In-lab versus web-based eye-tracking in decision-making: A systematic comparison on multiple devices using discrete-choice experiments.

Sebastian Muñoz[1*], Vladimir Maksimenko [3*], Bastian Henriquez-Jara[2,4], Prateek Bansal[3], and & Omar D. Perez[1,4]

[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 [4] Complex Engineering Systems Institute (ISCI)
Chile

Author Note

*Authors contributed equally to this work. Correspondence regarding this article should be sent to Omar D. Perez (omar.perez.r@uchile.cl), Prateek Bansal (prateekb@nus.edu.sg) or Bastian Henriquez-Jara (bastian.henriquez@uchile.cl).

Acknowledgments: We thank Liting Yuan for her help with data collection for these experiments. **Funding:** This research was partially funded by ANID-PIA/PUENTE AFB220003, ANID-SIA 85220023, ANID-FONDECYT 1231027, and MoE Tier 1 Grant (A-8002145-00-00). The data collection was supported by the Presidential Young Professorship Grant at the National University of Singapore.

Declaration of interest: Authors declare to have no conflicts of interest associated with this research.

Abstract

Web-based eye-tracking has attracted increasing attention in recent years due to its ability to reliably capture alternative-based visual attention patterns in decision-making tasks. However, previous studies have primarily focused on validating webcam-based eye-tracking in perceptual and cognitive tasks, leaving open the question of whether it can be reliably applied to studies of choice behavior and economic decision making. In this study, we systematically compared the performance of the EyeLink 1000 Plus eye-tracker with a web-based solution (WebGazer) in discrete choice experiments conducted across different electronic devices, namely, monitor, laptop, tablet, and mobile. We found that the webcam-based system performed similarly well across monitor and laptop settings, but showed reduced reliability in the mobile condition and when the task was made more complex. Our findings provide the first systematic evaluation of the merits of web-based eye-tracking in decision-making research, offering critical insights into its viability for online behavioral studies.

Keywords: eye-tracking, attention, discrete choice

In-lab versus web-based eye-tracking in decision-making: A systematic comparison on multiple devices using discrete-choice experiments.

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 timeframe—typically a few hours.

From its inception, a primary concern in this regard 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. Despite these challenges, however, a seminal study conducted about a decade ago by Crump and colleagues (2013) demonstrated a high replication rate of classic experiments in psychology utilizing the M-Turk online platform, lending support to the reliability of online experimentation and encouraging its applicability across different fields of study.

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, which enables the analysis of visual attention and various 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; Krajbich *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 diverse experimental paradigms.

Although these results are promising, they have focused primarily on validating

webcam-based eye tracking in perceptual and cognitive tasks, leaving open the question of whether WebGazer can be reliably applied to studies of choice behavior and economic decision-making, a research domain where attention plays a fundamental role in explaining human choices, and which has been of great value to inform policy-makers in estimating 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 and Rose, 2024; Louviere and Hensher, 1982). Unlike prior studies which have largely examined raw gaze metrics, we employ theoretically grounded behavioral models that incorporate attention as a key explanatory variable. In addition, we systematically manipulate two critical factors: task complexity and display size. Task complexity was varied across experiments, allowing us to assess whether attentional patterns differ depending on cognitive demands. Display size, on the other hand, was manipulated within subjects, enabling us to examine how different viewing conditions, such as those experienced by subjects when using different electronic devices, impact fixation behavior and parameter estimation. Our findings thus provide the first systematic evaluation of web-based eye-tracking in decision-making research, offering critical insights into its viability for online behavioral studies.

Methods

Experimental Task

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 do people ascribe to each attribute) based on observed choices.

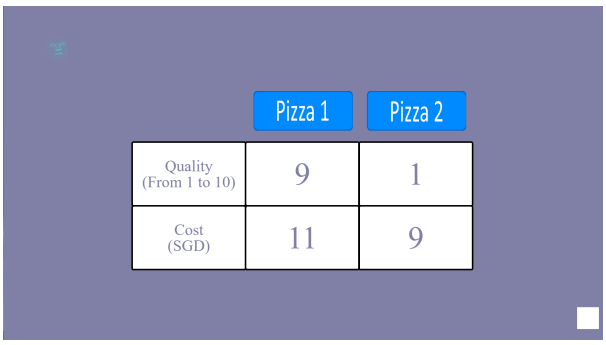
Experiment 1 featured a simple binary choice paradigm in which participants selected between two alternatives of pizzas based on two attributes: quality and cost. Quality was presented using a scale with five possible values $[1, 3, 5, 7, 9]$ (similar to the way various foods are presented in online web sites), 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 could assess how individuals weight quality relative to price in their decision-making. The simplicity of this 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. 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 attention allocation and choice behavior in the two eye-tracking methods.

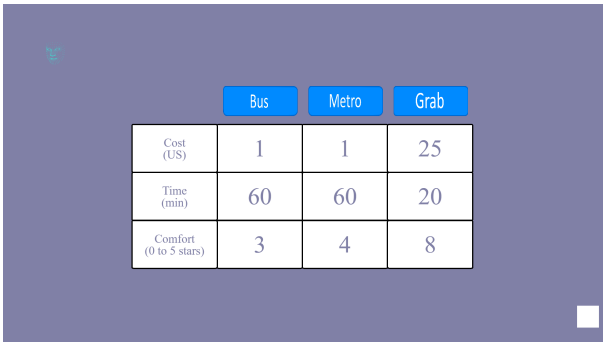
To evaluate the robustness of eye-tracking measures across different devices, both experiments were conducted under multiple display sizes, mimicking monitor, laptop, tablet, and mobile devices. This within-subject manipulation enabled us to assess how screen size impacts gaze tracking quality and decision model parameter estimation between the two eye-tracking methods.

Participants and Apparatus

Forty participants (14 males, 26 females; mean age = 23.9, SD = 6.66) were recruited for Experiment 1, and a separate group of 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



(a) *Experiment 1: Binary choice task*



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

Figure 1

Experimental design of choice tasks for Experiment 1 (Panel A) and Experiment 2 (Panel B). *Participants completed 12 choice trials per display condition (monitor, laptop, tablet, and mobile). The display layout was adjusted to fit the screen size of each device. Here, the laptop condition is shown.*

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 a compensation of S$10 for their participation.

