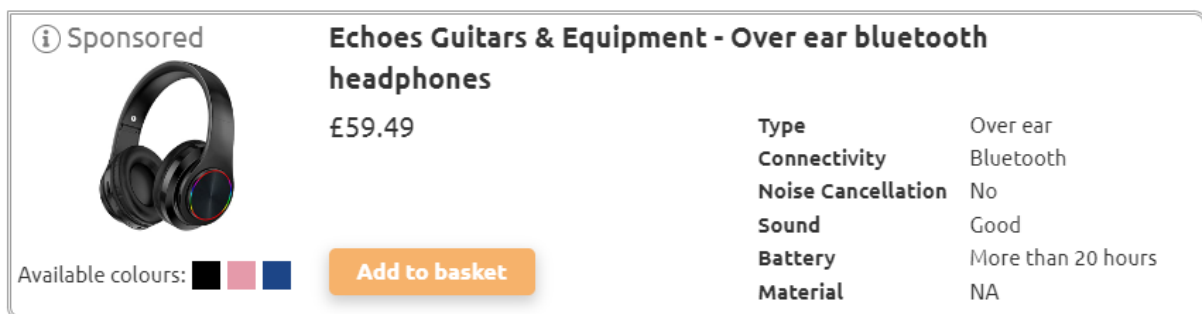


Figure 14: Example of 'sponsored' messaging

- **'iWebs' Choice Messaging:** Trial arm 6 incorporated the idea of 'best-sellers' or 'top-choice' messaging commonly seen on e-commerce platforms, using the phrasing 'iWebs Choice.' These boxes were listed at position 1, 2, and 5 on the site, indicating that the product was highly rated by the e-commerce site. We chose positions 1,2 and 5 as this reflected similar positions to the online platforms we looked at initially.

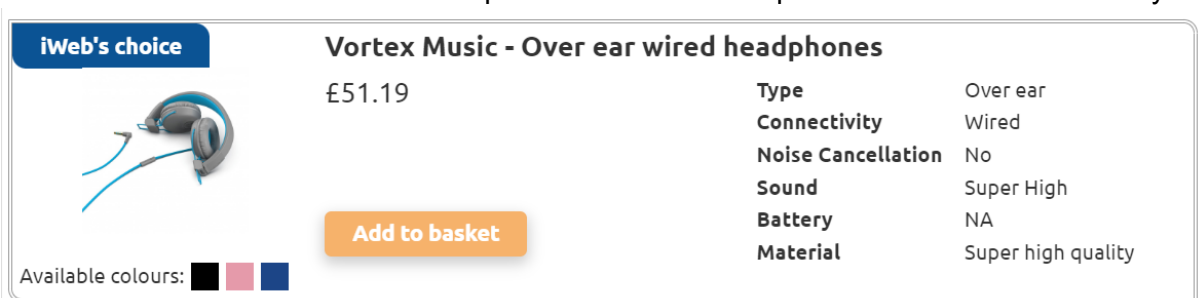


Figure 15: Example of 'iWebs' choice messaging

- **Sort Function:** Finally, trial arm 7 included a sort option in the commercially-focused algorithm. This option appeared in the top right corner of the page and offered participants three sort options: Default (commercially-focused algorithm), Price: Low to high, and Price: High to low.

