



In-lab versus web-based eye-tracking in decision-making: A systematic comparison on multiple display-size conditions mimicking common electronic devices

Sebastián Muñoz¹ · Vladimir Maksimenko³ · Bastian Henriquez-Jara^{2,4} · Prateek Bansal³ · Omar David Perez^{1,4}

Received: 3 April 2025 / Accepted: 1 October 2025 / Published online: 10 November 2025
© The Psychonomic Society, Inc. 2025

Abstract

Eye-tracking has gained considerable attention across multiple research domains. Recently, web-based eye-tracking has become feasible, demonstrating reliable performance in perceptual and cognitive tasks. However, its systematic evaluation in decision-making remains unknown. Here we compare a laboratory-based eye tracker (the EyeLink 1000 Plus) with a webcam-based method (WebGazer) across two discrete-choice experiments. We systematically manipulated display size to approximate common device classes (monitor, laptop, tablet, mobile) and task complexity (simple vs. complex choice matrices). We find that on larger displays and simpler tasks, WebGazer produces gaze patterns and parameter inferences from computational models of behavior comparable to EyeLink. However, reliability diminishes on smaller displays and with more complex choice matrices. These results provide the first systematic evaluation of web-based eye tracking for decision-making research and offer practical guidance regarding its viability for online behavioral studies.

Keywords Eye-tracking · Attention · Discrete choice

Introduction

Over the past decade, various academic fields have shown an increasing interest in conducting behavioral experiments online. This trend is primarily attributable to three key

factors: scalability, cost-effectiveness, and the COVID-19 pandemic. Crowdsourcing platforms such as Amazon Mechanical Turk and, more recently, Prolific have enabled researchers to collect large amounts of data that meet specific inclusion criteria. These platforms facilitate the efficient collection of substantial volumes of data within a brief time-frame – typically a few hours.

From its inception, a primary concern has been the degree of reliability of online compared to laboratory experiments, as it is natural to anticipate that results may exhibit greater variability in online settings due to several factors such as participants' country, time of day, dropout rates, inattention, compensation for participation, potential for cheating, technological constraints, among others (Crump et al., 2013). While some of these issues can be addressed through methods such as implementing attention checks or ensuring participation at specific times of day, other challenges are less easily resolved. However, a seminal study conducted about a decade ago by Crump et al. (2013) demonstrated that these potential issues were less concerning than previously expected, obtaining a high replication rate of classic psychology experiments using the MTurk online platform, lending support to the reliability of online experimentation and encouraging its broader application across different fields of study.

Sebastián Muñoz and Vladimir Maksimenko contributed equally to this work.

✉ Omar David Perez
omar.perez.r@uchile.cl
Bastian Henriquez-Jara
bastian.henriquez@uchile.cl
Prateek Bansal
prateekb@nus.edu.sg

- ¹ Department of Industrial Engineering, Universidad de Chile, Santiago, Chile
- ² Department of Civil Engineering, Universidad de Chile, Santiago, Chile
- ³ Behavioural Cognitive Science Lab, National University of Singapore, Singapore, Singapore
- ⁴ Complex Engineering Systems Institute (ISCI), Santiago, Chile

A critical advantage of laboratory-based experimentation over its online counterpart is the ability to collect psychophysiological data alongside behavioral measures. One such method, which has become increasingly prevalent, is eye-tracking. This method enables the analysis of visual attention and eye movement patterns that correlate with internal brain mechanisms and neurotransmitter activity involved in learning and decision-making processes (Cavanagh et al., 2014; Dayan et al., 2000; Pool et al., 2019). The application of eye-tracking in laboratory settings has yielded significant insights into these underlying mechanisms in both humans and animals across diverse fields, including psychology, neuroscience, marketing, engineering, and economics (Bansal et al., 2024; Borozan et al., 2022; Cavanagh et al., 2014; Kraljich et al., 2010; Le Pelley et al., 2015; Martinovici et al., 2023; Pool et al., 2019; Reutskaja et al., 2011; Shimojo et al., 2003).

Prompted by the emergence of computer software specifically designed to leverage built-in cameras for remote gaze tracking, the question of whether eye-tracking can be reliably implemented in online experimentation to produce results comparable to those obtained in laboratory settings has gained increasing attention in recent years. One of these tools is WebGazer, a JavaScript-based eye-tracking library that uses common webcams to infer gaze locations in real time, self-calibrates by observing user interactions, and runs entirely in the client browser without requiring video data to be sent to a server (Papoutsaki et al., 2016).

The performance of WebGazer has been systematically evaluated in a number of recent studies across various research fields. For example, Steffan et al. (2024) validated WebGazer for early childhood research, comparing its performance with in-lab methods and highlighting its potential despite higher attrition rates. A similar approach was followed by Hutt et al. (2023), who employed WebGazer to track mind-wandering, demonstrating its ability to predict cognitive states based on gaze patterns. Slim et al. (2024) extended this approach by conducting a direct comparison between WebGazer and in-lab eye-tracking methods across multiple paradigms, including language processing, visual attention, and cognitive load assessment. Their study suggested that while WebGazer may exhibit greater variability compared to traditional laboratory-based systems, it was still capable of capturing broad attentional patterns across a range of experimental paradigms.

While prior findings are encouraging for web-based eye-tracking methods, research to date has focused mainly on validating webcam-based tracking in perceptual and cognitive tasks. This leaves open the question of whether WebGazer can be reliably applied to studies of choice behavior and economic decision-making – a domain in which attention plays a central role in explaining human choices and has been highly valuable for informing policy-makers

about market preferences. In this paper, we address this gap by applying WebGazer to discrete choice experiments (DCEs), a widely used method for studying decision-making in applied economics, marketing, psychology, and engineering (Bliemer & Rose, 2024; Louviere & Hensher, 1982).

In DCEs, regions of interest (ROIs) are clearly defined across trials and easy to identify, allowing reliable estimation of behavioral parameters in models that incorporate attentional factors in subjects' decisions. However, it is not clear if the same parameters will be identifiable or similar when using WebGazer, given its poorer temporal and spatial resolution relative to physical eye trackers. This concern is amplified in more complex tasks, where ROIs (e.g., attribute values of different options) are close together and the number of ROIs increases, possibly making it more difficult for web-based eye trackers in general. Critically, we show this to be the case not only at the level of raw gaze metrics, but at the model level: inferences about WTP/SVTTS estimates are also affected by these factors, affecting conclusions about the latent processes driving behavior in decision-making tasks.

Unlike prior studies that have focused mainly on raw gaze metrics, we also employ behavioral models that incorporate attention as a key explanatory variable. In addition, we manipulate two critical factors: task complexity and display size. Task complexity varied across experiments to test whether attentional patterns shift under different cognitive demands, while display size was varied within subjects to assess how viewing conditions, such as those encountered on different electronic devices, influence fixation behavior and parameter estimates. Because display size alters the visual angle of the choice matrix on the retina, and prior work shows that stimulus size can affect eye-movement patterns (e.g., smaller stimuli eliciting more centralized viewing with fewer, longer fixations), we use it to approximate device classes and evaluate the robustness of the different eye-tracking systems. Our findings thus provide the first systematic evaluation of web-based eye tracking in decision-making, providing boundary conditions for its viability in online behavioral studies.

Experimental methods

Choice tasks

To examine the reliability of web-based and laboratory-based eye-tracking in decision-making, we designed two discrete choice experiments (DCEs). DCEs are widely used in applied economics, marketing, engineering (for example, in transportation research) and psychology to investigate how individuals make trade-offs between competing attributes of different available alternatives (Bansal et al., 2024). This

experimental framework allows researchers to systematically vary attribute levels and estimate preference parameters (i.e., how much importance people ascribe to each attribute) based on observed choices.

Experiment 1 (pizza choices) featured a simple binary choice paradigm in which participants selected between two alternatives of pizzas based on two attributes: quality and cost. Quality was presented using a scale with five possible values [1, 3, 5, 7, 9], similar to the way various foods are presented in online websites, while cost was displayed as a monetary value [5, 7, 9, 11, 13, 15] Singapore dollars (SGD). By systematically varying these attributes across trials, we were able to assess how individuals weight quality relative to price in their decision-making. The simplicity of this (2×2) design facilitated the extraction of clean fixation patterns while maintaining ecological validity.

Experiment 2 increased task complexity by presenting participants with three transportation alternatives, each characterized by three attributes: cost, travel time, and comfort (a 3×3 design). Cost was displayed as a fare price, travel time as a duration in minutes, and comfort as a categorical descriptor (e.g., low, medium, high). This multi-attribute, multi-alternative design allowed us to examine how increased cognitive demands may influence gaze tracking and behavioral parameters in the two eye-tracking methods.

Critically, to test robustness across device classes, we ran both experiments at four display-size conditions (monitor, laptop, tablet, mobile) chosen to mimic typical use on those devices. This within-subject manipulation allowed us to assess how display size affects gaze-tracking quality and model-based behavioral parameters for both eye-tracking methods (Fig. 1).

Participants and apparatus

Forty participants (14 males, 26 females; mean age = 23.9, SD = 6.66) were recruited for Experiment 1, and 40 participants (14 males, 26 females; mean age = 22.3, SD = 2.35) participated in Experiment 2. All participants were students or staff at the National University of Singapore (NUS) and were recruited through the NUS Student Work Scheme. The sample size was determined based on prior research employing similar decision-making paradigms and ensured sufficient statistical power for model estimation (Bansal et al., 2024).

Participants were pre-screened to exclude subjects wearing glasses that could interfere with eye-tracking measurements due to reflection artifacts. All participants provided informed consent before participation. The study protocol was approved by the Institutional Review Board of the National University of Singapore (NUS-IRB-2023-1055). Each participant received compensation of S\$10 for their participation.

Procedure

The experimental procedure followed four main steps, identical for both experiments (Fig. 2). We first calibrated the external eye-tracker, followed by the calibration of the webcam-based eye-tracker. After calibration, participants were presented with the instructions of the task, after which they started the experiment. Before each version of the experiment (i.e., variations of display size), a recalibration was performed separately for the two eye-tracking methods.

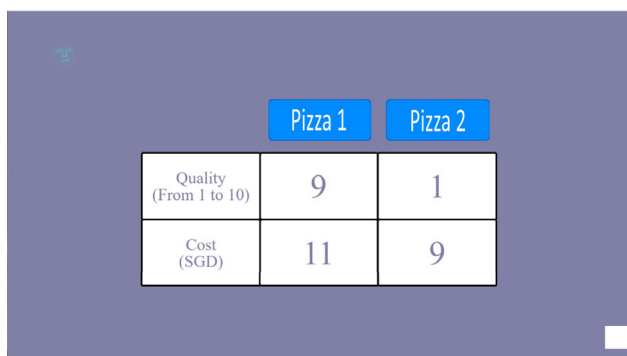
Participants were asked to make their choices with a mouse click (pizzas in Experiment 1; transport mode in Experiment 2) and completed four display-size choice tasks (monitor, laptop, tablet, mobile), implemented by scaling stimuli and setting viewing distance to approximate typical visual angles for those device classes; we did not test actual physical devices. The order of the four display-size choice tasks was randomized across participants. Within each task, participants completed 12 trials¹ presented in random order. The attribute (up-down) position was randomized across trials for each participant; alternative positions (left-right) were held constant within each experiment and across participants. We did not include practice or catch trials. Each task lasted about 3–5 min; we recalibrated both eye trackers before each task and offered short breaks between them. Including instructions, setup, recalibrations, and breaks, the total session time was about 40 min.

For each participant, we recorded gaze data using the EyeLink 1000 Plus and WebGazer simultaneously. EyeLink was calibrated first, followed by WebGazer, and both were recalibrated before each display-size task (EyeLink was run in Remote Mode. Following SR Research guidance, we used a 13-point calibration.) Each task lasted approximately 3–5 min; including instructions, setup, recalibrations, and short breaks between tasks, the total session time was about 40 min.

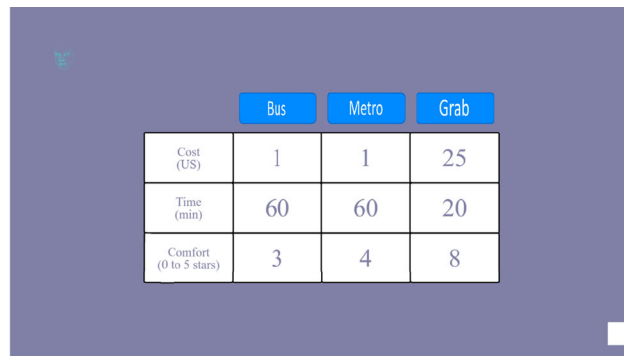
In Experiment 1 (pizza choices), we generated a *D-efficient* design using *Ngene* (ChoiceMetrics; standalone design software)², avoiding dominant alternatives and ensuring that attribute levels were well balanced within and across the 12 choice trials (i.e., each level appears approximately equally often and level combinations are close to orthogonal). In Experiment 2 (transport choices), we used an optimal fractional factorial design with the R package *choiceDes*, again yielding 12 trials per task with balanced levels and low collinearity among attributes. The same 12-choice trials were used in each display-size task. The order of presentation of trials was randomized for each participant in a task.

¹ In the economics and transportation literature, each choice made by subjects is usually referred to as a “choice task”; we use the term “trial” here, following the terminology employed in psychology and neuroscience.

² <https://www.choice-metrics.com/ngene>



	Pizza 1	Pizza 2
Quality (From 1 to 10)	9	1
Cost (SGD)	11	9

(a) *Experiment 1: Binary choice task*


	Bus	Metro	Grab
Cost (US)	1	1	25
Time (min)	60	60	20
Comfort (0 to 5 stars)	3	4	8

(b) *Experiment 2: Multi-attribute choice task*

Fig. 1 Experimental design of choice tasks for Experiment 1 (pizza choices) (A) and Experiment 2 (transport mode choices) (B). Participants completed 12 choice trials per display condition (monitor, laptop, tablet, and mobile) while their gaze was tracked simultaneously by both

EyeLink and WebGazer methods of eye tracking. The display layout was adjusted to mimic the screen size of each device. Here, the laptop condition is shown

Eye-tracker calibration

For all subjects, the two eye-tracking methods were calibrated sequentially, starting from the EyeLink, followed by WebGazer. This process was repeated between the first and second experiments.

EyeLink The EyeLink was calibrated using a three-step process with a standard 13-point fixation method, where participants fixated on a series of targets presented sequentially on the screen. The system adjusted its settings to map the participant's gaze coordinates accurately across the display. Following calibration, participants fixated on predefined targets to assess accuracy. The recorded gaze positions were compared with expected target locations, ensuring tracking

error remained within an acceptable range (typically $< 0.5^\circ$ of visual angle). A drift check was also performed to maintain accuracy throughout the session: participants focused on a reference point, and any deviations beyond a set threshold triggered a recalibration routine before proceeding.

Gaze data were recorded at 1000 Hz using an EyeLink 1000 Plus (SR Research, Ottawa, Canada) and the built-in webcam of each device. Choice tasks were presented on 1920×1080 resolution, 25-inch monitors, with participants seated approximately 60 cm from the screen. To mimic natural viewing conditions for webcam-based eye-tracking, we did not use a chin rest in these experiments.