Procedure

The experimental procedure followed four main steps, identical for both experiments. 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 re-calibration was performed separately for the two eye-tracking methods.

Participants completed 12 trials per experiment version, with the attribute order randomized across trials. The order of alternatives was kept fixed throughout the whole

experiment and was the same for all subjects in each experiment. The sequence of display sizes was also randomized across participants. Participants were asked to use a mouse click to make their choices (pizzas in Experiment 1; transport mode in Experiment 2). Each version lasted approximately 3–5 minutes, and the full experiment took about 40 minutes, including 5-minute breaks between versions.

Eye-tracker calibration

Eye-tracking methods were calibrated sequentially, starting from the EyeLink, followed by WebGazer. 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 $< 0.5°$ 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 recalibration 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.

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

Screen and choice display dimensions

Our interest was to manipulate display size to test how each eye-tracking method would generate records of fixations, dwell-times 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 1 summarizes the typical viewing distances and screen dimensions for each device.

Table 1
Screen sizes and viewing distances for different display conditions.

Device	Viewing distance (cm)	Screen size (cm)	Usable display (deg)
Smartphone	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

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:

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

Similarly, the height of each cell was calculated as:

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

Table 2 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.

Table 2
Choice cell dimensions across different devices

Device	Alternatives	Attributes	Cell Width (deg)	Cell Height (deg)
Smartphone	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
Smartphone	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

Results

In the present paper we aimed to address three main aims:

- The capability of the EyeLink and WebGazer to reliably record fixation counts and durations across various attributes and alternatives;
- The consistency of parameter estimates from decision-making models that integrate attention mechanisms in discrete choice experiments (DCEs) across both eye-tracking methodologies;
- The robustness of these findings across four distinct devices, namely monitor, laptop, tablet, and mobile.

To achieve these ends, we first analyzed fixation counts and dwell times (fixation duration), followed by parameter estimation in the choice model. Statistical analyzes and visualizations were performed in R using the RStudio IDE RStudio Team, 2020 for behavioral modeling; Python was also used for data visualization and fixation analysis. Parameter estimation was performed using the Apollo choice modeling package in R (Hess and Palma, 2019). Scripts for all analyses can be found at http://osf.io/aeuwx/.

Statistical analysis

To test if the number of fixations and duration of fixations (i.e., dwell times) differed across alternatives, for each experiment we ran a repeated measures ANOVA with eye-tracking method (EyeLink, WebGazer), display size (monitor, laptop, tablet, mobile), and alternative (2 alternatives in Experiment 1; 3 alternatives in Experiment 2) as within-subject factors. When a significant effect was found, we performed pairwise comparisons by t-tests corrected with Holm's method.

In Experiment 1 (pizza choices), we discarded 3 participants from the analysis of fixation duration across alternatives and 3 from the analysis across attributes, because they did not have fixations in one of the display conditions. For the same reason, in Experiment 2 (travel mode choices), we discarded 10 participants from the analysis of fixation duration across alternatives and 5 from the analysis across attributes.

Analysis of fixations

Experiment 1

Fixation counts

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 Figure 2 (panels A and 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 at pizza1 on average, but that this difference varied across display sizes.

Post-hoc tests (see Suppl. Table 2) 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 number of fixations to alternatives was reliably detected