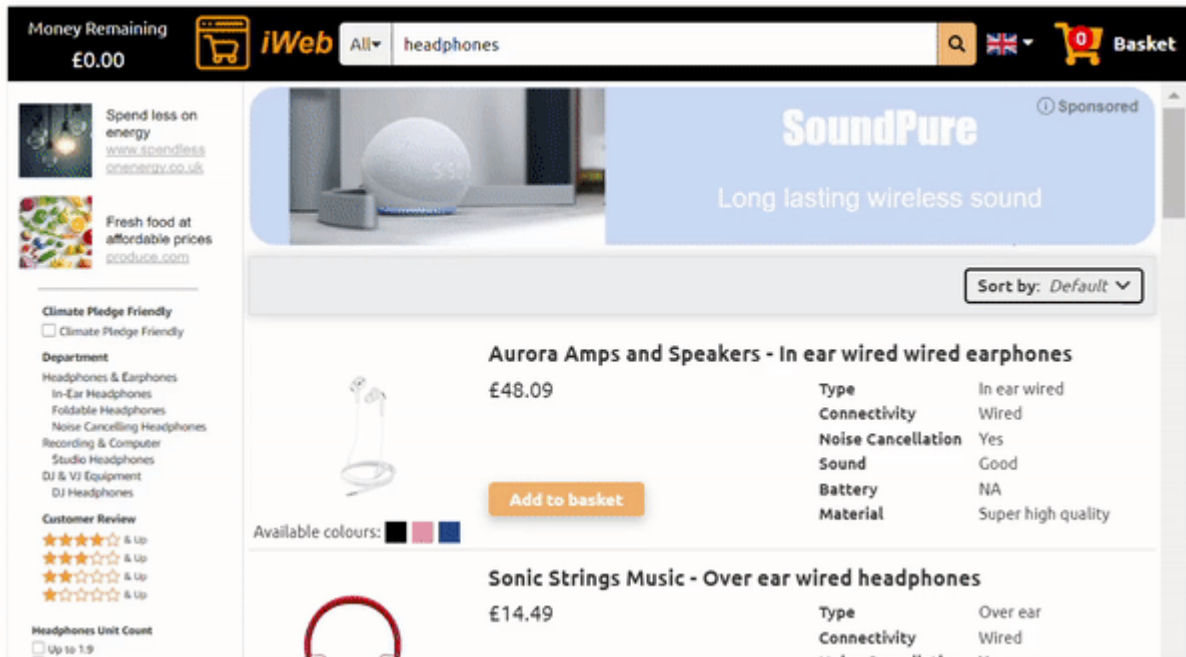


Figure 16: Overview the sort function

All of these additional features were added to specific trial arms, with the intent of testing their impact on our main outcome measures.

Appendix 5: Analysis Plan

The analysis strategy for this study was designed to assess the impact of different algorithms and different website features often used in conjunction with algorithms on consumer decision-making and potential overspending. We controlled for the following covariates, age, gender, ethnicity, noise cancellation preference⁴², and image colour to ensure that any observed effects were not confounded by these factors. We report the findings for all outcomes as treatment effects.

Primary Outcome 1 - Selection of top-scoring product

⁴² Noise cancellation was the only preference that was controlled for, as it was the only elicited preference that was also unbalanced in terms of price. This means, that all else equal, people who have a preference for noise cancellation technology will be less likely to find a product that matches their preferences, as the product that fits all of their preferences will be more likely to be outside of their price range.

The primary outcome for this study is the proportion of participants who select the top-scoring product⁴³ that is available to them. In addition to the covariates listed above we use total budget⁴⁴ as additional covariates in this regression.

The primary outcome is binary, and therefore we used a logistic regression to assess the effect of our treatment on this outcome.

$$Y_i \sim \text{bernoulli}(p_i); \text{logit}(p_i) = \alpha + \beta_1 T_i + \beta_2 \text{Cov}_i + \epsilon_i$$

Where:

- Y_i is a binary indicator for the participant i indicating the top-scoring product selection
- p_i is the probability of a positive selection of the top-scoring products for participant i
- α is the constant
- $\beta_1 T_i$ is a categorical indicator of treatment assignment for participant i .
- $\beta_2 \text{Cov}_i$ is vector of covariates used as controls in the regression
- ϵ_i is the error term for participant i

Primary Outcome 2 - Product proximity score

For the second primary outcome, we analysed a measure of proximity to a participant's top-scoring product. In developing this score, we made use of the consumer-focused algorithm scoring, outlined in Sections 3.3.3 and 3.4.2, and assumed that the product with the highest score in this algorithm was the top-scoring product for a consumer. This outcome measures the difference between a participants' top-scoring product score, and the product they selected, meaning, a participant who selected their top-scoring product will receive a score of zero. As an example, a participant whose top-scoring product had a consumer-focused algorithm score of 20, but selected a product with a score of 15, will receive a proximity score of -5. The lower the score, the worse off consumers were. This outcome is continuous, and was analysed for treatment effects using OLS regression.

$$Y_i = \alpha + \beta_1 T_i + \beta_2 \text{Cov}_i + \epsilon_i$$

Where:

- Y_i is the secondary outcome measure
- α is the constant
- $\beta_1 T_i$ is a categorical indicator of treatment assignment for participant i
- $\beta_2 \text{Cov}_i$ is a vector of covariates used as controls in the regression

⁴³ The top-scoring product is the product with the highest score determined by the consumer-focused algorithm

⁴⁴ We control for both the elicited budget and the budget available to spend on products in the simulation

- ϵ_i is the error term for participant i

Secondary Outcome - Financial Impact

In addition to the primary variable, we also analysed two secondary outcomes. The first was a measure of financial impact, this value is determined by dividing any overspend or underspend from a participant (overspend is any pound spent over the value of the top-scoring product) by the elicited budget. This measure allows us to determine the financial impact on each consumer generated in each experimental arm. This outcome is continuous, and was analysed for treatment effects using OLS regression. In addition to the covariates listed in the introduction to this section we use product price as an additional covariate in the regression.