WebGazer For the web-based eye-tracking, we used WebGazer's standard method. Participants were presented with

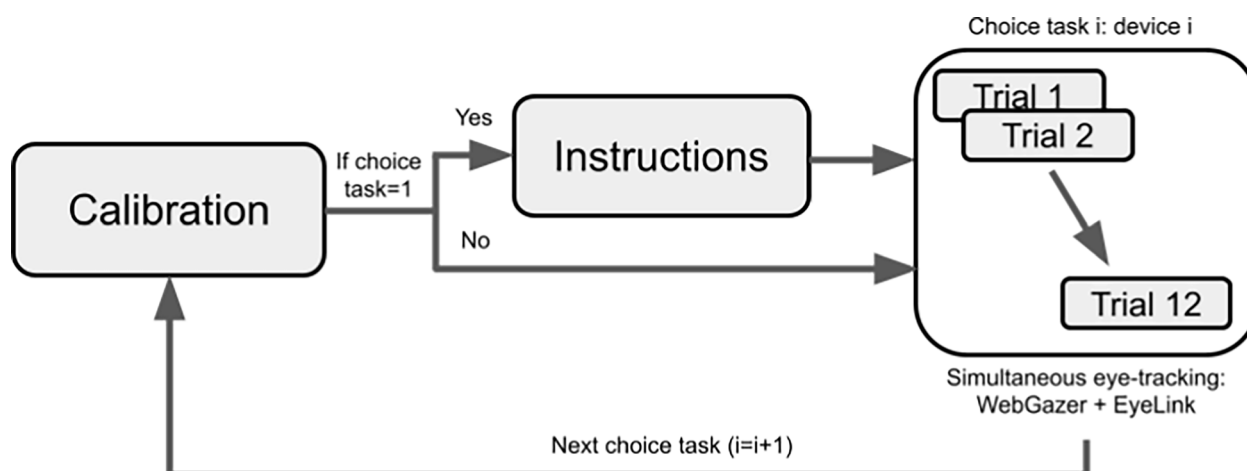


Fig. 2 Experimental procedure. For each participant, calibration of each eye-tracker (EyeLink first, then WebGazer) was followed by a single set of instructions for all experiments. Each experiment comprised a

12-trial choice task (2x2 for the simplest version of Experiment 1; 3x3 for the more complex version of Experiment 2)

a sequence of eight circles, each appearing one at a time in a random order along the edges of the screen. Each circle remained visible for one second before disappearing, at which point a simulated mouse click was generated. Following this initial calibration sequence, a final fixation check was performed. A single circle appeared at the center of the screen for three seconds, during which WebGazer computed the proportion of estimated gaze points that fell within this central region. For calibration success, we used the default WebGazer's threshold criterion. All participants passed calibration in both experiments.

Statistical methods and behavioral modeling

In this paper we aimed to compare the data produced by EyeLink and WebGazer in terms of (1) fixation counts and durations across attributes and alternatives; (2) the consistency of parameter estimates from decision-making models that integrate attention mechanisms in discrete choice experiments (DCEs); (3) the robustness of these findings across four distinct display sizes mimicking monitor, laptop, tablet, and mobile devices. To these ends, we first report raw gaze metrics (number and duration of fixations), followed by parameter estimations of a behavioral model that incorporates attentional factors in subjects' decision process.

Statistical analyses of fixations and behavioral modeling were performed in R using the RStudio IDE (RStudio Team, 2020); R and Python were used for data visualization and fixation analysis. Parameter estimation for the behavioral model was performed using the Apollo choice modeling package in R (Hess & Palma, 2019).

Fixation analysis

We analyzed gaze at two ROIs, separately, one for the alternative-level cells (Exp. 1: two alternatives; Exp. 2: three alternatives), and one for the attribute-level cells. (Experiment 1: Quality, Cost; Experiment 2: Time, Comfort, Cost). Figure 3 shows a representative choice screen with gaze data for Experiment 1 (pizza choices) in the monitor condition, separated by eye tracking method.

To analyze gaze data, for each experiment, we ran two separate repeated-measures ANOVAs (one for alternatives and one for attributes as ROIs). Each ANOVA included as within-subject factors the tracking method (EyeLink, WebGazer), the display (monitor, laptop, tablet, mobile) and the ROI identity (Alternatives: 2 or 3 levels, depending on the experiment; Attributes: 2 or 3 levels, depending on the experiment). When a main effect or interaction was significant, we ran Holm-corrected paired *t* tests on the relevant marginal means and report the direction of the effect (which method/display/ROI is higher).

ROIs and gaze metrics

We defined ROIs as rectangles matching the visible table cells that contained the decision information (attribute values and alternative cells), defined in stimulus coordinates and proportionally scaled for each display size. We report two fixation-based ROI measures averaged across trials per each display condition: (i) *Fixation count*: the number of fixation events in each ROI (in the figures, we plot the proportion of fixations per trial, $\text{count}_{\text{ROI}}/\text{count}_{\text{trial}}$). (ii) *Mean relative fixation duration*: the average duration of fixation events in each ROI. Trials with no fixations to a given ROI are undefined for mean duration and were treated as missing for that outcome (not zero). We use *relative fixation duration* in the plots to refer to the mean fixation duration in a given ROI divided by the mean fixation duration of the same trial, producing a unitless ratio centered near 1.³

Inclusion criteria

We applied the same rule to both eye-tracking methods. We kept all completed trials with a valid gaze. Fixation duration is defined only when a given ROI received at least one fixation; if a participant had no fixations on an ROI in any display condition, we excluded that participant from the fixation-duration ANOVA for that analysis. In Experiment 1 (pizza choices), we discarded three participants from the alternatives analysis and three from the attributes analysis. In Experiment 2 (travel mode choices), we discarded ten participants from the alternatives analysis and five from the attributes analysis. These exclusions apply only to the fixation-duration analyses; all other analyses include the full sample. As a robustness check, we re-ran the analysis at the trial level using a linear mixed-effects model in *lmer* that keeps all participants and treats those ROI-trials as missing; the results were consistent with the ANOVA results (see Appendix).

Behavioral modeling

Our second aim was to test whether the two eye-tracking methods would yield comparable parameter estimations from a discrete-choice behavioral model. To this end, we estimated subjects' preferences using an integrated choice and latent variable model (ICLV; Piras et al., 2020; Vij & Walker, 2016). These models extend the traditional random utility maximization (RUM) framework (McFadden, 1972, see next section) by incorporating latent variables, such as attention and perception, which are assumed to be not directly observable. ICLV models are widely applied in marketing,

³ EyeLink fixations were taken from the device's native parser; WebGazer data were parsed offline so that both methods yield comparable fixation events.

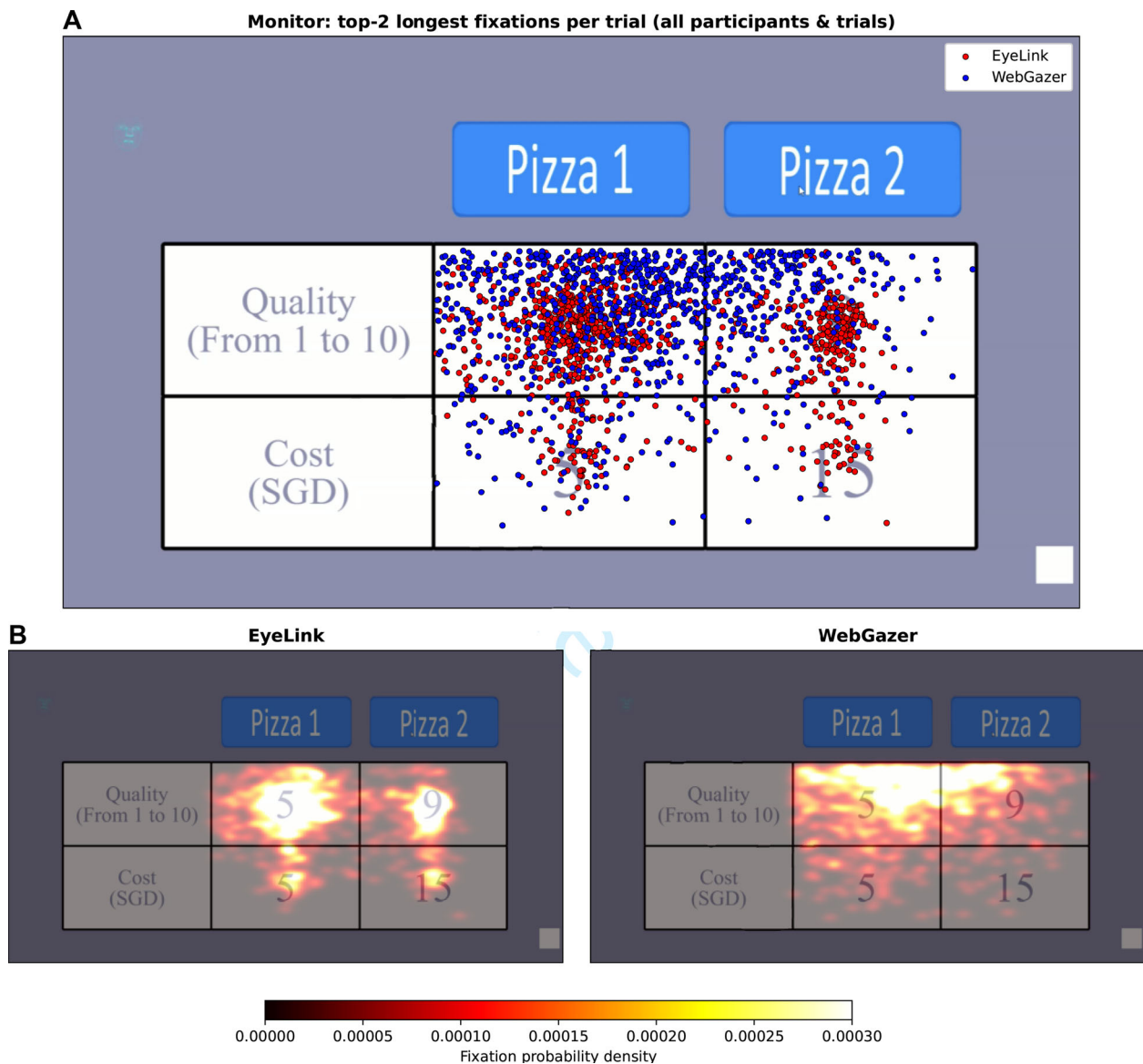


Fig. 3 Fixation distributions. **A.** Two longest fixations per trial (centroids), pooled across all participants for Experiment 1 (pizza choices); red = EyeLink, blue = WebGazer. **B.** Fixation heatmaps shown separately for EyeLink and WebGazer. Brighter colors indicate higher

fixation probability. Rectangles mark ROIs (alternative and attribute cells). Numbers inside cells are example attribute levels from a representative screen

engineering, transportation research and behavioral economics, making them an ideal tool for integrating visual attention data into discrete choice experiments (Bansal et al., 2024).

In this study, we used a latent information processing (LIP) model (Krucien et al., 2017), which explicitly links visual attention to decision weights in DCEs. If the two eye-tracking methods provide comparable data, parameter estimates from the LIP model should remain stable between them and be consistent across display sizes. Below, we briefly outline the structure of the RUM model and its extension in the LIP framework.

Random utility maximization (RUM)

In standard discrete choice models, the utility U_{nj} of alternative j for individual n consists of a systematic component V_{nj} and an unobserved stochastic term ε_{nj} :

$$U_{nj} = V_{nj} + \varepsilon_{nj}, \quad (1)$$

where the deterministic utility is typically defined as a linear function of attributes:

$$V_{nj} = \sum_{k=1}^K \beta_k x_{nj,k}, \quad (2)$$

where x_{njk} represents the value of attribute k , and β_k is its associated weight. Assuming the error term follows an extreme value type I (Gumbel) distribution, the probability of choosing alternative i follows a multinomial logit (MNL) form:

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_{j \in C} \exp(V_{nj})}. \quad (3)$$

The parameters β_k are estimated using maximum simulated likelihood estimation using the Apollo package in R (see [Appendix](#) for modeling details).

Latent information processing (LIP) model

In the LIP model, visual attention is explicitly modeled as a latent process that influences decision weights (the importance of each attribute in the final decision). The utility function is thus modified to include a latent information processing (IP) factor:

$$U_{nj} = \sum_{k=1}^K \tilde{\beta}_k x_{njk} + \varepsilon_{nj}, \quad (4)$$

where the adjusted coefficients $\tilde{\beta}_k$ depend on individual attention:

$$\tilde{\beta}_k = \beta_k + \alpha_k I P_n. \quad (5)$$

Here, $I P_n$ is a latent variable representing an individual's information processing, and α_k quantifies how attention influences decision weights. The latent process $I P_n$ is linked to observed fixation data via a measurement equation. In this way, the LIP model allows us to examine whether differences in visual attention lead to systematic variations in choice behavior. If the two eye-tracking methods yield consistent estimates, the inferred relationships between visual attention and choice preferences should remain stable across the two eye-tracking methods. A detailed mathematical derivation of the LIP model, including estimation procedures, is provided in the [Appendix](#).

Results

Statistical analysis of fixations

Experiment 1: pizza choices

Fixation count

Alternatives We first tested whether the number of fixations on the two different alternatives (pizza1, pizza2) captured by the EyeLink and WebGazer differed, and whether this difference remained stable across display sizes. The results

are shown in Fig. 4 A–B. We found a significant main effect of alternative ($F(1, 39) = 175.30, p < .001$), and a significant interaction between alternative and display size ($F(3, 117) = 18.41, p < .001$), suggesting that participants fixated more on pizza1 ($M = 0.299, SE = 0.005$) than on pizza2 ($M = 0.186, SE = 0.004$) on average, but that this difference varied across display sizes. Post hoc tests (see Fig. 4) showed that the difference in fixations between pizza1 and pizza2 was consistently significant for all display sizes with the EyeLink. With WebGazer, however, the difference was not significant in the mobile condition. These results show that the difference in the number of fixations to alternatives was reliably detected by both eye-tracking methods, with the exception of WebGazer in the mobile condition.

Attributes We then examined whether the number of fixations differed between attributes (quality and cost). We found no significant main effect of attribute ($F(1, 39) = 0.03, p = .873$), nor any interaction involving attribute and method or display size. This suggests that participants did not show a preference in fixation for one attribute over the others, and that this pattern was consistent across methods and display sizes (see Fig. 4 C–D).

Fixation durations

Alternatives Next, we tested whether fixation durations differed across alternatives. There was a significant main effect of eye-tracking method ($F(1, 36) = 39.62, p < .001$), suggesting that, on average, slightly longer fixations were estimated by WebGazer ($M = 1.002$ s, $SE = 0.002$) than by EyeLink ($M = 0.968$ s, $SE = 0.004$). No other significant main effects or interactions were found (Fig. 5 A–B).

Attributes Finally, we analyzed whether the duration of fixations on different attributes (quality and cost) differed, and whether this pattern remained stable across display sizes and eye-tracking methods. We found a significant effect of eye-tracking method ($F(1, 36) = 40.31, p < .001$), reflecting slightly longer fixations detected by Webcam-based eye-tracker, but no other significant main effects or interactions (all $ps > .1$). These results suggest that participants spent similar amounts of time fixating on each attribute, and that this pattern remained consistent across both eye-tracking methods and display sizes (Fig. 5 C–D).