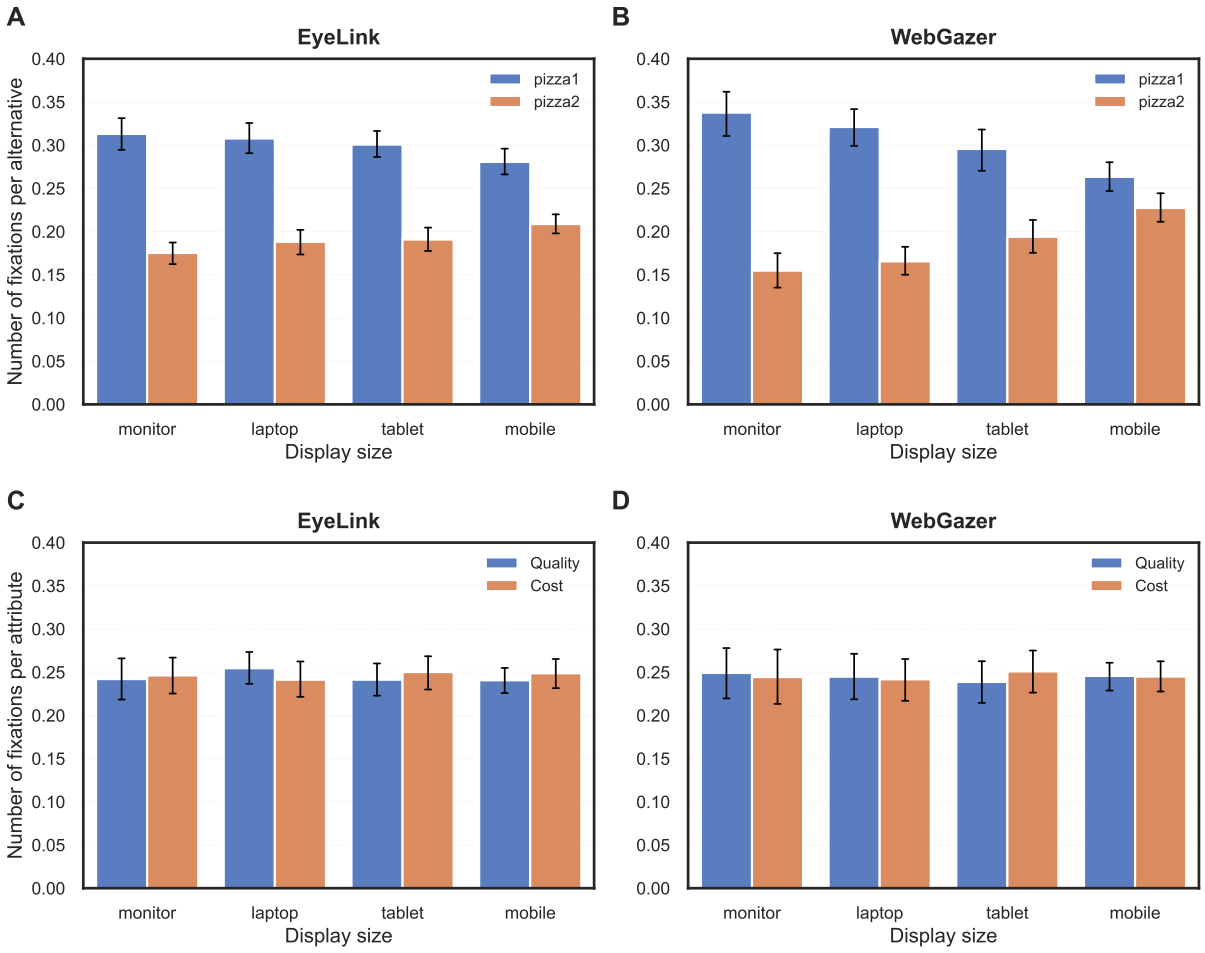


Figure 2

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).*

by both eye-tracking methods, with the exception of WebGazer in the mobile condition.

We then examined whether the number of fixations differed between attributes (quality and cost). We found no significant main effect of attribute ($F(1, 39) = 0.03, p < .001$), nor any interaction involving attribute and method or display size. This suggests that participants did not show preference in fixation for one attribute over the others, and that this pattern was consistent across methods and display sizes (see

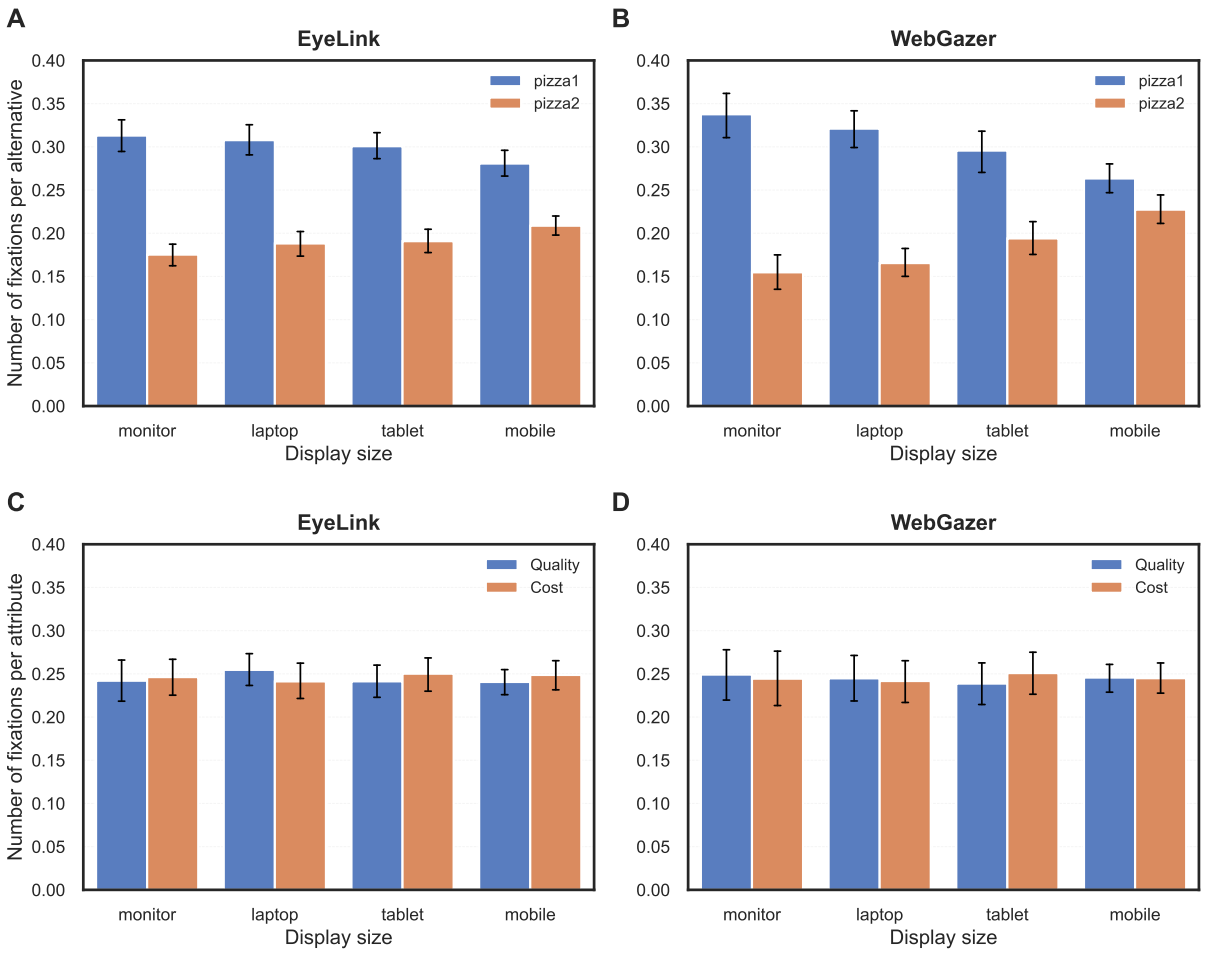


Figure 3

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).*

Figure 2, panels C and D).

Fixation durations

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, dwell times were higher for one eye-tracking method over the other. No other significant main effects or interactions were found.

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 ($F(1, 36) = 40.31, p < .001$), 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.

Experiment 2

Fixation counts

We tested whether the number of fixations on the three alternatives (metro, bus, grab) captured by EyeLink and WebGazer differed, and whether these differences remained stable across display sizes.

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. We also found a significant three-way interaction between method, alternative and display size ($F(6, 234) = 0.588, p = 8.748$). Post-hoc tests (see Suppl. Table 7) 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.

We next examined whether the number of fixations differed between the three attributes (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. No other main effects or interactions reached significance. These results suggest that participants allocated attention unevenly across attributes, but that this allocation was robust across different experimental conditions and eye-tracking methods.