Overspend is the amount of money above the amount they would have needed to spend to get their top-scoring product. This is not necessarily in the consumer's interest as they are spending money they don't need to. Underspend is the amount of money consumers are spending below the amount they are willing to, but receiving inferior goods is an indicator of an inefficient supply and demand market, which we will explore.

$$Y_i = \alpha + \beta_1 T_i + \beta_2 Cov_i + \epsilon_i$$

Where:

- Y_i is the measure of financial impact
- α is the constant
- $\beta_1 T_i$ is a categorical indicator of treatment assignment for participant i
- $\beta_2 Cov_i$ is a vector of covariates used as controls in the regression
- ϵ_i is the error term for participant i

Exploratory Outcome - Time spent searching on the simulated environment

The second exploratory outcome we look at for descriptive comparison between treatment arms is time in seconds consumers spend in the simulated e-commerce environment. We find this to be of particular interest to determine if there is any evidence of additional burden to searching caused by the experimental arms.

$$Y_i = \alpha + \beta_1 T_i + \beta_2 Cov_i + \epsilon_i$$

Where:

- Y_i is the exploratory outcome measure of time spent in the simulation
- α is the constant

- $\beta_1 T_i$ is a categorical indicator of treatment assignment for participant i
- $\beta_2 Cov_i$ is a vector of covariates used as controls in the regression
- ϵ_i is the error term for participant i

Appendix 6: Statistical Clustering Methodology

K-medoids is a statistical clustering algorithm that is similar to K-means clustering but uses a different method to determine the cluster centres. In K-medoids, the cluster centre is represented by one of the actual data points, whereas in K-means, the cluster centre is the mean of all the data points in the cluster. We used K-medoids over K-means because it is less sensitive to outliers, as the cluster centre is always represented by one of the actual data points, rather than a calculated mean.

The algorithm works by randomly selecting k data points to serve as the initial cluster centres. It then iteratively assigns each data point to the nearest cluster centre, and then calculates the total distance between each data point and its assigned cluster centre. It then selects a new data point to be the centre of the cluster that has the minimum total distance. This process is repeated until the algorithm converges, or until a specified number of iterations is reached.

Appendix 7: Analysis of participants who did not select their top-scoring product

One potential drawback of this study, as previously elaborated, is the reliance on participants' self-reported preferences, which may not always align with their genuine preferences. To counterbalance this, we performed several robustness tests, as described below.

We presented participants with their top-scoring product - as dictated by their declared preferences - after the e-commerce simulation, only if they hadn't initially chosen it. This involved approximately 97.43% of the random ranking group, 80.68% of the consumer-oriented algorithm group, 93.45% of the commercially-oriented algorithm group and 98.47% in the income-based algorithm group. We then inquired if, given the opportunity, they would swap their original choice for their top-rated product.

43% of these participants who didn't initially opt for their top-scoring product stated that they wouldn't choose it, suggesting a discrepancy between their self-reported and revealed preferences. On the other hand, 57% of participants expressed a willingness to choose the top-scoring product, broken down into: 67% in the random ranking group, 49% in the consumer-focused algorithm, 57% in the commercially-focused algorithm, and 64% in the income-based algorithm. This supports the theory that more participants are inclined to switch products in algorithms that make it difficult to discover their top-scoring product. Furthermore, the consumer-focused and commercially-focused algorithms had higher proximity scores, which could explain why many of these respondents were sufficiently satisfied with their choice and didn't want to switch to the top-scoring product.

A further concern was how to interpret the underspend, since participants who 'underspent' could feel satisfied with their chosen product due to perceived savings. An analysis of the subset of participants who spent less than the cost of their top-scoring product (n=2,993) revealed no difference in the propensity to switch, with 57% wishing to switch to the top-scoring product.

We carried out mean comparisons for our primary and secondary outcomes to verify if the results fluctuated when excluding participants unwilling to switch from the product they chose in the simulation, even if it wasn't the top-scoring one (34.7% in random arm, 60.5% in consumer-focused algorithm, 46.7% in commercially-focused algorithm and 37% of people in income-based algorithm). The results, presented in Table 13, confirm that the statistical significance of our findings do not change.