Experiment 2: Transport mode choices

Fixation counts

Alternatives In Experiment 2, we found a significant main effect of alternative ($F(2, 78) = 111.167, p < .001$), indicating that participants fixated more on some alternatives than others. Metro ($M = 0.166, SE = 0.003$) attracted more

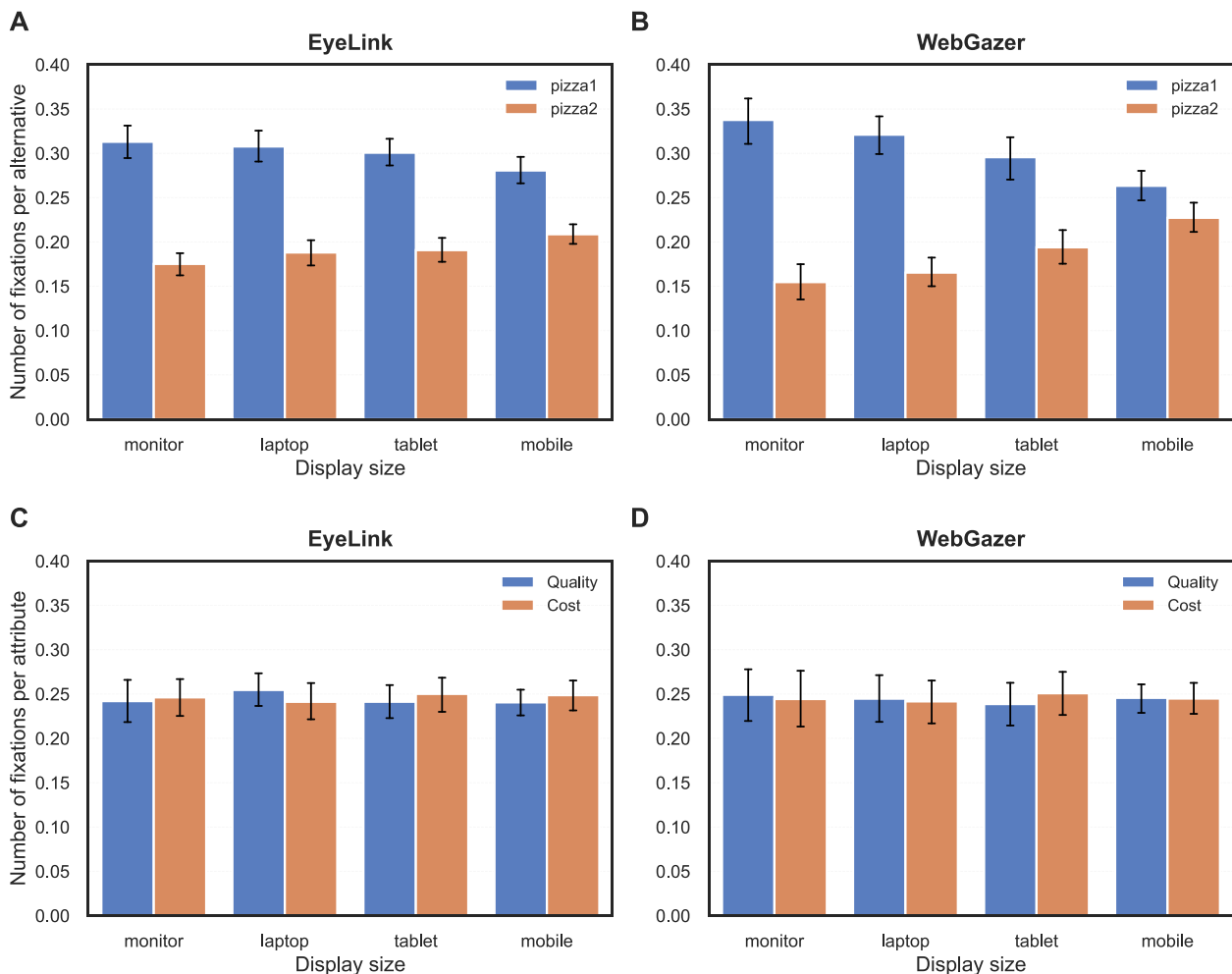


Fig. 4 Fixation counts across alternatives (**A**: EyeLink, **B**: WebGazer) and attributes (**C**: EyeLink, **D**: WebGazer) by display size for Experiment 1 (pizza choices). Data are presented as mean values across participants with 95% within-subject confidence intervals (based on 1000 bootstrap iterations)

fixations than Bus ($M = 0.088$, $SE = 0.005$; $t(39) = 12.66$, $p < .001$, uncorrected) and Grab ($M = 0.067$, $SE = 0.005$; $t(39) = 17.46$, $p < .001$, uncorrected). Bus attracted more fixations than Grab ($t(39) = 2.37$, $p = .022$, uncorrected). We also found a significant interaction between alternative and display size ($F(6, 234) = 4.217$, $p < .001$), and a three-way interaction between method, alternative and display size ($F(6, 234) = 8.747$, $p < .001$). Post hoc tests showed that metro attracted more fixations than bus and grab in the monitor condition, but this difference flattened in the laptop, tablet, and mobile settings, where bus and grab did not differ significantly. This suggests that WebGazer reliably detects alternative-based fixations in large displays, but this reliability may be affected in smaller displays (Fig. 6 A–B).

Attributes We next examined whether the number of fixations differed between the three attributes of Experiment 2 (comfort, travel time, cost) and whether this pattern remained stable across eye-tracking methods and display sizes. The

analysis revealed a significant main effect of attribute, ($F(2, 78) = 8.25$, $p < .001$), suggesting that some attributes consistently attracted more fixations than others. Travel time ($M = 0.113$, $SE = 0.002$) attracted more fixations than Comfort ($M = 0.097$, $SE = 0.002$; $t(39) = 3.80$, $p < .001$, uncorrected) and Cost ($M = 0.106$, $SE = 0.002$; $t(39) = 2.80$, $p = .008$, uncorrected). There was no significant difference in fixation count between comfort and cost ($t(39) = 1.41$, $p = .168$, uncorrected). No other main effects or interactions reached significance. These results suggest that participants' attention was distributed unevenly across attributes, but that this allocation was robust across different display conditions and eye-tracking methods (Fig. 6 C–D).

Fixation durations

Alternatives We next analyzed whether the duration of fixations on different alternatives (metro, bus, grab) differed, and

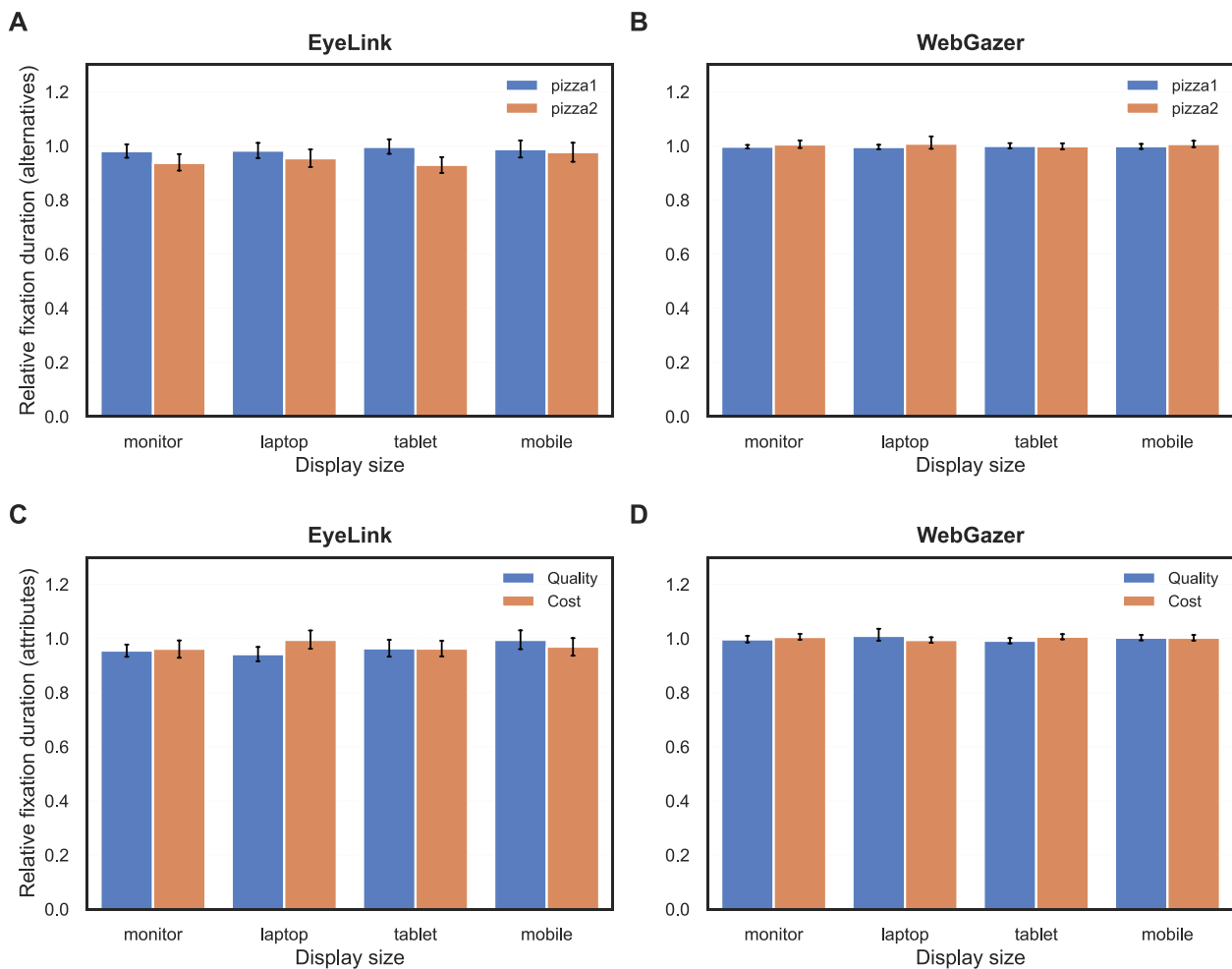


Fig. 5 Fixation durations across alternatives (**A**: EyeLink, **B**: webcam) and attributes (**C**: EyeLink, **D**: webcam) by display size for Experiment 1 (pizza choices). Data are presented as mean values across participants with 95% within-subject confidence intervals (based on 1000 bootstrap iterations)

whether this pattern remained stable across display sizes and eye-tracking methods.

We found a significant main effect of alternative ($F(2, 58) = 10.51, p < .001$) and eye-tracking method ($F(1, 29) = 21.76, p < .001$). Metro ($M = 1.006, SE = 0.003$) attracted longer fixation durations than Bus ($M = 0.960, SE = 0.007; t(39) = 5.301, p < .001$, uncorrected) and Grab ($M = 0.978, SE = 0.007; t(39) = 3.487, p = .001$, uncorrected). There was no reliable difference in fixation duration between Bus and Grab (Bus: $M = 0.960, SE = 0.007$; Grab: $M = 0.978, SE = 0.007; t(39) = 1.682, p = .101$, uncorrected).

Fixations were slightly longer for WebGazer ($M = 1.001$ s, $SE = 0.002$) than for EyeLink ($M = 0.961$ s, $SE = 0.006$). Furthermore, there was a significant interaction between alternative and eye-tracker ($F(2, 58) = 10.38, p < .001$), suggesting that fixation duration differs across alternatives depending on the eye-tracking method. No other main effects or interactions were found (all $ps > .31$).

Post hoc tests showed that fixation durations differed significantly between metro and bus and between metro and grab with EyeLink, but not with WebGazer. This indicates that fixation duration differences across alternatives were captured reliably by EyeLink but not by the WebGazer (Fig. 7 A–B).

Attributes Finally, we tested whether the duration of fixations on different attributes (comfort, travel, cost) differed and whether this pattern remained stable across display sizes and eye-tracking methods. We found a significant main effect of eye-tracking method ($F(1, 34) = 32.08, p < .001$), reflecting slightly longer fixations detected by Webcam-based eye-tracker. No other significant main effects or interactions were found (all $ps > .24$). These results suggest that participants spent similar amounts of time fixating on each attribute, regardless of the display size, but that the eye-tracking method has an effect on the recorded durations (Fig. 7 C–D).

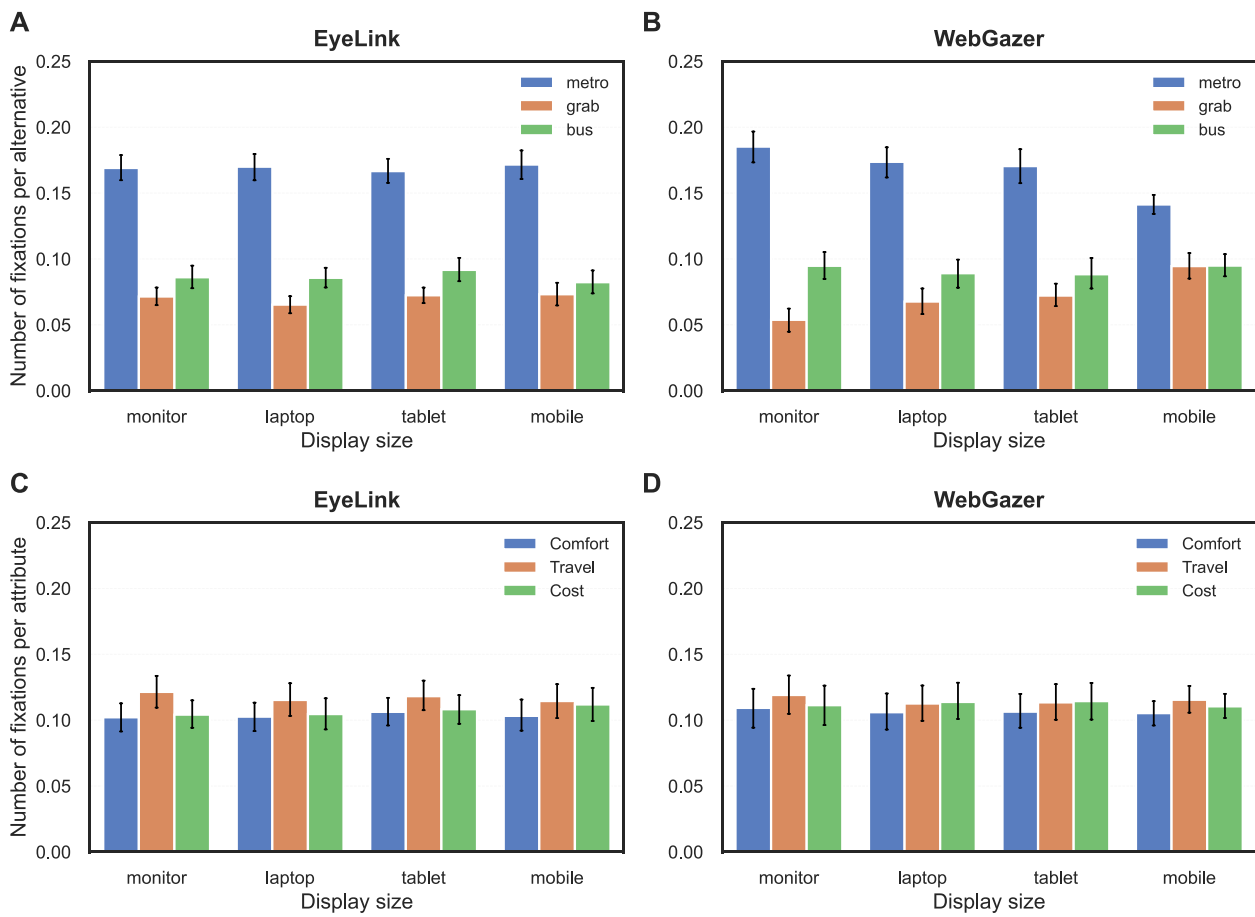


Fig. 6 Fixation counts across alternatives (**A**: EyeLink, **B**: WebGazer) and attributes (**C**: EyeLink, **D**: WebGazer) by display size for Experiment 2 (transport choices). Data are presented as mean values across participants with 95% confidence intervals (based on 1000 bootstrap iterations)

Modeling results

To check the consistency in parameter estimation across the two eye-tracking methods, we compared two measures of willingness to pay (WTP): i) the WTP for relevant attributes (pizza quality in the first experiment, i.e., $-\frac{\tilde{\beta}_{\text{quality}}}{\tilde{\beta}_{\text{cost}}}$, and travel-time reduction in the second experiment, i.e., $\frac{\tilde{\beta}_{\text{time}}}{\tilde{\beta}_{\text{cost}}}$), and ii) the contribution of the latent variable to utility ($\alpha_k I P_n$, in each experiment).

WTP represents the maximum amount an individual is willing to sacrifice (e.g., in monetary terms or time) to obtain a unit improvement in a specific attribute. In the context of our experiments, WTP for pizza quality reflects how much more participants are willing to pay for higher-quality pizza, while WTP for travel-time reduction indicates how much they value saving time. The latter measure is known as the Subjective Value of Travel Time Savings (SVTTS). Both metrics are defined as random variables; therefore, we compare their posterior distributions across individuals (Train,

2009). If the EyeLink and WebGazer data are consistent, the distributions should show a significant degree of overlap between them. Moreover, the parameter estimates should be comparable between the two methods. To this end, we fit the model and compare parameter estimations across the two eye-tracking methods. Provided EyeLink and WebGazer yield comparable estimates, *t* tests should be non-significant when comparing them. Results of WTP in Experiment 1 and SVTTS in Experiment 2 are shown in Singapore dollars (SGD).