Fixation durations

We next analyzed whether the duration of fixations on different alternatives (metro, bus, grab) differed, and 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 ($F(1, 29) = 21.76$,0 $p < .001$). Furthermore, there was a significant interaction between alternative and eye-tracker type, ($F(2, 58) = 10.38$, $p < .001$), 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 (see Supplemental Tables) 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. 5A, B). 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 ($F(1, 34) = 32.08$, $p < .001$), but no other 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

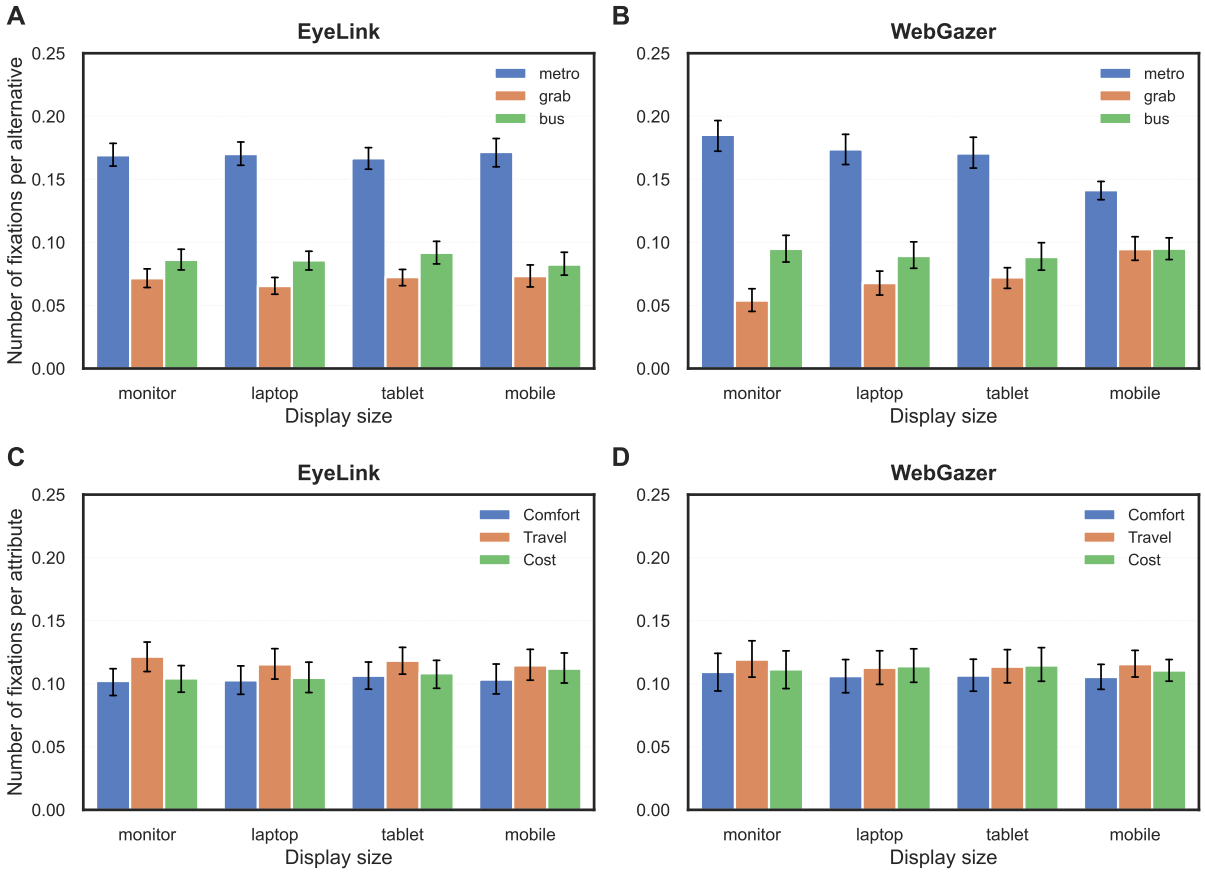


Figure 4

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).*

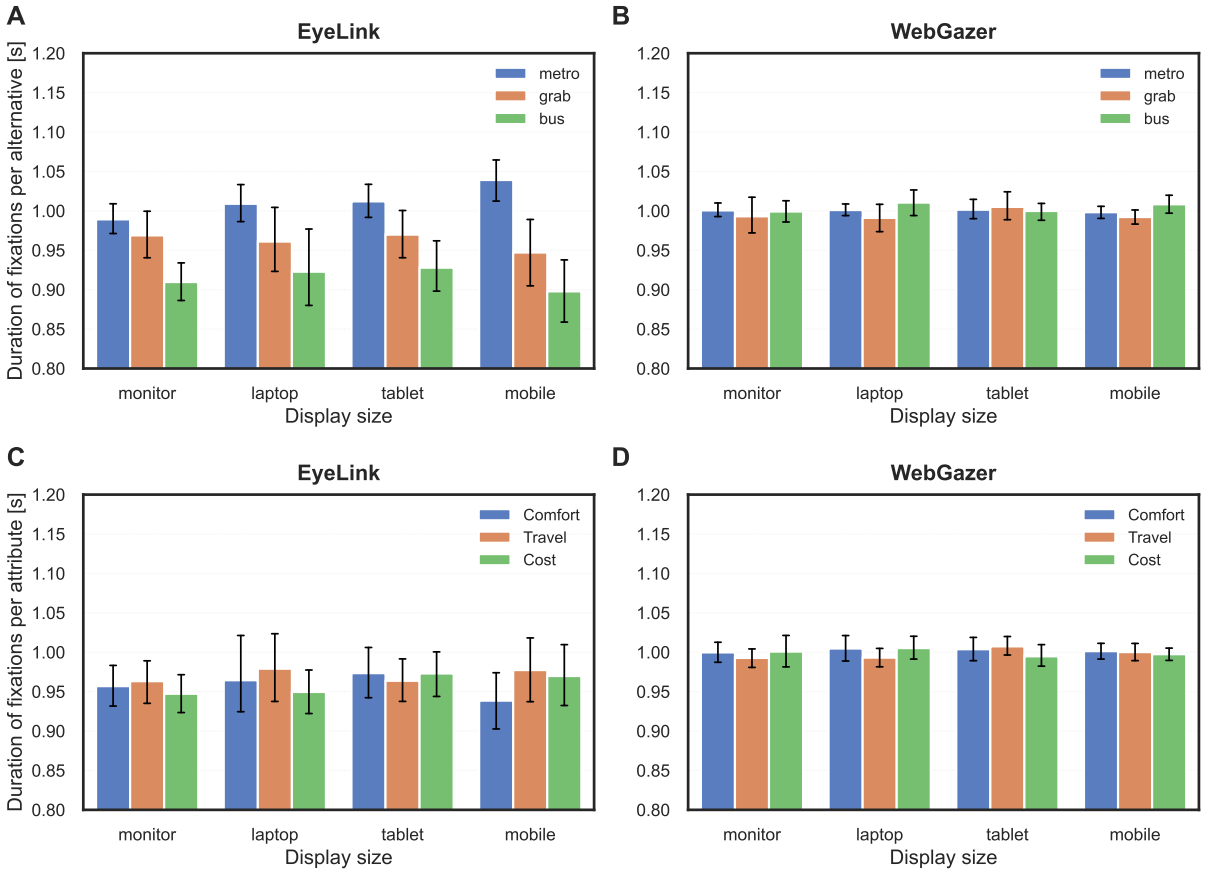


Figure 5

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).*

display size, but that the eye-tracking method employed has an effect on the recorded durations (Figure 5C, D).

Behavioral Modeling

Our second aim was to test if 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 and 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, 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}:

$$U_{nj} = V_{nj} + \varepsilon_{nj}, \tag{1}$$

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

$$V_{nj} = \sum_{k=1}^{K} \beta_k x_{njk}, \tag{2}$$

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:

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

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:

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

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

$$\tilde{\beta}_k = \beta_k + \alpha_k IP_n. \tag{5}$$

Here, IP_n is a latent variable representing an individual's information processing, and α_k quantifies how attention influences decision weights. The latent process IP_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.

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 ($\alpha_k IP_n$, 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. 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 between them. Results of WTP in Experiment 1 and SVTTS in Experiment 2 are shown in Singapore dollars (SGD).

Table 3

Comparison of Posterior Means and Distributions Between WebGazer and EyeLink

Device	Mean (95% CI)		t	p	KS	p	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 4
Comparison of distribution shapes for the latent variable between WebGazer and EyeLink across devices and attributes.

Device	Attribute	KS	p	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).

Experiment 1

Table 3 (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 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 4 (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 difference in the distributions between the two eye-tracking methods.

Experiment 2

Table 3 (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 4 (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 (Krajbich *et al.*, 2010; Le Pelley *et al.*, 2015; Reutskaja *et al.*, 2011; Shimojo *et al.*, 2003). 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). Given these trends, there is increasing interest in whether eye-tracking methodologies can be successfully adapted to online platforms. 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. We explored the results obtained with these two eye-trackers under display sizes similar to monitors, laptops, tablets and mobile devices. To test if task complexity would affect the results across eye-trackers and devices, we ran two experiments. In Experiment 1, participants chose between two pizza options that had two attributes each whereas Experiment 2 included three alternatives with three attributes, making the areas of interest smaller than in Experiment 1 for all conditions.

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 webcam eye-tracking. In Experiment 2, the metro option consistently received the highest number of fixations across all devices. The EyeLink system