Table 13: Primary & secondary outcome measures when excluding people who would not switch to the top-scoring product

	Random ranking	Consumer-focused algorithm	Commercially-focused algorithm	Income-based algorithm
% selecting top-scored product	4%	32%	14%	7%
Financial impact (£)	£16.27	-£3.03	-£7.18	-£11.65
Proximity Score	-3.24	-0.71	-1.10	-4.47

Green and red shading indicate statistically highest/lowest value in row at a 95% confidence level ($p < 0.05$)
All statistical tests are conducted with the random algorithm as the comparison group

Finally, we questioned participants who didn't select their top-scoring product about specific preferences that their chosen product didn't meet. For instance, if a participant chose wireless headphones over their stated preference for wired headphones, we sought to understand why. The results of these queries are summarised in the following table.

Table 14: Reasons for not selecting a product that matched their original preferences

	Over-ear vs in-ear vs earbuds (n=822)	Wireless vs wired (n=326)	Noise cancellation technology (n=1977)
% who did not select a product in line with their specified preference	10%	4%	25%

I changed my mind after seeing the products	43%	34%	30%
I think the product I chose was good enough, considering value for money	25%	28%	34%
I couldn't find the product that matched my preferences	21%	22%	21%
I couldn't afford the headphones/earphones I wanted	16%	14%	18%
I made a mistake in my preferences earlier	11%	9%	7%
I didn't actually have a preference here	9%	6%	8%
I did not want to spend more time looking for my top-scoring product, so I chose the best product I could within a reasonable time	7%	9%	9%
Other reasons	5%	7%	5%

No statistical testing conducted

Appendix 8: Sample composition and Preferences

Table 15: Overview of the sample composition of the participants in the survey

Demographic	Category	Demographic	Category
Age	18-24: 14%	Region	South & East: 31%
	25-54: 52%		Midlands: 17%
	55+: 34%		London: 14%
Gender	Male: 50%		North: 25%
	Female: 50%		Scotland, Wales & NI: 14%
Ethnicity	White: 86%		
	BAME: 14%		

Income	Below £40k: 52%	Universal Credit	Yes: 19%
	Above £40k: 48%		Mental health challenges (last 12 months)

Appendix 9: Sample Preferences

Table 16: Overview of consumers' stated preferences, elicited prior to the online shopping simulation

Preference elicited	Question	Answer Options
Headphone type preference	When listening to music/podcasts using headphones, which of the following headphone types do you prefer?	Over-the-ear headphones: 19% In-ear wired earphones: 31% Earpods: 50%
Wireless or wired	Which of the following do you prefer?	Wireless over ear headphones: 64% Wired over ear headphones: 36%
Noise cancellation	Which of the following do you prefer?	Noise cancelling: 73% Non-noise cancelling: 27%
Ideal budget	Imagine you had lost or broken your current headphones/earphones, and wanted to purchase a new pair. What's the approximate amount you'd be willing to pay?	Mean: £103 Median: £80
Maximum budget	Realistically, what is the maximum amount you'd be willing to pay for the right headphones/earphones?	Mean: £134 Median: £100
Ranking of attributes	Please rank the following product attributes in terms of how important they are to you when purchasing headphones/earphones.	Ranking of attribute importance: 1. The sound quality of the product 2. Product is within budget 3. Product is an over the ear headphone/in-ear wired earphone/earpods style earphone

4. Whether the product is wired/wireless (bluetooth)
5. Battery life
6. Having noise cancellation technology
7. The material quality of the cable/wire for wired earphones/headphones

Appendix 10: Algorithm Diagnostics and Additional Summary Statistics

Here we provide an overview of some diagnostic results, which were used to further examine how the algorithms performed in practice. In Table 17 we examine the median rank of the top-scoring product for each algorithm, the median rank of the product a participant selected, the % of top-scoring products shown on the first page, by algorithm, and the % of participants selecting a product on the first page, by algorithm. These results show that a) the algorithms worked as intended, and b), that the more consumer preferences were taken into account, the earlier participants selected a product.