Experiment 1: Pizza choices

Table 1 (top) shows the parameter estimates and overlapping degrees between the posterior distributions for QWTP in Experiment 1. As can be seen in the table, the distributions were consistent across all devices; the overlapping degree between them was above 0.5 in all cases, with the highest overlap observed for the largest display size (monitor; 0.9). Kolmogorov–Smirnov (KS) tests confirmed that the distributions were not significantly different (all *ps* > 0.12). The

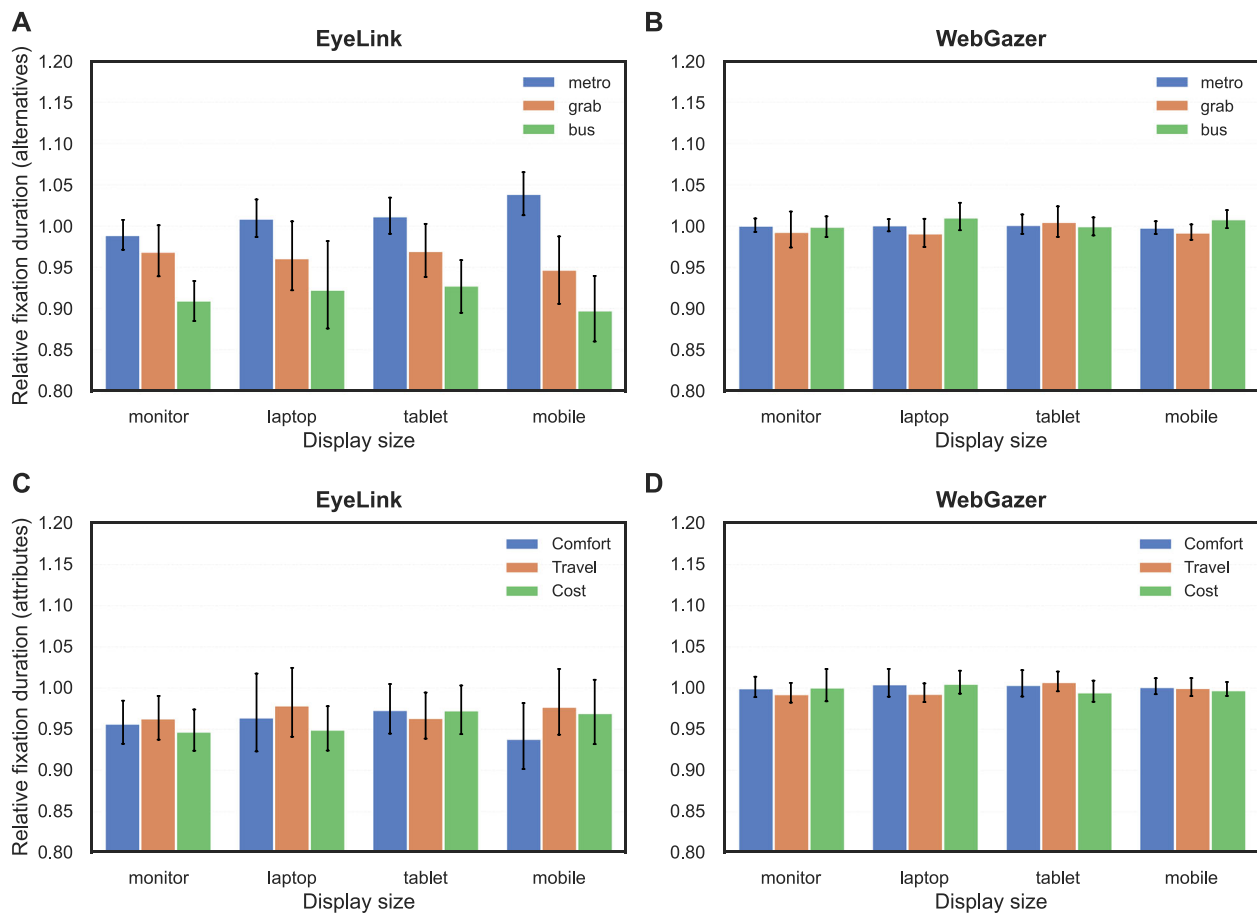


Fig. 7 Fixation durations across alternatives (**A**: EyeLink, **B**: WebGazer) and attributes (**C**: EyeLink, **D**: WebGazer) by display size for Experiment 2 (transport choices). Data are presented as mean values across participants with 95% confidence intervals (based on 1000 bootstrap iterations)

Table 1 Comparison of posterior means and distributions between WebGazer and EyeLink

Device	Mean (95% CI)		<i>t</i>	<i>p</i>	KS	<i>p</i>	Overlap
	WebGazer	EyeLink					
Exp. 1							
QWTP							
Monitor	2.82 (2.55, 3.10)	3.19 (2.50, 3.87)	-0.97	.34	.10	.99	.90
Laptop	3.34 (2.70, 3.97)	2.68 (1.94, 3.42)	1.32	.19	.18	.55	.69
Tablet	2.35 (2.19, 2.51)	2.30 (2.19, 2.41)	0.51	.62	.23	.21	.77
Mobile	2.31 (2.24, 2.37)	2.31 (2.19, 2.43)	-0.08	.93	.26	.12	.59
Exp. 2							
SVTTS							
Monitor	0.50 (0.40, 0.60)	0.46 (0.36, 0.56)	0.53	.60	.17	.71	.94
Laptop	0.47 (0.36, 0.59)	0.69 (0.65, 0.73)	-3.56	.001	.67	<.001	.34
Tablet	0.73 (0.70, 0.76)	0.59 (0.40, 0.78)	1.39	.17	.78	<.001	.18
Mobile	0.78 (0.77, 0.80)	0.47 (0.38, 0.55)	7.51	<.001	.86	<.001	.14

Note. QWTP = quality willingness to pay; SVTTS = subjective value of travel time savings. All *p* values < .05 are considered significant and are shown in bold. KS = Kolmogorov–Smirnov test statistic. Parenthesis indicate 95% confidence intervals

Table 2 Comparison of distribution shapes for the latent variable between WebGazer and EyeLink across devices and attributes

Device	Attribute	KS	<i>p</i>	Overlap
Exp. 1				
Monitor	Cost	0.270	0.107	0.644
Monitor	Quality	0.209	0.302	0.716
Laptop	Cost	0.154	0.752	0.895
Laptop	Quality	0.154	0.752	0.837
Tablet	Cost	0.335	0.019	0.507
Tablet	Quality	0.235	0.198	0.685
Mobile	Cost	0.184	0.545	0.812
Mobile	Quality	0.132	0.903	0.869
Exp. 2				
Monitor	Comfort	0.444	0.001	0.250
Monitor	Cost	0.222	0.340	0.855
Monitor	Time	0.278	0.125	0.726
Laptop	Comfort	0.306	0.069	0.610
Laptop	Cost	0.417	0.003	0.133
Laptop	Time	0.500	<.001	0.254
Tablet	Comfort	0.270	0.135	0.611
Tablet	Cost	0.514	<.001	0.043
Tablet	Time	0.432	0.002	0.425
Mobile	Comfort	0.306	0.069	0.538
Mobile	Cost	0.500	<.001	0.064
Mobile	Time	0.444	0.001	0.313

Note. Bold *p* values indicate statistically significant differences ($p < .05$) in distribution shape (KS test) between WebGazer and EyeLink. Overlap ranges from 0 (no overlap) to 1 (identical distributions)

parameter estimates of QWTP were also consistent, as we found no evidence of a significant difference between the two eye-tracking methods for each display size (all $ps > .19$).

The contribution of the latent variable to utility was highly consistent across display sizes (see Table 2 top). Overlapping values were above 0.64 in all conditions, and no significant differences were found in the distribution shapes. The only exception was the cost attribute in the tablet condition, for which the KS test was significant ($p = .019$), indicating a difference in the distributions between the two eye-tracking methods.

Experiment 2: Transport mode choices

Table 1 (bottom) shows the parameter estimates for SVTTS and the distribution overlaps in Experiment 2. Unlike the simpler scenario in Experiment 1, the more complex design involving three alternatives and three attributes showed consistency between eye-trackers only in the monitor condition, with no significant differences and high distributional overlap (KS test: $p = .71$; overlap: .94) and comparable parameter estimates ($p = .60$).

The contribution of the latent variable to utility was also more consistent in the monitor condition (see Table 2 bottom), with no significant differences in distribution shape for the cost and time attributes (KS tests: $p = .340$ and $p = .125$; overlaps: 0.855 and 0.726, respectively). However, the comfort attribute showed a significant difference ($p = .001$, overlap = 0.25).

For other display sizes, the results were more nuanced. In the laptop, tablet, and mobile conditions, at least one attribute showed significantly different distributions between the two eye-trackers (KS tests: all $p < .05$), with overlap values below 0.5. The cost attribute showed the lowest overlap for tablet (0.043) and mobile (0.064), with significant KS tests in both cases ($p < .001$). Similar patterns were observed for the time attribute. Overlap for the comfort attribute remained moderate across all display sizes (0.25 to 0.61).

Discussion

The use of eye-tracking has gained widespread popularity across multiple disciplines, including psychology, neuroscience, marketing, and economics, due to its ability to provide fine-grained insights into visual attention and cognitive processes in learning and decision-making (Shimojo et al., 2003; Krajbich et al., 2010; Reutskaja et al., 2011; Le Pelley et al., 2015). At the same time, the rise of online experimentation has enabled researchers to collect large-scale behavioral data efficiently and cost-effectively, extending the scope of experimental research beyond traditional laboratory settings (Crump et al., 2013; Rodd, 2024). The present study contributes to this growing literature by systematically comparing the performance of a high-precision laboratory-based eye-tracker (EyeLink 1000 Plus) with a web-based solution (WebGazer) in decision-making tasks conducted across different display size conditions mimicking the electronic devices users can find in the market.

Our first aim was to assess the reliability of EyeLink and WebGazer in capturing fixation counts. In both experiments, we observed significant heterogeneity in fixation counts across alternatives. In Experiment 1, participants consistently fixated more on one option, pizza1, a pattern that held across all devices with the exception of the mobile display when using WebGazer. In Experiment 2, the metro option consistently received the highest number of fixations across all display conditions. The EyeLink system consistently captured stable patterns across all display sizes. WebGazer, on the other hand, performed similarly well across monitor, laptop, and tablet settings but exhibited reduced reliability in the mobile condition.

A notable difference between the two experiments emerged when we analyzed which alternatives attracted more attention.

This suggests that alternative-based visual attention is task-dependent, potentially influenced by prior expectations or inherent preferences related to the choice domain. However, it is also possible that the position of alternatives on the screen played a role in driving attention. For instance, in left-to-right reading cultures, alternatives placed on the left side of the screen may naturally attract more attention, independent of their intrinsic attributes.

Regarding attribute attention, both methods showed no differences in attention in Experiment 1. In Experiment 2, EyeLink showed an effect in favor of travel time, an effect that was not present on WebGazer, suggesting attribute-based attention may not be consistently captured across eye-tracking methods. By contrast, participants' attention to specific attributes did not vary systematically across display conditions or eye-tracking methods. This is in line with previous studies indicating that attribute-based attention is generally less sensitive to device and display differences compared to alternative-based attention. One possibility is that participants relied on compensatory decision strategies (McFadden, 1972), distributing their attention evenly across attributes rather than focusing on a subset of dominant attributes.

Finally, we examined fixation durations across alternatives and attributes. While some significant differences in fixation durations were observed – particularly for alternatives with EyeLink on larger displays – these effects were generally weaker and less consistent than for fixation counts, and were largely absent with WebGazer. This suggests that the time spent fixating on each alternative or attribute was relatively stable across display conditions. Importantly, fixation durations were longest for metro compared to the other alternatives, a trend that mirrored the fixation count results. This indicates that alternatives receiving more fixations also tend to receive longer fixation durations, reinforcing the view that fixation count is a stronger indicator of decision salience than fixation duration. By contrast, fixation durations did not vary across attributes, suggesting that both methods may lack sensitivity to detect subtle differences in this measure. In web-based settings, noise from device characteristics and participant movement may also obscure small temporal variations. Indeed, prior work shows that eye trackers with lower spatial precision, similar to WebGazer, can sometimes wrongly assign fixations when ROIs are small or closely spaced. Combined with the coarse temporal resolution of webcam sampling, which tends to merge short fixations, these factors likely contribute to the weaker performance of WebGazer relative to EyeLink in some conditions (Dalmaijer, 2014).

Overall, the findings from Experiment 1 in which there is a simple 2 (alternatives) x 2 (attributes) choice presented, support the conclusion that webcam-based eye-tracking can

reliably capture alternative-based visual attention patterns in online decision-making studies conducted on monitor, laptop, and tablet devices, extending the utility of web-based eye-tracking to complex decision environments where attention to both alternative and attribute dimensions plays a crucial role in preference formation. However, the decrease in reliability in the mobile setting suggests caution in using WebGazer for studies where fine-grained gaze precision is required on small screens. For the more complex design of Experiment 2 in which there are three alternatives and three attributes presented, WebGazer is only able to capture alternative-based attention in the largest display (monitor).

Our second goal was to assess the potential of web-based eye-tracking to model decision-making processes. Here, we employed two canonical attention-based models and tested whether the estimated parameters differed significantly across experimental contexts and devices. Our results indicate that WebGazer is generally reliable for modeling behavior. In Experiment 1, the distributions of willingness to pay (WTP) for a key attribute were consistent across all devices, with overlapping degrees above 0.5 and no significant differences according to Kolmogorov-Smirnov tests ($p > 0.1$). Similarly, the contribution of the latent variable ($\alpha_k I P_n$) to utility was highly consistent for large-screen devices (monitor, laptop, and tablet).

In the more complex experiment (Experiment 2), the results were more nuanced. While the distributions of the SVTTS were consistent for the monitor, smaller devices (laptop, tablet, and mobile) showed significantly lower consistency, with a maximum overlapping degree of .34. This discrepancy is likely due to the smaller areas of interest in the more complex experimental design, which reduces the accuracy of WebGazer. However, the latent variable's contribution to utility remained consistent only for the monitor condition, suggesting that WebGazer can still reliably model attention-based decision processes in complex tasks when used with larger screens. A summary of all our results is presented in Table 3.

Applications of webcam-based eye tracking in decision-making contexts have been scarce. A notable exception is a study by Yang and Krajbich (2021), which employed WebGazer in a simpler two-alternative forced-choice task, which goal was to be compatible with a particular type of theory of decision making: the attentional drift-diffusion model (Krajbich et al., 2010). Unlike this prior work, our study systematically evaluates WebGazer across multiple devices in the context of a well established, more complex decision-making paradigm, offering a more comprehensive assessment of its reliability for modeling choice behavior.

Most research comparing the two eye-tracking methods has focused on perceptual and cognitive tasks, and has

reached similar conclusions: with coarse ROIs and simple displays, webcam data is able to reproduce key effects, although with some loss of precision and smaller effect sizes (Hutt et al., 2023; Slim et al., 2024; Steffan et al., 2024). Our main contribution in this paper is to show a similar pattern in economic decision-making, explicitly manipulating display size to mimic eye tracking experiences in different devices and task complexity. Our results are consistent with this previous literature, showing that WebGazer data are reliable on larger screens and simpler tasks (few attributes/alternatives), and that performance is adversely affected as smaller screens and more complex tasks are employed. Critically, we show this to be the case not only at the level of raw gaze metrics, but at the model level: inferences about WTP/SVTTS estimates are also affected by these factors, affecting conclusions about the latent processes driving behavior in decision-making tasks. Taken together, these results provide boundary conditions and concrete guidelines for when WebGazer could be successfully applied in DCEs, and when researchers should expect weaker effects and/or less reliable inferences about decision-making processes.

In summary (Table 3), this paper provides the first systematic evidence of the merits of web-based eye-tracking for decision-making research. Fixation counts and durations, the cardinal markers of attention, were largely consistent across EyeLink and WebGazer, with the exception of small display sizes mimicking mobile devices. Model fitting was also consistent, with WebGazer reliably capturing both preference and attention-based heterogeneity. Our findings demonstrate that WebGazer holds significant promise for testing decision-making processes online with large sample sizes, opening new grounds for studying individual differences in attention and preference formation. However, our results also emphasize the importance of considering device-specific constraints when designing online experiments, particularly for mobile devices and smaller tablet models, which may introduce additional noise in gaze data. Future studies should explore whether methodological refinements – such as increased calibration procedures, individualized gaze correction, or more sophisticated gaze-processing algorithms – could enhance the sensitivity of webcam-based eye-tracking for capturing fine-grained attention patterns.