consistently captured stable patterns across all display sizes, confirming its robustness across device types. WebGazer, on the other hand, performed similarly well across monitor, laptop, and tablet settings but exhibited reduced reliability in the mobile condition. This suggests that while webcam-based eye-tracking is effective for larger displays, its accuracy declines when tracking on smaller screens.

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.

Second, we analyzed fixation counts across attributes. In Experiment 1, both methods showed no differences in attention to attributes. In Experiment 2, EyeLink showed an effect in favor of travel time, but WebGazer was not able to capture this, suggesting attribute-based attention may not be consistently captured across eye-tracking methods. These results suggest that 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 more evenly across attributes rather than focusing on a single dominant one.

Finally, we examined fixation durations across alternatives and attributes. While some significant differences in fixation durations were observed — particularly for alternatives using EyeLink in larger displays — these effects were generally weaker and less consistent than for fixation counts, and largely absent with WebGazer. This suggests that

the time spent fixating on each alternative or attribute remained stable, regardless of whether participants were choosing between options in different domains. Importantly, fixation durations were longest for metro than the other alternatives, a trend that mirrored the fixation counts results. This suggests that alternatives receiving more fixations also tend to receive longer dwell times, reinforcing the notion that fixation count is likely a stronger indicator of decision salience than fixation duration. However, the absence of variation in fixation durations across attributes suggests that both eye-tracking methods may lack the sensitivity to capture subtle differences in dwell times, particularly in web-based settings, where noise from device characteristics and participant movement may obscure small temporal variations.

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 is 3 alternatives and 3 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 models based on attention and tested whether the parameters estimated 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 ($\alpha_k IP_n$) 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 was 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, suggesting that WebGazer can still reliably model attention-based decision processes in complex tasks when used with larger screens.

Even though there have been numerous studies testing the reliability of web-based eye-tracking, our study differs from prior experiments testing WebGazer. Previous work has largely focused on perceptual and cognitive tasks (Hutt *et al.*, 2023; Slim *et al.*, 2024; Steffan *et al.*, 2024), demonstrating that WebGazer can capture broad attentional patterns despite higher attrition rates and greater variability compared to in-lab systems. However, applications in decision-making contexts have been scarce. A notable exception is the study by Yang and Krajbich (2021), which employed WebGazer in a simpler two-alternative forced-choice task. Unlike 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.

In summary, 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.

References

Bansal, P., Kim, E.-J., and Ozdemir, S. (June 2024). Discrete choice experiments with eye-tracking: How far we have come and ways forward. *Journal of Choice Modelling*, **51**: 100478.

Bliemer, M. C. and Rose, J. M. (2024). Designing and conducting stated choice experiments. *Handbook of choice modelling*. Edward Elgar Publishing: 172–205.

Borozan, M., Loreta, C., and Riccardo, P. (Sept. 2022). Eye-tracking for the study of financial decision-making: A systematic review of the literature. *Journal of Behavioral and Experimental Finance*, **35**: 100702.

Cavanagh, J. F. *et al.* (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General*, **143**. Publisher: American Psychological Association: 1476–1488.

Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, **8**. Publisher: Public Library of Science San Francisco, USA: e57410.

Dayan, P, Kakade, S, and Montague, P. R. (Nov. 2000). Learning and selective attention. *Nature neuroscience*, **3 Suppl**: 1218–1223.

Hess, S. and Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of choice modelling*, **32**. Publisher: Elsevier: 100170.

Hutt, S. *et al.* (Jan. 2023). Webcam-based eye tracking to detect mind wandering and comprehension errors. en. *Behavior Research Methods*, **56**: 1–17.

Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, **13**. ISBN: 1546-1726 (Electronic)$\$n1097-6256 (Linking): 1292–1298.

Krucien, N., Ryan, M., and Hermens, F. (2017). Visual attention in multi-attributes choices: What can eye-tracking tell us? *Journal of Economic Behavior & Organization*, **135**. Publisher: Elsevier: 251–267.

Le Pelley, M. E. *et al.* (2015). When goals conflict with values: Counterproductive attentional and oculomotor capture by reward-related stimuli. *Journal of Experimental Psychology: General*, **144**: 158–171.

Louviere, J. J. and 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., and Erdem, T. (Aug. 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**. Publisher: Springer US ISBN: 4156201805.

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. (Feb. 2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, **134**: 104472.

RStudio Team (2020). *RStudio: Integrated Development Environment for R*. Place: Boston, MA.

Shimojo, S. *et al.* (Dec. 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**. Publisher: MIT Press 255 Main Street, 9th Floor, Cambridge, Massachusetts 02142, USA . . . : 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**. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/infa.12564: 31–55.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Vij, A. and Walker, J. L. (2016). How, when and why integrated choice and latent variable models are latently useful. *Transportation Research Part B: Methodological*, **90**. Publisher: Elsevier: 192–217.

Yang, X. and Krajbich, I. (Nov. 2021). Webcam-based online eye-tracking for behavioral research. en. *Judgment and Decision Making*, **16**: 1485–1505.

Appendix

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:

$$U_{nj} = \sum_{k=1}^{K} \beta_k x_{njk} + \varepsilon_{nj}. \tag{A1}$$

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

$$U_{nj} = \sum_{k=1}^{K} \tilde{\beta}_k x_{njk} + \varepsilon_{nj}, \tag{A2}$$

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

$$\tilde{\beta}_k = \beta_k + \alpha_k IP_n. \tag{A3}$$

The latent information processing variable IP_n follows a normal distribution:

$$IP_n \sim N(\gamma_{IP}, \sigma_{IP}). \tag{A4}$$

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:

$$VA_{nk} = \gamma_0 + \gamma_k IP_n + \omega_{nk}, \tag{A5}$$

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:

$$P_{ni}(x_{ni}, \beta|\eta_{ni}) = \frac{exp(\sum_{k=1}^{K} \beta_k x_{nik} + \alpha_k IP_n)}{\sum_{j \in \mathcal{C}} exp(\sum_{k=1}^{K} \beta_k x_{njk} + \alpha_k IP_n)}. \tag{A6}$$

Likelihood Function

The likelihood function integrates over the latent variable:

$$L = \prod_n \int_\eta \left(\prod_k f_{nk}(\sigma_k, \gamma|\eta_{nk}) \prod_j P_{nj}(x_{nj}, \beta|\eta_{nk})^{y_{nj}} \right) d\eta. \tag{A7}$$

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

$$SLL = \sum_n \ln \sum_{r=1}^{R} \frac{1}{R} \left(\prod_k f_{nk}(\sigma_k, \gamma|\eta_{nk}^r) \prod_j P_{nj}(x_{nj}, \beta|\eta_{nk}^r)^{y_{nj}} \right). \tag{A8}$$

Posterior distributions for WTP and SVTTS

Experiment 1 (pizza choices)