Table 17: Ranking order diagnostics median position and proportion on first page

	Median Rank of product <i>(by treatment, out of 432 products)</i>		% of products listed on 1st page <i>(by treatment)</i>	
	Top-scoring Product	Product Selected	Top-scoring Product	Product Selected
Random ranking	214	10	4%	67%
Consumer-focused algorithm	1	5	100%	88%
commercially-focused algorithm	14	6	55%	83%
Income-based	159	10	3%	69%

No statistical testing conducted

Appendix 11: Additional analysis of proximity score variable

While also considering the possibility that there might be more than one ‘top-scoring’ product for a consumer, and that penalising participants who choose a product very close to their top score product may impact the results of the study. To test this, we analysed what may have happened with our primary outcome if we expanded what we consider an ‘top-scoring’ product. To do this we ran 3 separate means tests on our primary outcome, % of participants

selecting a top-scoring product, by changing our definition of what a top-scoring product is. First, of participants who didn't select their top-scoring product, we take the 5% with the highest proximity scores, and consider them to have selected a 'top-scoring' product. We then do this again at the 12.5% and 20% levels. The results are shown in Table 18. As expected, we see increases in all algorithms in terms of % of people selecting their top-scoring product, and our significance tests show the same results as our primary analysis.

Table 18: Broadened top product criteria analysis

% selecting the 'top-scoring' product	Random ranking	Consumer-focused algorithm	Commercially-focused algorithm	Income-based algorithm
% selecting top-scoring product (for reference)	3%	19%	7%	2%
Top 5% of proximity scores considered top-scoring	4%	28%	11%	3%
Top 12.5% of proximity scores considered top-scoring	5%	37%	19%	4%
Top 20% of proximity scores considered top-scoring	7%	44%	28%	5%

Green and red shading indicate statistically highest/lowest value in row at a 95% confidence level ($p < 0.05$) All statistical tests are conducted with the random algorithm as the comparison group

Appendix 12: Sub group analysis

Table 19: Proportion of respondents selecting their top-scoring product by subgroup and algorithm

	Age			Gender		Ethnicity		Mental health challenges		Digital confidence		Universal credit	
	Under 25	25-54	55+	Female	Male	BAME	White	No	Yes	Low	High	No	Yes
Consumer-Focused Algorithm	26%	18%	19%	19%	20%	19%	20%	20%	19%	20%	18%	19%	22%
commercially-focused algorithm	5%	6%	8%	7%	6%	6%	7%	7%	6%	6%	7%	7%	6%
income-based algorithm	0%	2%	2%	2%	1%	0%	2%	2%	0%	1%	2%	2%	1%

No statistical testing conducted, bolding indicates largest variances

Table 20: Average financial impact by subgroup and algorithm

	Age			Gender		Ethnicity		Mental health challenges		Digital confidence		Universal credit	
	Under 25	25-54	55+	Female	Male	BAME	White	No	Yes	Low	High	No	Yes
Consumer-Focused Algorithm	-£2.87	-£0.48	-£7.29	-£2.43	-£3.65	-£2.97	-£3.18	-£4.02	-£0.48	-£1.55	-£4.74	-£3.53	-£1.51
commercially-focused algorithm	-£4.81	-£7.57	-£9.88	-£5.77	£10.15	-£9.24	-£7.76	-£8.67	-£5.76	-£5.30	£10.84	-£7.74	-£9.00
income-based algorithm	£17.35	£19.89	£12.59	£13.92	£19.11	£22.55	£16.07	£20.40	-£6.75	£18.55	£15.74	£21.45	£1.42

No statistical testing conducted, bolding indicates largest variances

Table 21: Proximity score by subgroup and algorithm

	Age			Gender		Ethnicity		Mental health challenges		Digital confidence		Universal Income	
	Under 25	25-54	55+	Female	Male	BAME	White	No	Yes	Low	High	No	Yes
Consumer-Focused Algorithm	-0.61	-0.78	-0.87	-0.78	-0.79	-0.80	-0.79	-0.85	-0.60	-0.74	-0.84	-0.83	-0.63
commercially-focused algorithm	-1.21	-1.08	-1.11	-1.09	-1.13	-1.16	-1.10	-1.13	-1.03	-1.09	-1.13	-1.11	-1.11
income-based algorithm	-5.66	-4.87	-4.40	-4.79	-4.86	-5.14	-4.76	-4.86	-4.73	-5.14	-4.49	-4.83	-4.85

No statistical testing conducted, bolding indicates largest variances

This publication is available from: www.gov.uk/dsit

If you need a version of this document in a more accessible format, please email alt.formats@dsit.gov.uk. Please tell us what format you need. It will help us if you say what assistive technology you use.