Table 3 Convergence/divergence between EyeLink (EL) and WebGazer (WG) across tasks and displays

Outcome	Simple task (Exp. 1: 2 alts × 2 atts)	Complex task (Exp. 2: 3 alts × 3 atts)
Alternative fixation counts	EL stable across displays; WG matches on monitor/laptop/tablet, weaker on mobile	Same direction; WG weakest on mobile; EL patterns stable across displays
Attribute fixation counts	No attribute differences (both methods)	Attribute effect present; no method or display interactions (both methods capture it)
Alternative fixation durations	<i>Method effect only</i> (EL ≠ WG); no display effects	<i>Method + alternative + method × alternative</i> ; EL: metro > bus/grab; WG: no differences; no display effects
Attribute fixation durations	<i>Method effect only</i> ; no attribute or display effects	<i>Method effect only</i> ; no attribute or display effects
Model: QWTP (Exp. 1)	Converges across all devices (non-sig tests; overlaps ≥ .59)	—
Model: SVTTS (Exp. 2)	—	Converges on <i>monitor</i> (high overlap); <i>laptop/tablet/mobile</i> diverge (lower overlaps; several tests significant)

Notes. WG = WebGazer; EL = EyeLink. “Converges” denotes practical equivalence/high distribution overlap of parameter estimates. Displays approximate monitor, laptop, tablet, and mobile conditions

Appendix

Fixation distributions for experiments and display sizes

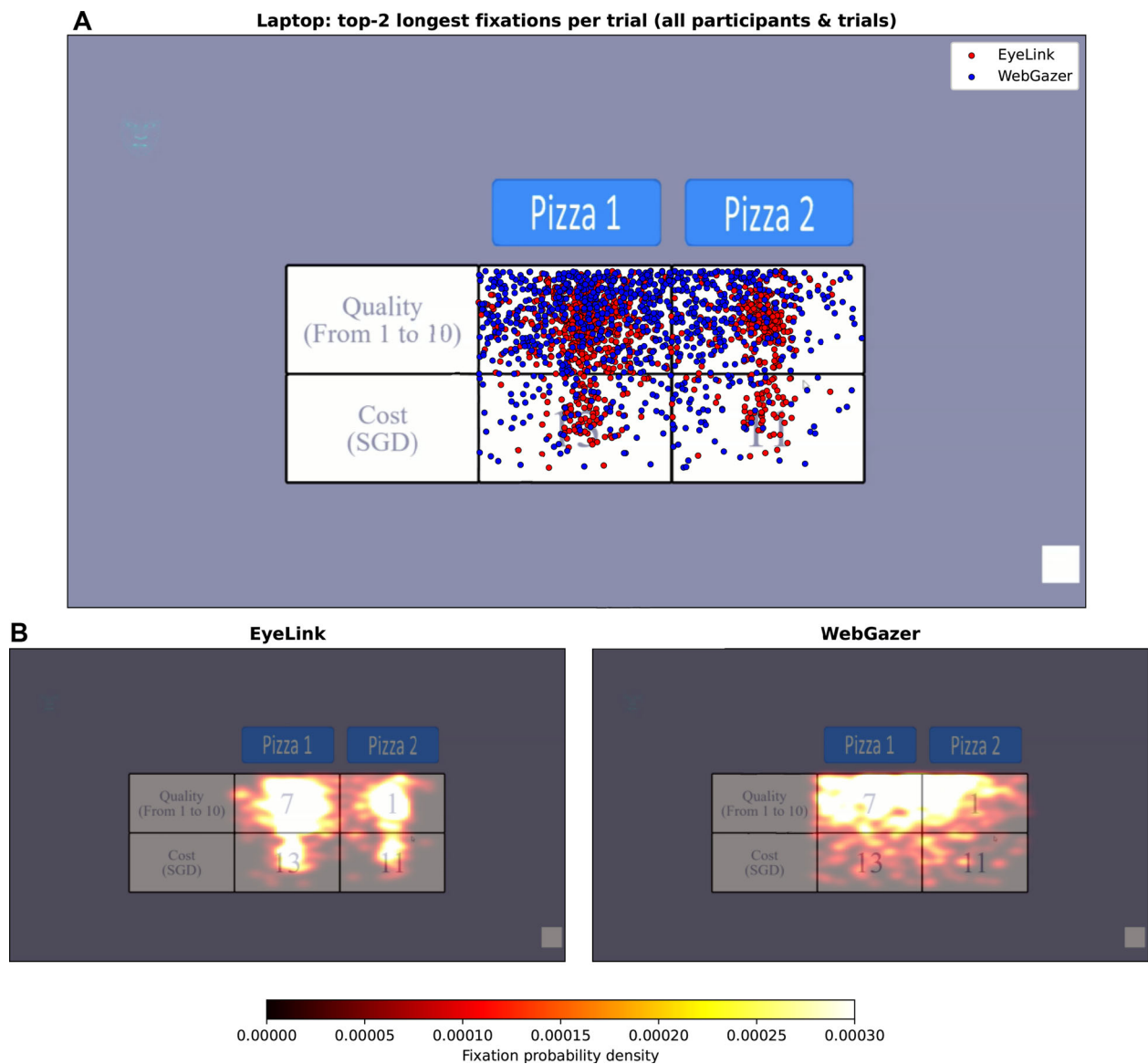


Fig. 8 Fixation distributions for Experiment 1 (pizza choices), laptop display condition: **A**. Two longest fixations per trial (centroids), pooled across all participants; red = EyeLink, blue = WebGazer. **B**. Fixation heatmaps shown separately for EyeLink and WebGazer. Brighter colors indicate higher fixation probability. Rectangles mark ROIs (alternative and attribute cells). Numbers inside cells are example attribute levels from a representative screen

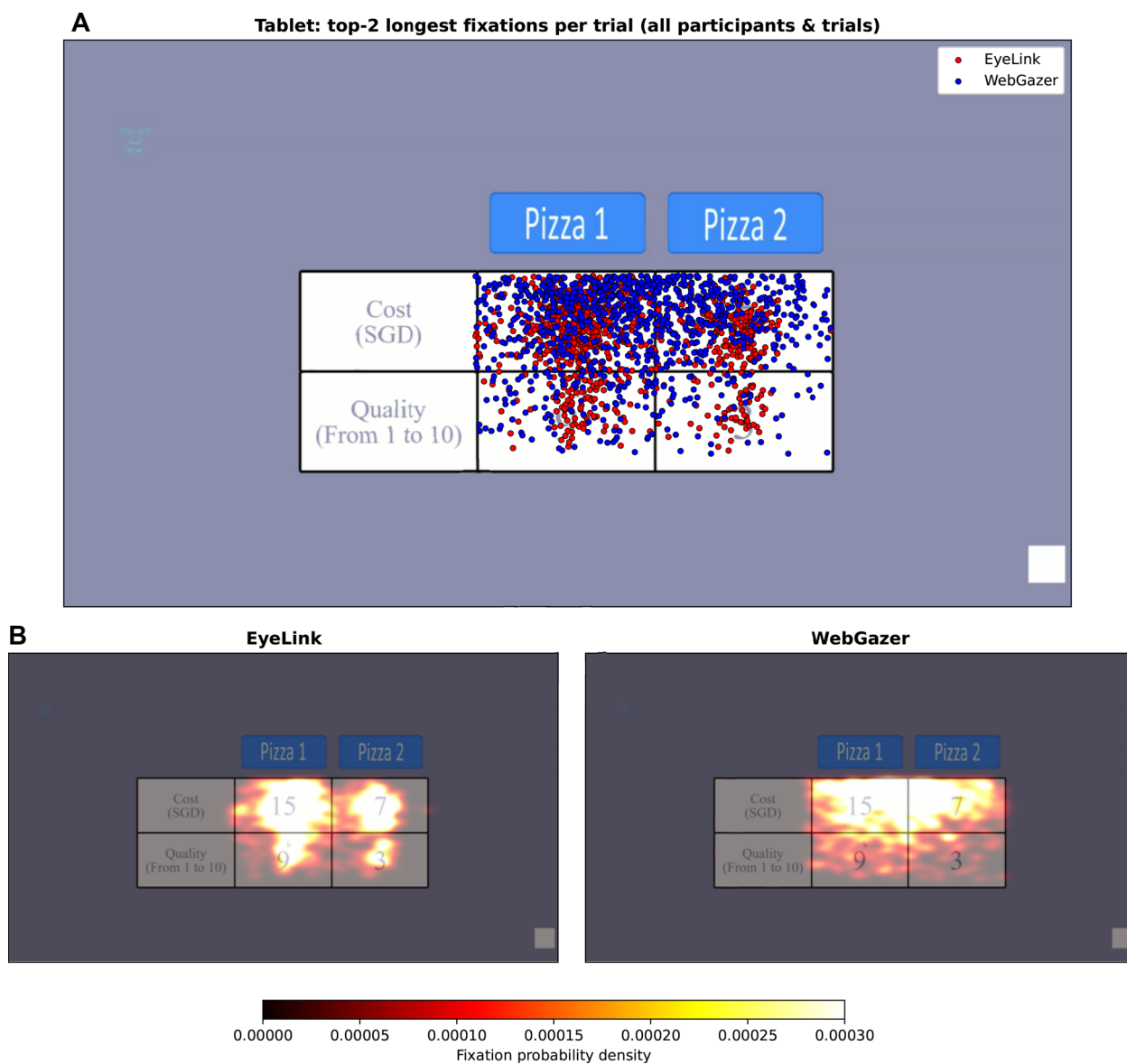


Fig. 9 Fixation distributions for Experiment 1 (pizza choices), tablet display condition. **A.** Two longest fixations per trial (centroids), pooled across all participants; red = EyeLink, blue = WebGazer. **B.** Fixation heatmaps shown separately for EyeLink and WebGazer. Brighter colors

indicate higher fixation probability. Rectangles mark ROIs (alternative and attribute cells). Numbers inside cells are example attribute levels from a representative screen

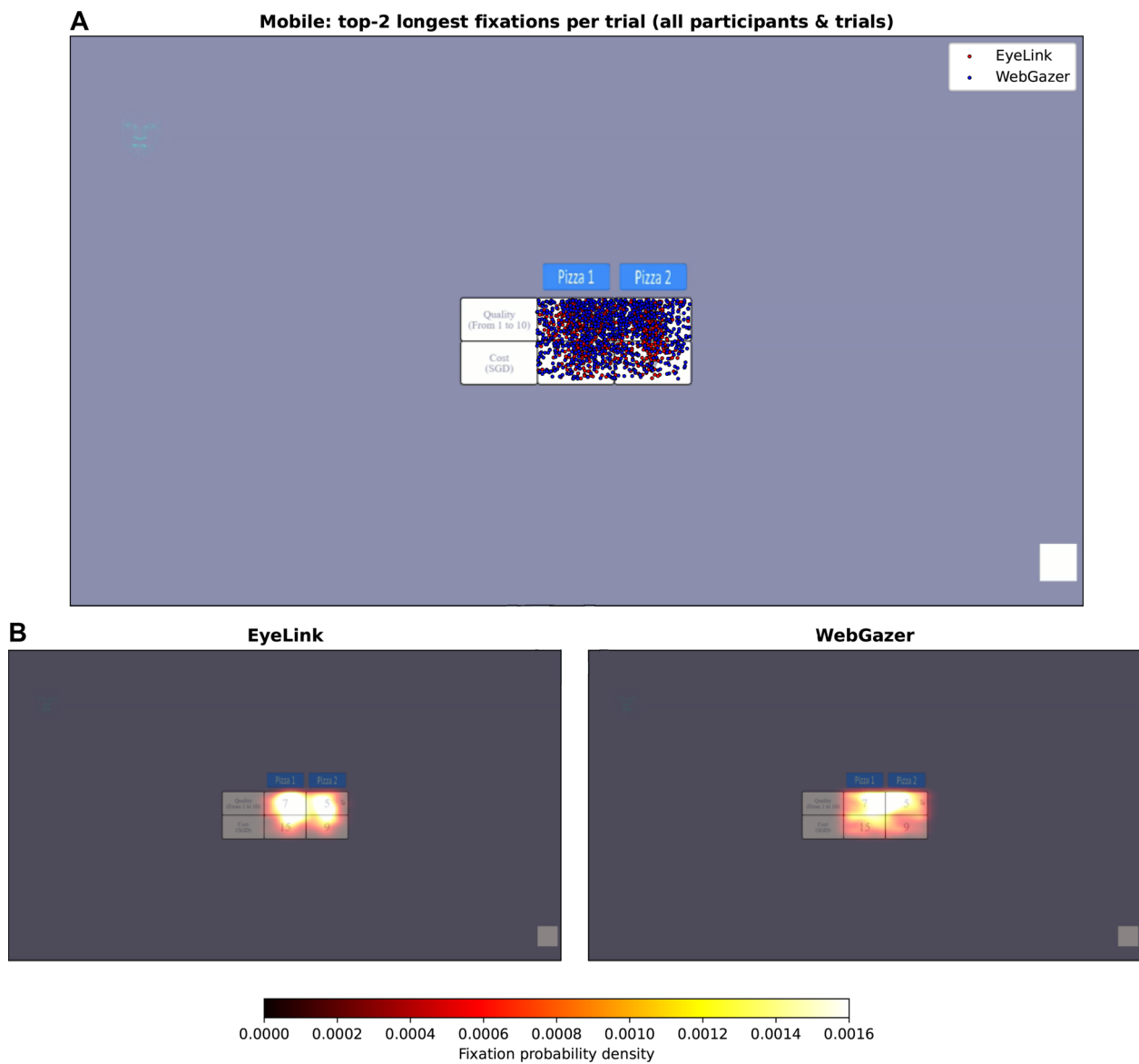


Fig. 10 Fixation distributions for Experiment 1 (pizza choices), mobile display condition: **A**. Two longest fixations per trial (centroids), pooled across all participants; red = EyeLink, blue = WebGazer. **B**. Fixation heatmaps shown separately for EyeLink and WebGazer. Brighter colors

indicate higher fixation probability. Rectangles mark ROIs (alternative and attribute cells). Numbers inside cells are example attribute levels from a representative screen