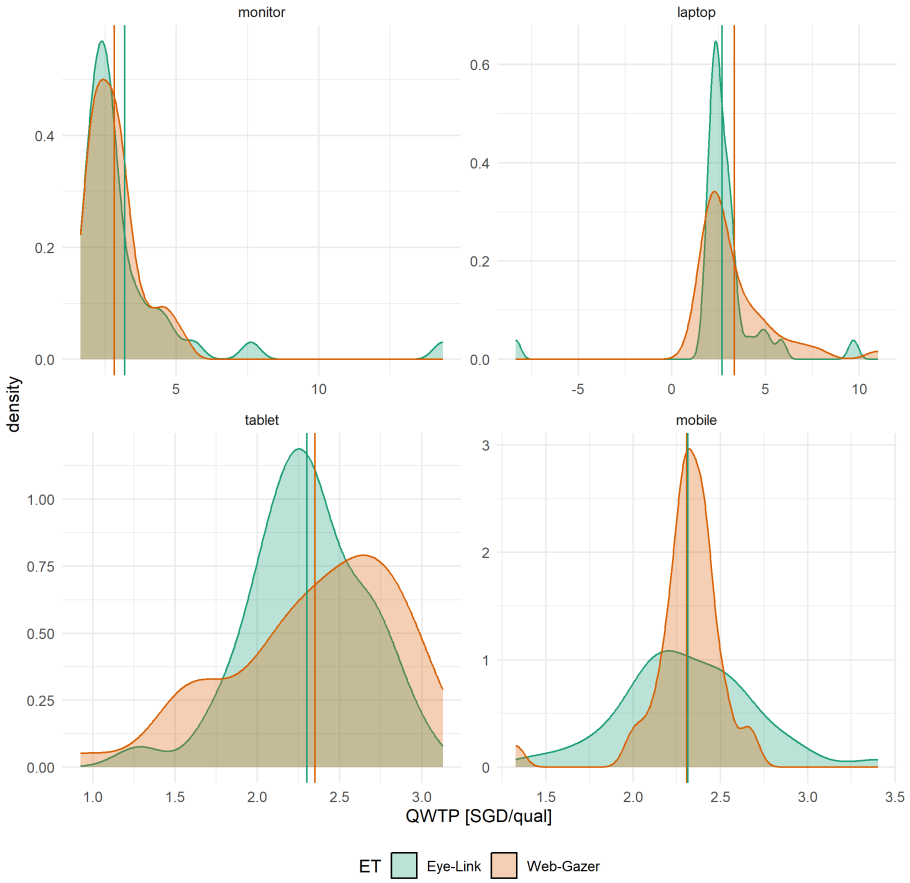


Figure A1

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

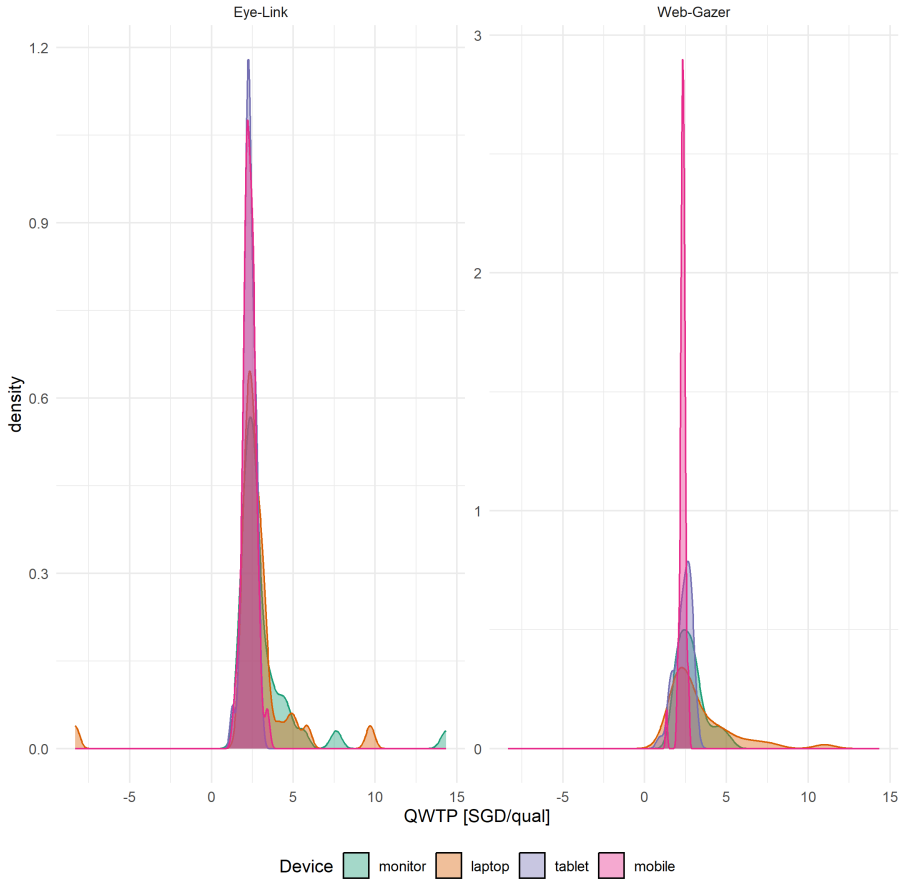


Figure A2

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

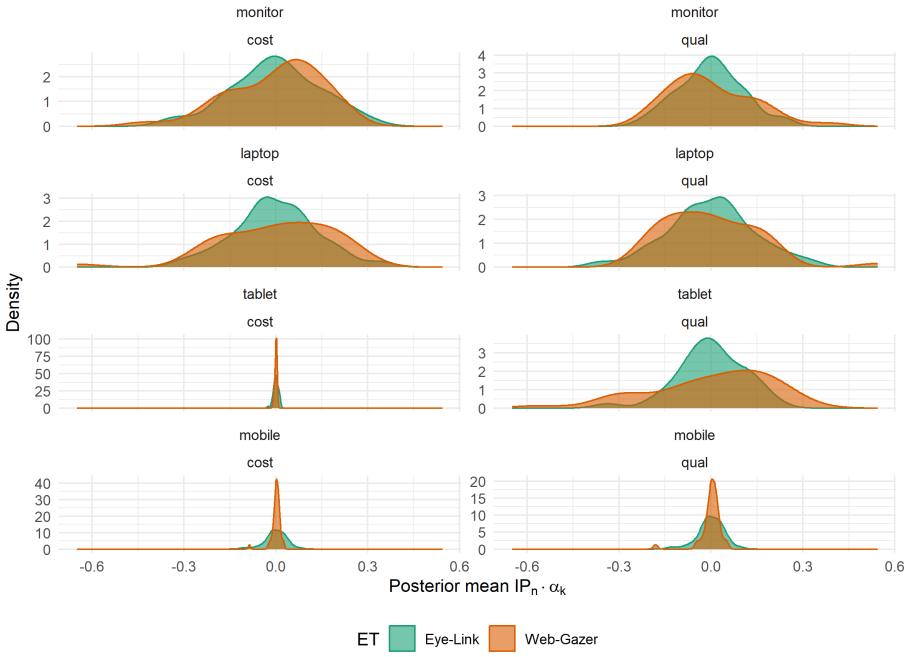


Figure A3

Posterior distributions of $\alpha_k IP_n$, grouped by display size for Experiment 1.

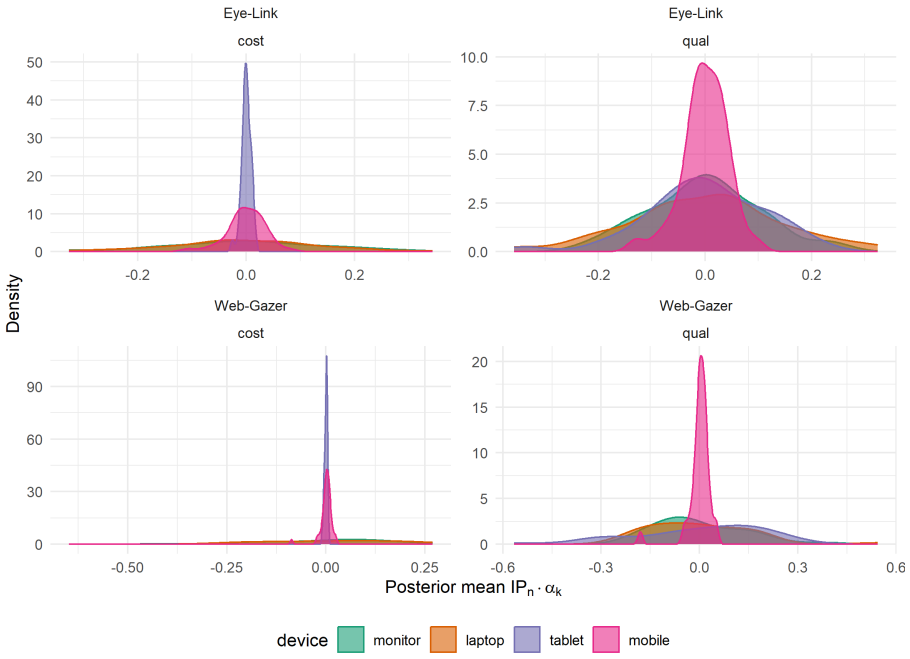


Figure A4

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

Experiment 2 (transport choices)