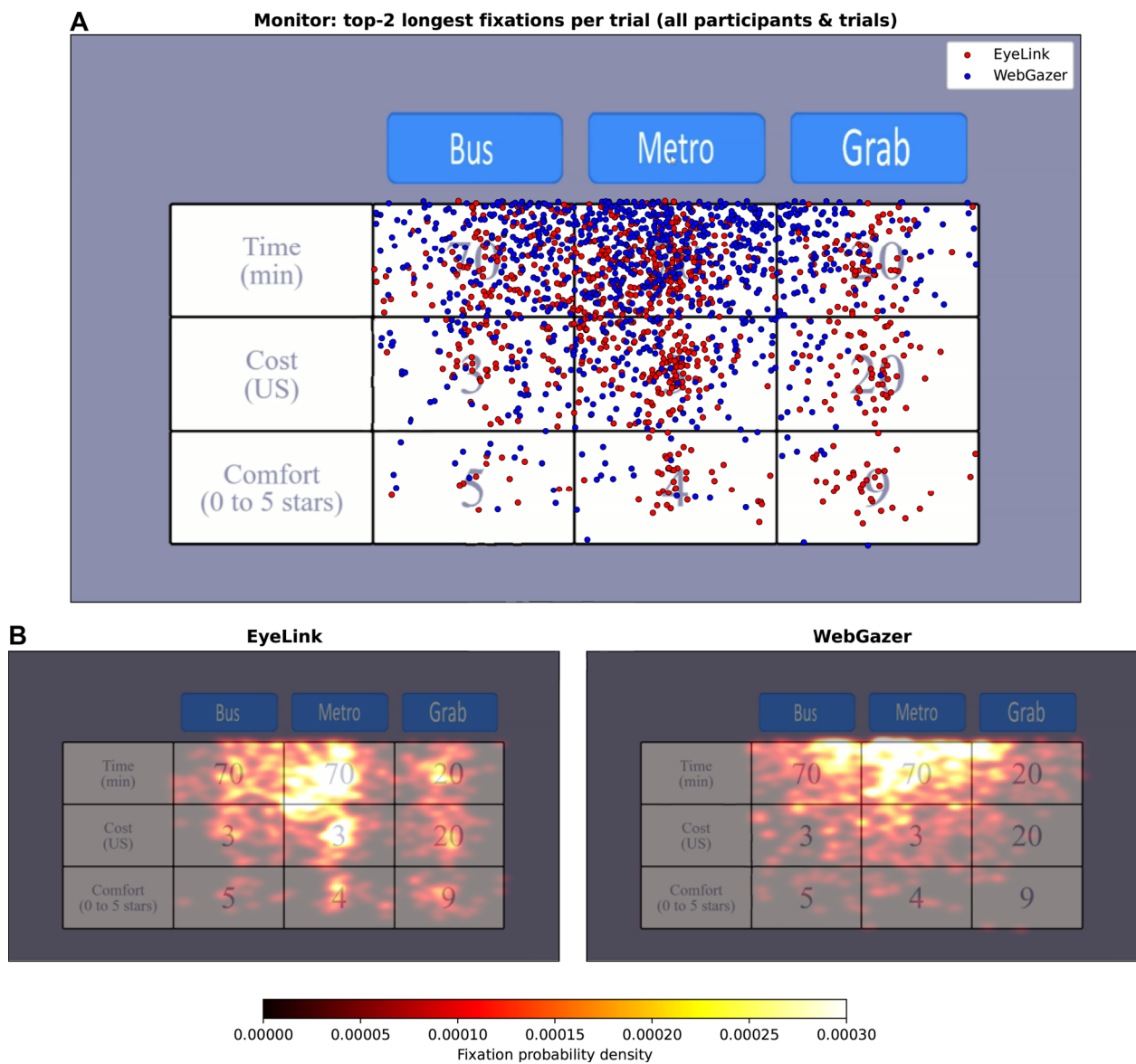


Fig. 11 Fixation distributions for Experiment 2 (travel mode choices), monitor display condition: **A**. Two longest fixations per trial (centroids), pooled across all participants; red = EyeLink, blue = WebGazer. **B**. Fixation heatmaps shown separately for EyeLink and WebGazer. Brighter

colors indicate higher fixation probability. Rectangles mark ROIs (alternative and attribute cells). Numbers inside cells are example attribute levels from a representative screen

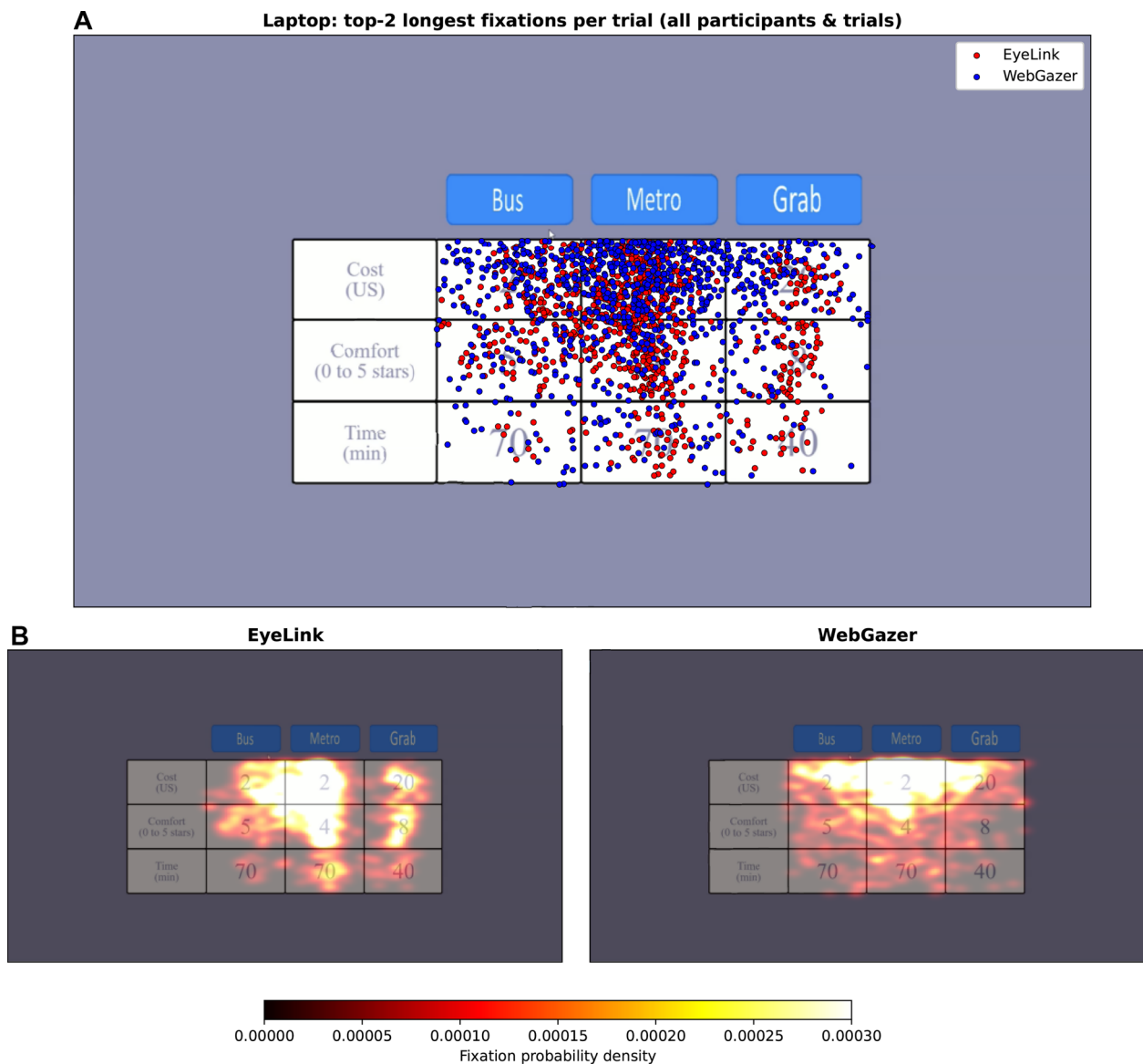


Fig. 12 Fixation distributions for Experiment 2 (travel mode choices), laptop display condition: **A**. Two longest fixations per trial (centroids), pooled across all participants; *red* = EyeLink, *blue* = WebGazer. **B**. Fixation heatmaps shown separately for EyeLink and WebGazer. *Brighter*

colors indicate higher fixation probability. *Rectangles* mark ROIs (alternative and attribute cells). *Numbers inside cells* are example attribute levels from a representative screen

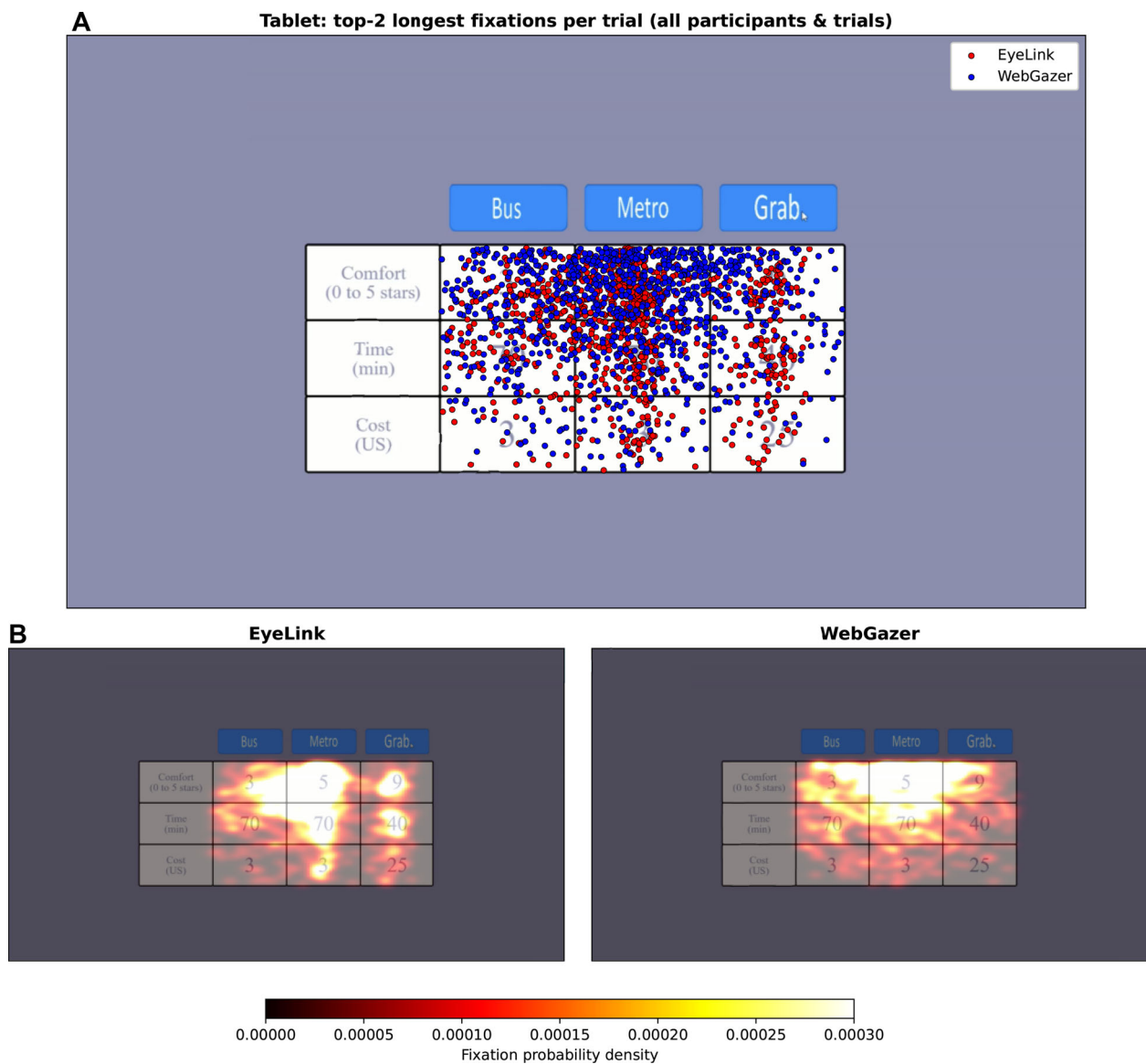


Fig. 13 Fixation distributions for Experiment 2 (travel mode choices), tablet display condition: **A**. Two longest fixations per trial (centroids), pooled across all participants; red = EyeLink, blue = WebGazer. **B**. Fixation heatmaps shown separately for EyeLink and WebGazer. Brighter

colors indicate higher fixation probability. Rectangles mark ROIs (alternative and attribute cells). Numbers inside cells are example attribute levels from a representative screen

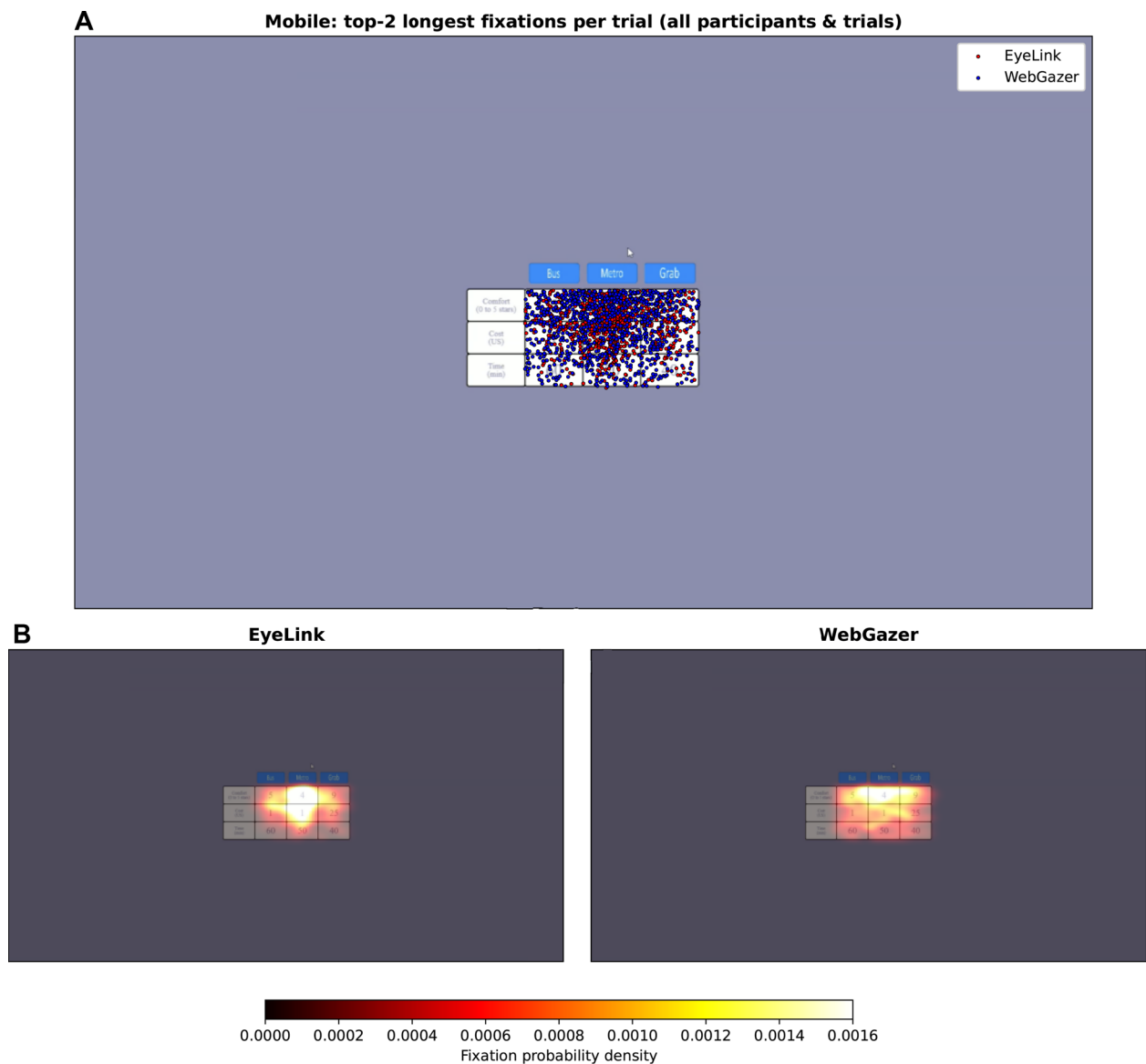


Fig. 14 Fixation distributions for Experiment 2 (travel mode choices), mobile display condition: **A**. Two longest fixations per trial (centroids), pooled across all participants; *red* = EyeLink, *blue* = WebGazer. **B**. Fixation heatmaps shown separately for EyeLink and WebGazer. *Brighter colors* indicate higher fixation probability. *Rectangles* mark ROIs (alternative and attribute cells). *Numbers inside cells* are example attribute levels from a representative screen

Technical details of the latent information processing (LIP) model

The LIP model extends the integrated choice and latent variable (ICLV) framework by explicitly linking decision weights to information processing. Below, we present the full derivation of the model and estimation approach.

Utility specification

The standard Random Utility Maximization (RUM) model assumes that the utility of alternative j for individual n is

given by:

$$U_{nj} = \sum_{k=1}^K \beta_k x_{njk} + \varepsilon_{nj}. \quad (6)$$

In contrast, the LIP model modifies this specification by incorporating a latent attention variable:

$$U_{nj} = \sum_{k=1}^K \tilde{\beta}_k x_{njk} + \varepsilon_{nj}, \quad (7)$$

where the adjusted parameter $\tilde{\beta}_k$ is defined as:

$$\tilde{\beta}_k = \beta_k + \alpha_k I P_n. \quad (8)$$

The latent information processing variable IP_n follows a normal distribution:

$$IP_n \sim N(\gamma_{IP}, \sigma_{IP}). \quad (9)$$

Measurement equations for visual attention

The latent attention variable IP_n is inferred from observed visual attention data, denoted as VA_{nk} . The relationship between IP_n and visual attention indicators is formulated through measurement equations:

$$VA_{nk} = \gamma_0 + \gamma_k IP_n + \omega_{nk}, \quad (10)$$

where $\omega_{nk} \sim N(0, \sigma_k)$ represents measurement error.

Choice probabilities

The probability of choosing alternative i , conditional on IP_n , is given by:

$$P_{ni}(x_{ni}, \beta | \eta_{ni}) = \frac{\exp(\sum_{k=1}^K \beta_k x_{nik} + \alpha_k IP_n)}{\sum_{j \in C} \exp(\sum_{k=1}^K \beta_k x_{nj k} + \alpha_k IP_n)}. \quad (11)$$

Likelihood function

The likelihood function integrates over the latent variable:

$$L = \prod_n \int_{\eta} \left(\prod_k f_{nk}(\sigma_k, \gamma | \eta_{nk}) \prod_j P_{nj}(x_{nj}, \beta | \eta_{nk})^{y_{nj}} \right) d\eta. \quad (12)$$

The log-likelihood is estimated using simulation-based methods, taking R draws from a normal distribution:

$$SLL = \sum_n \ln \sum_{r=1}^R \frac{1}{R} \left(\prod_k f_{nk}(\sigma_k, \gamma | \eta_{nk}^r) \prod_j P_{nj}(x_{nj}, \beta | \eta_{nk}^r)^{y_{nj}} \right). \quad (13)$$

Posterior distributions for WTP and SVTTS

Experiment 1 (Pizza choices)

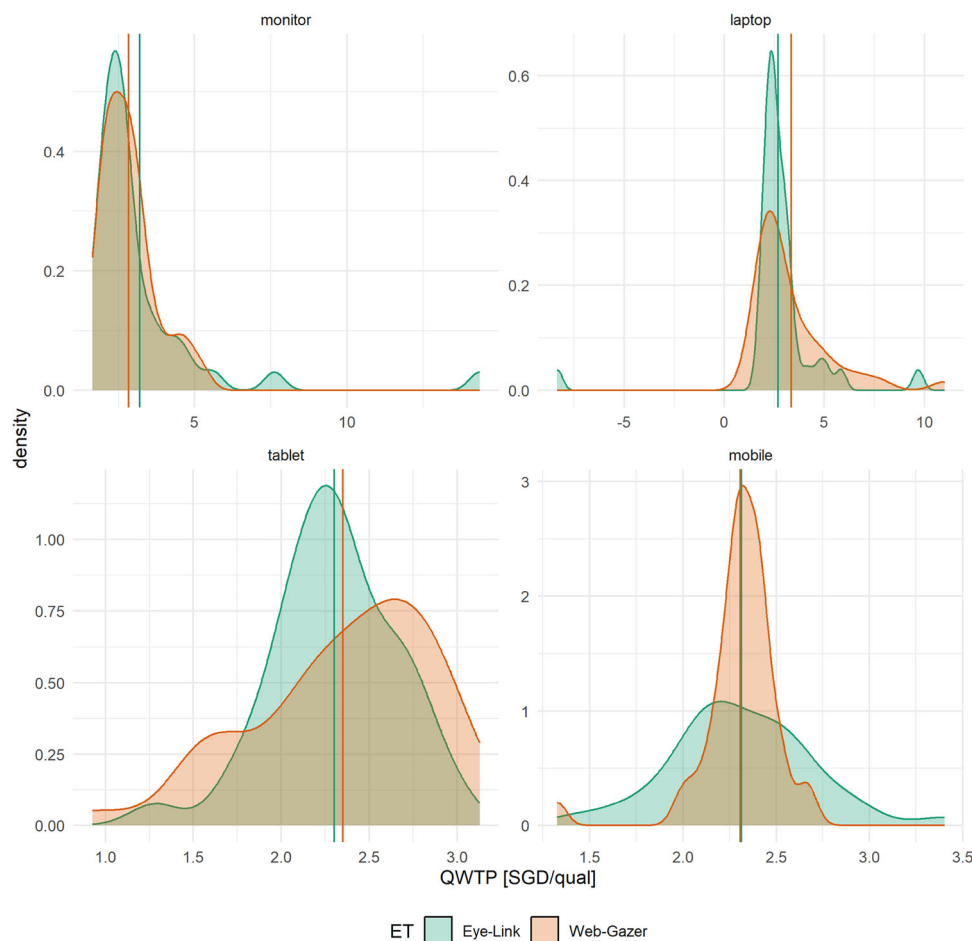


Fig. 15 Posterior distributions of QWTP, grouped by device, for Experiment 1

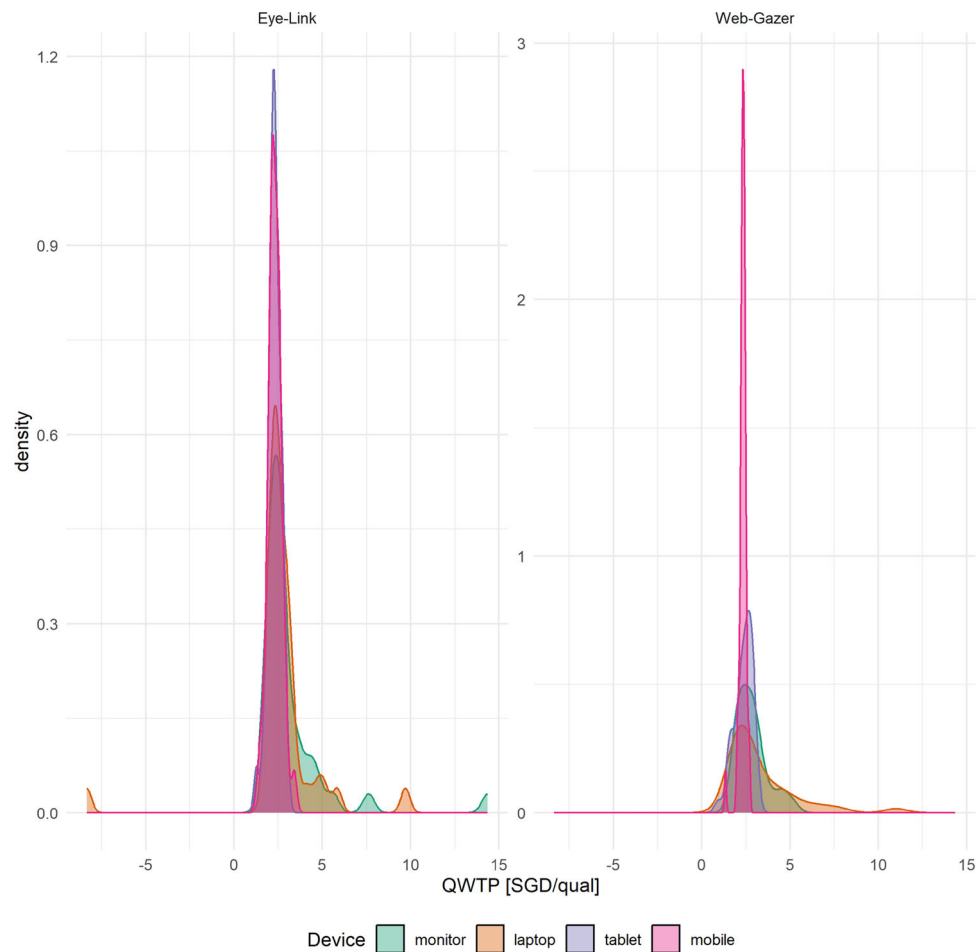


Fig. 16 Posterior distributions of QWTP, grouped by eye-tracking method, for Experiment 1

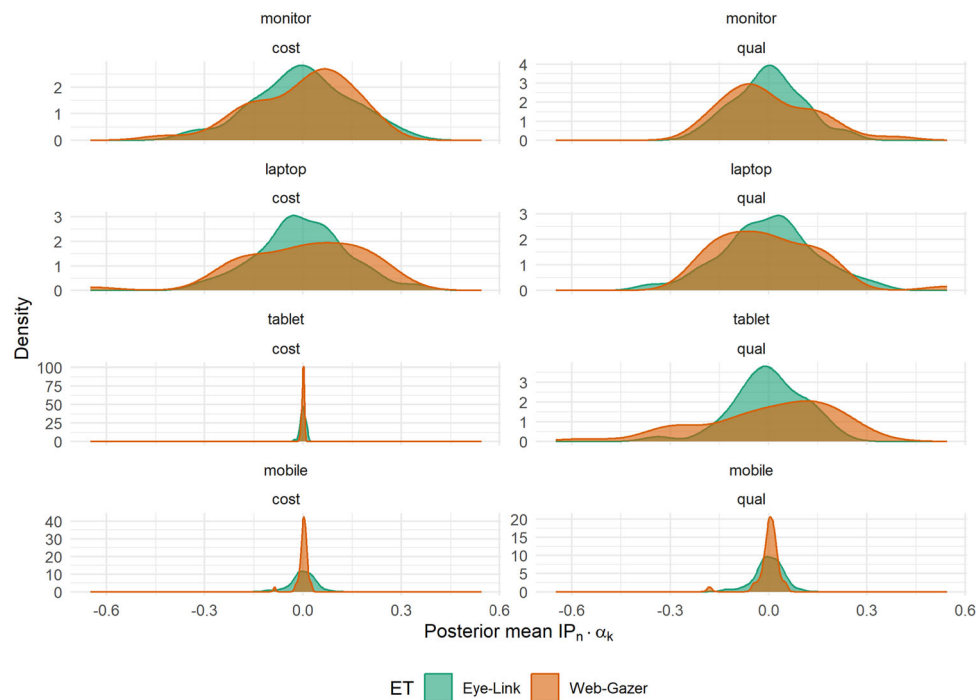


Fig. 17 Posterior distributions of $\alpha_k I P_n$, grouped by display size for Experiment 1

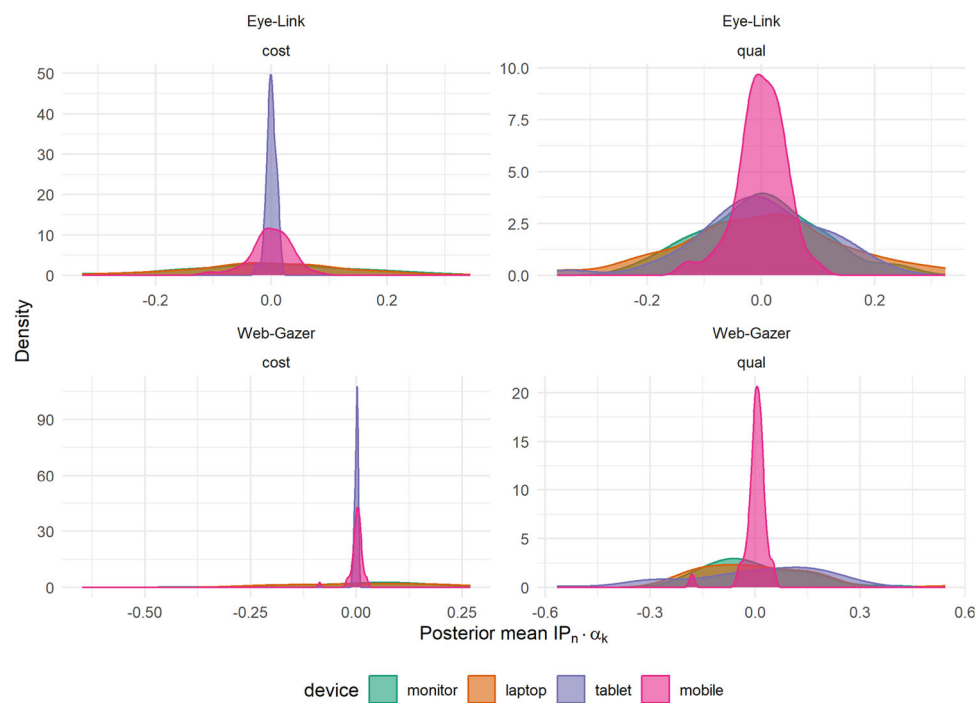


Fig. 18 Posterior distributions of $\alpha_k IP_n$, grouped by eye-tracker, for Experiment 1

Experiment 2 (Transport choices)

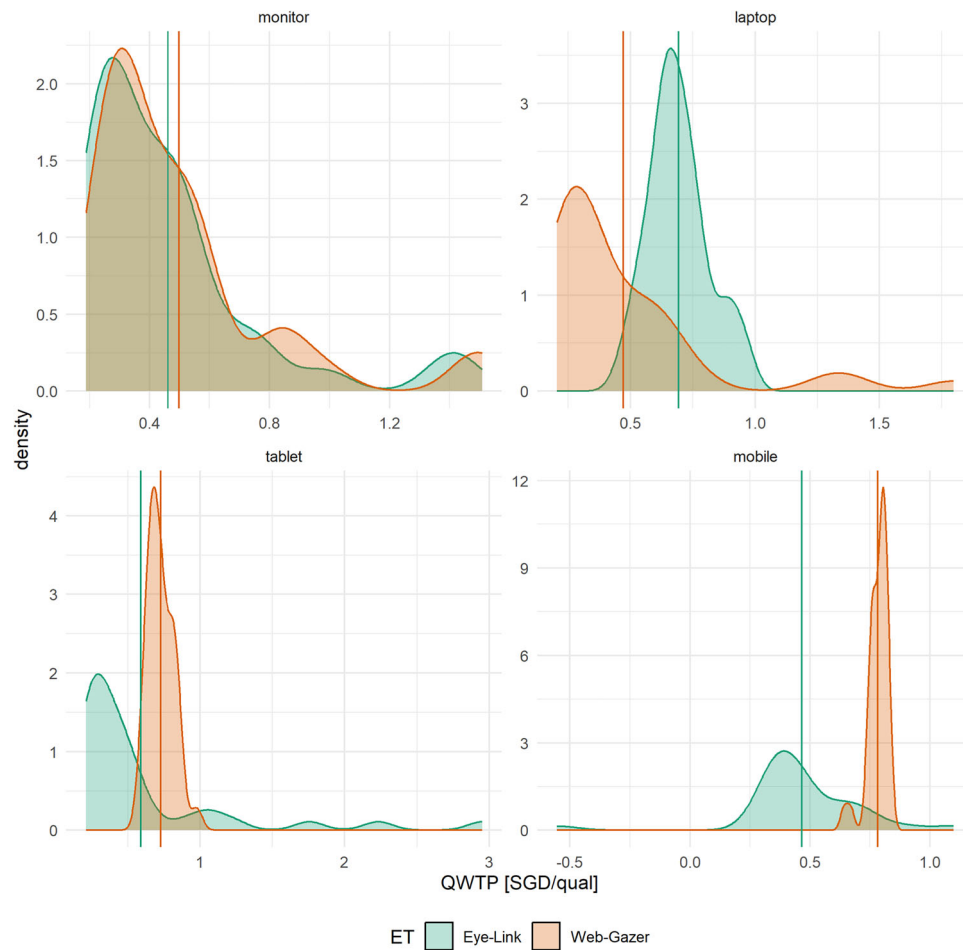


Fig. 19 Posterior distributions of SVTTS, grouped by device, for Experiment 2

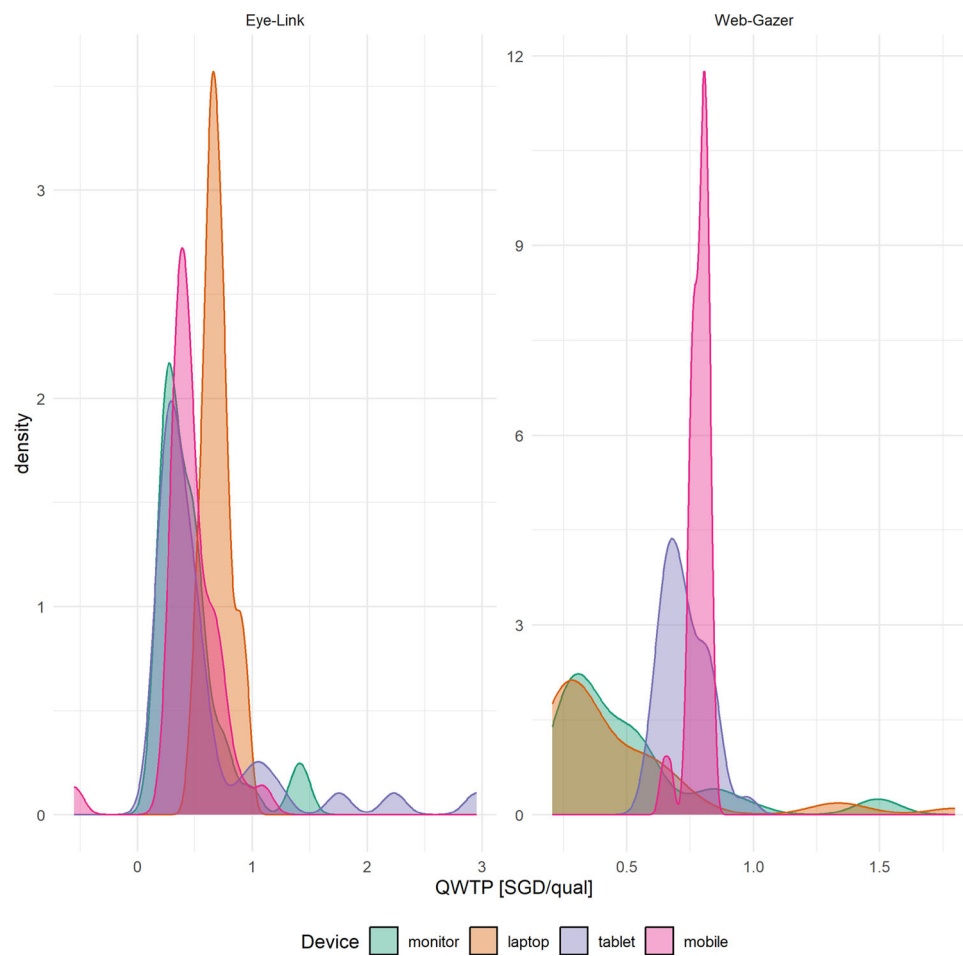


Fig. 20 Posterior distributions of SVTTS, grouped by eye-tracker, for Experiment 2

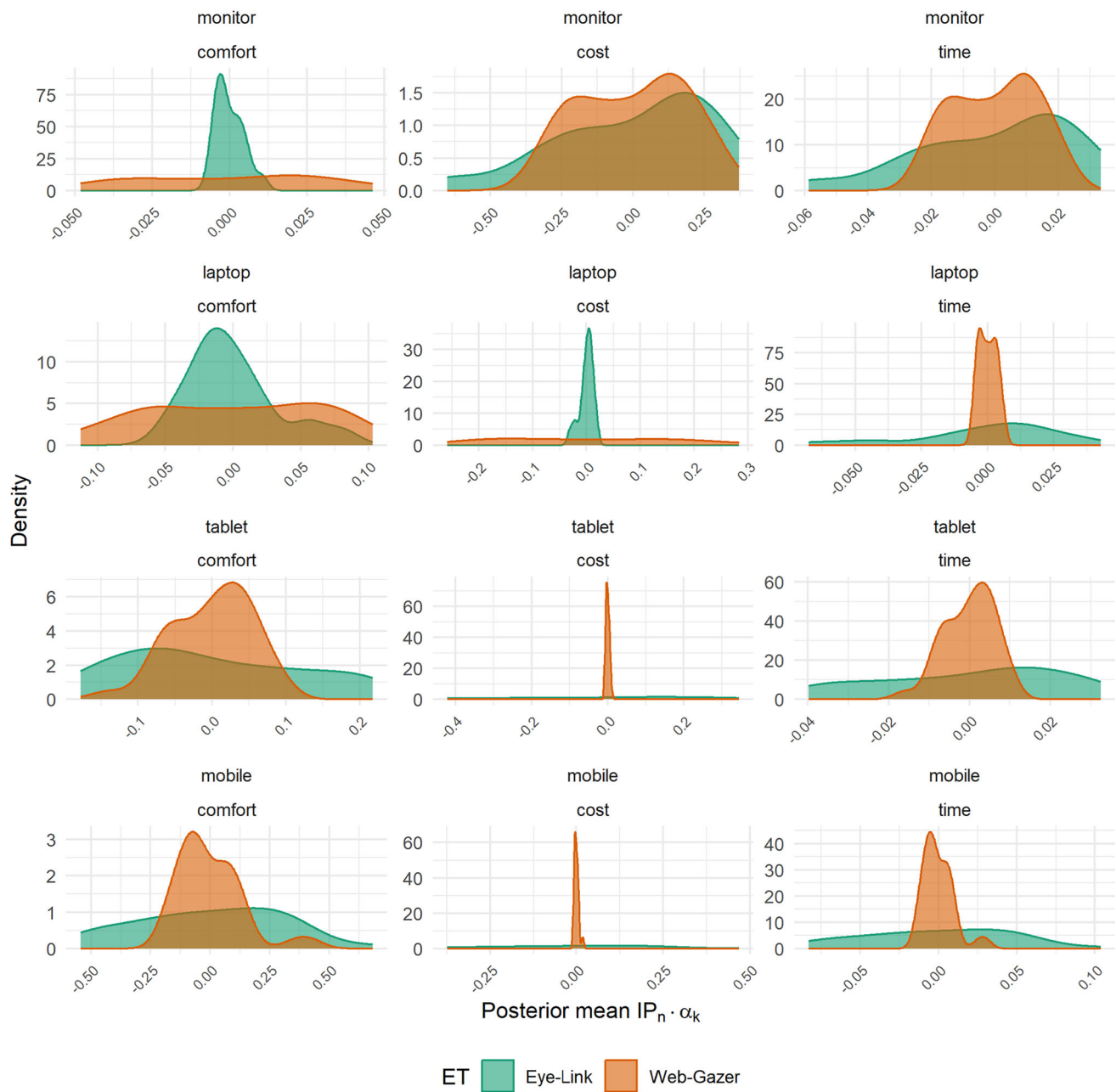


Fig. 21 Posterior distributions of $\alpha_k IP_n$, grouped by device, for Experiment 2

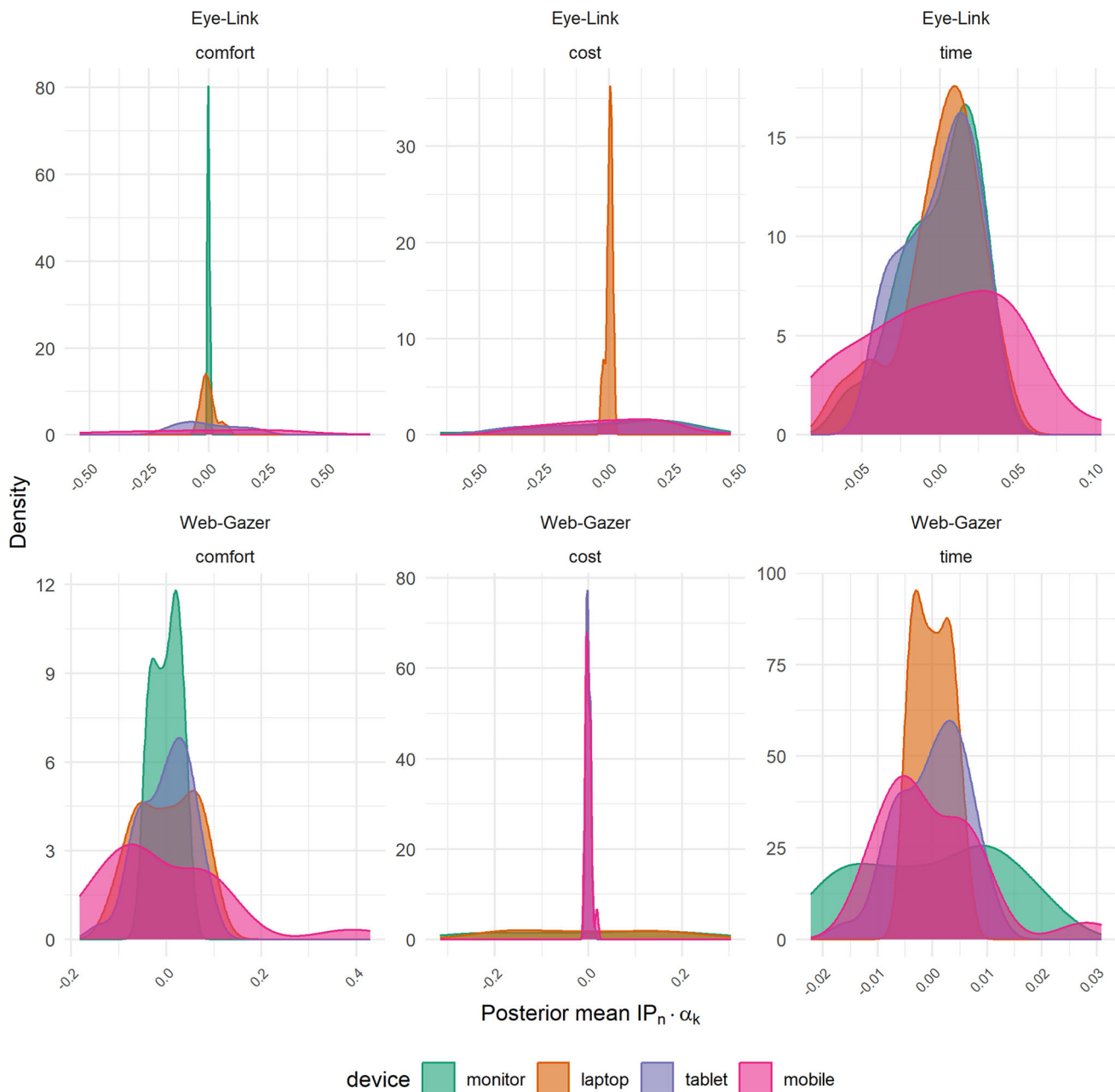


Fig. 22 Posterior distributions of $\alpha_k IP_n$, grouped by eye-tracker, for Experiment 2

Screen and choice display dimensions

Our interest was to manipulate display size to test how each eye-tracking method would generate records of fixations, fixation durations, and parameter estimation across different devices (monitor, laptop, mobile, tablet). Since screen size and viewing distance vary across devices, we ensured that the choice task presentation remained comparable across conditions. Table 4 summarizes the typical viewing distances and screen dimensions for each device.

To maintain consistency across devices, only 80% of the screen was allocated for choice presentation. The width and height of the choice area were calculated accordingly. Given a scenario with M alternatives and N attributes, the width of each choice cell was determined as:

$$\text{Cell width} = \frac{\text{Usable Display Width}}{M + 1}$$

Table 4 Screen sizes and viewing distances for different display conditions

Device	Viewing distance (cm)	Screen size (cm)	Usable display (deg)
Mobile	25–30	14 × 8	11.2 × 6.4
Tablet	30–45	22 × 12	25.6 × 14.4
Laptop	50–60	34 × 19	27.2 × 15.2
Monitor	50–76	56 × 32	37.6 × 22.4

Table 5 Choice cell dimensions across different devices

Device	Alternatives	Attributes	Cell width (deg)	Cell height (deg)
Mobile	2	2	3.7	2.2
Tablet	2	2	8.5	4.8
Laptop	2	2	9.2	5.1
Monitor	2	2	12.5	7.5
Mobile	3	3	2.8	1.6
Tablet	3	3	6.4	3.6
Laptop	3	3	6.8	3.8
Monitor	3	3	9.4	5.6

Similarly, the height of each cell was calculated as:

$$\text{Cell height} = \frac{\text{Usable Display Height}}{N + 1}$$

Table 5 presents the computed cell sizes for the different choice tasks. These calculations ensure that the choice matrices are comparable across devices while accounting for differences in screen size and viewing distance.

Funding This research was partially funded by ANID-PIA/PUENTE AFB220003, ANID-SIA 85220023, ANID-FONDECYT 1231027, the MoE Tier 1 Grant (A-8002145-00-00), and the Presidential Young Professorship Grant at the National University of Singapore.

Data Availability Data and materials are publicly available at: <http://osf.io/aeuwx/>

Code availability The analysis code used for this study is available at: <http://osf.io/aeuwx/>

Declarations

Conflicts of Interest/Competing Interests The authors declare that they have no conflicts of interest or competing interests.

Ethics approval All procedures performed in this study were in accordance with the ethical standards of the Institutional Review Board of the National University of Singapore (NUS-IRB-2023-1055) and with the 1964 Helsinki Declaration and its later amendments.

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication All authors have given their consent for publication.

Open Practices Statement The data and analysis scripts for this paper are publicly available at <http://osf.io/aeuwx/>.

References

- Bansal, P., Kim, E.-J., & Ozdemir, S. (2024). Discrete choice experiments with eye-tracking: How far we have come and ways forward. *Journal of Choice Modelling*, 51, Article 100478.
- Bliemer, M. C., & Rose, J. M. (2024). *Designing and conducting stated choice experiments* (pp. 172–205). Handbook of choice modelling. Cheltenham: Edward Elgar Publishing.
- Borozan, M., Loreta, C., & Riccardo, P. (2022). Eye-tracking for the study of financial decision-making: A systematic review of the literature. *Journal of Behavioral and Experimental Finance*, 35, Article 100702.
- Cavanagh, J. F., et al. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, 143, 1476–1488.
- Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *Plos One*, 8, Article e57410.
- Dalmaier E. (2014). Is the low-cost EyeTribe eye tracker any good for research? *PeerJ PrePrints* 2, e585v1. <https://doi.org/10.7287/peerj.preprints.585v1>
- Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, 3(Suppl), 1218–1223.
- Hess, S., & Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of Choice Modelling*, 32, Article 100170.
- Hutt, S., et al. (2023). Webcam-based eye tracking to detect mind wandering and comprehension errors. en. *Behavior Research Methods*, 56, 1–17.

- Krajovich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, 13, 1292–1298.
- Krucien, N., Ryan, M., & Hermens, F. (2017). Visual attention in multi-attributes choices: What can eye-tracking tell us? *Journal of Economic Behavior & Organization*, 135, 251–267.
- Le Pelley, M. E., et al. (2015). When goals conflict with values: Counter-productive attentional and oculomotor capture by reward-related stimuli. *Journal of Experimental Psychology: General*, 144, 158–171.
- Louviere, J. J., & Hensher, D. A. (1982). On the design and analysis of simulated choice or allocation experiments in travel choice modelling. *Transportation Research Record*, 890, 11–17.
- Martinovici, A., Pieters, R., & Erdem, T. (2023). Attention Trajectories Capture Utility Accumulation and Predict Brand Choice en. *Journal of Marketing Research*, 60, 625–645.
- McFadden, D. (1972). *Conditional logit analysis of qualitative choice behavior*.
- Papoutsaki, A. et al. (2016). WebGazer: Scalable Webcam Eye Tracking Using User Interactions. Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI). AAAI: 3839–3845.
- Piras, F. et al. (2020). Integrated Choice and Latent Variable model to evaluate the joint effectiveness of the introduction of a new light rail line and an informative measure. ETC Conference Papers 2020.
- Pool, E. R., et al. (2019). Behavioural evidence for parallel outcome-sensitive and outcome-insensitive Pavlovian learning systems in humans. *Nature Human Behaviour*, 3, 284–296.
- Reutskaja, E., et al. (2011). Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review*, 101, 900–926.
- Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, 134, Article 104472.
- RStudio Team. (2020). *RStudio: Integrated Development Environment for R*. Boston, MA: Place.
- Shimojo, S., et al. (2003). Gaze bias both reflects and influences preference. *Nature neuroscience*, 6, 1317–22.
- Slim, M. S., et al. (2024). Webcams as windows to the mind? A direct comparison between in-lab and web-based eye-tracking methods. *Open Mind*, 8, 1369–1424.
- Steffan, A., et al. (2024). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood en. *Infancy*, 29, 31–55. <https://doi.org/10.1111/inf.12564>
- Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge: Cambridge University Press.
- Vij, A., & Walker, J. L. (2016). How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, 90, 192–217.
- Yang, X., & Krajovich, I. (2021). Webcam-based online eye-tracking for behavioral research en. *Judgment and Decision Making*, 16, 1485–1505.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.