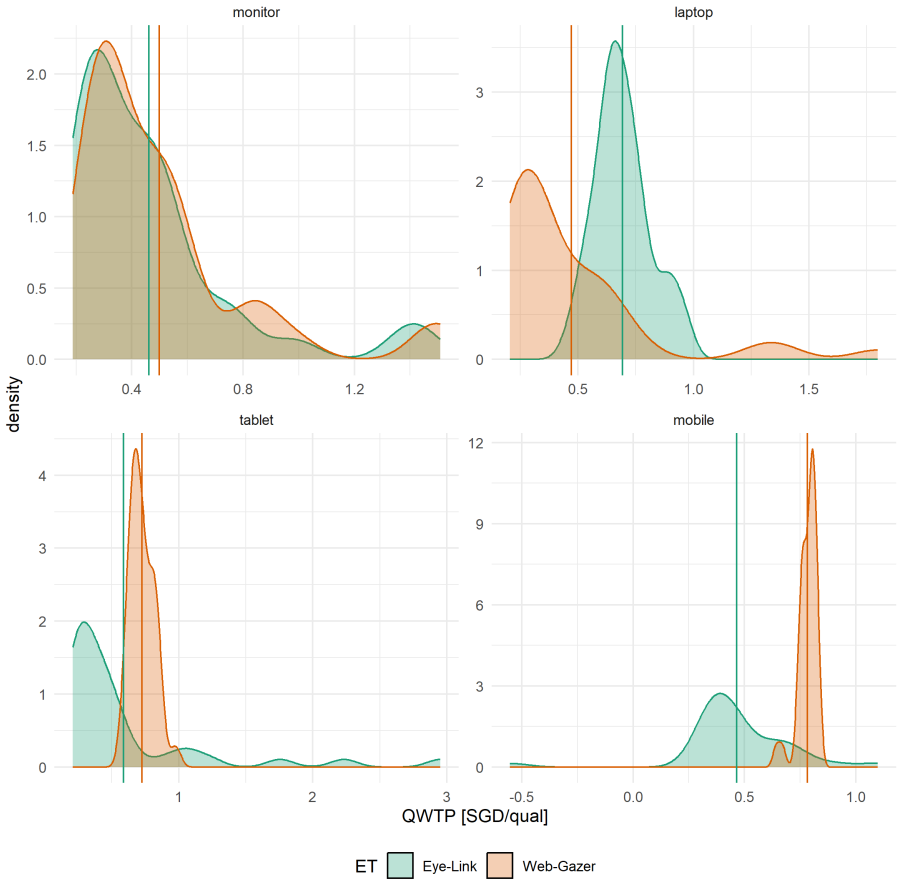


Figure A5
Posterior distributions of SVTTS, grouped by device, for Experiment 2.

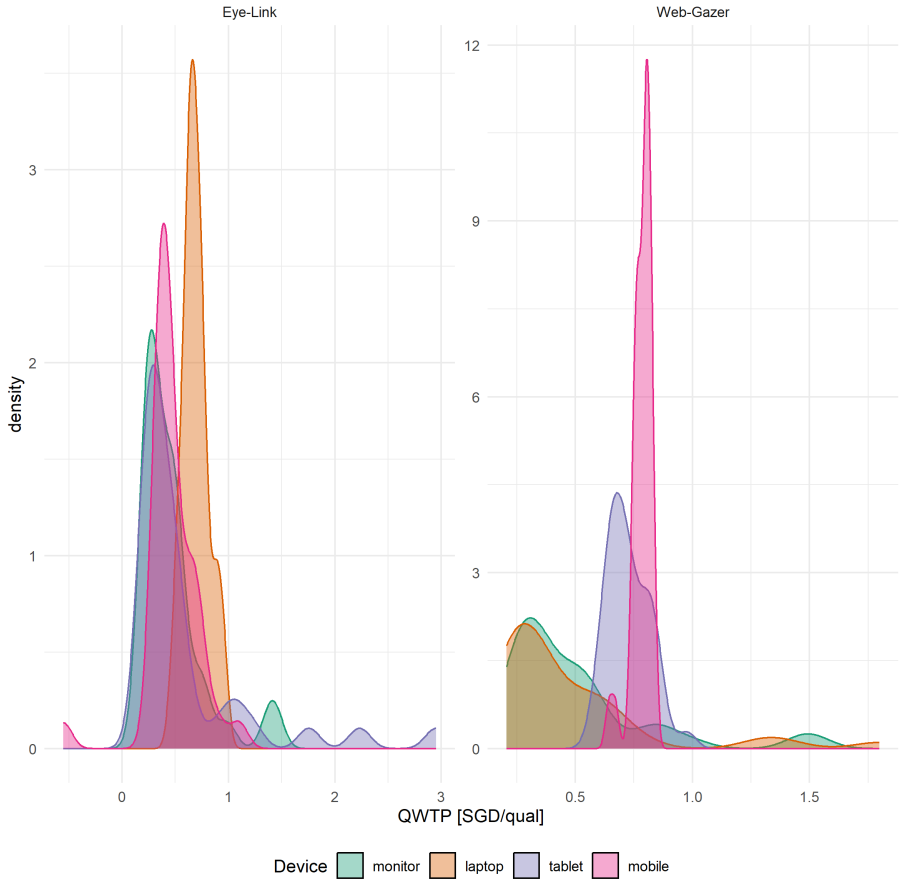


Figure A6

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

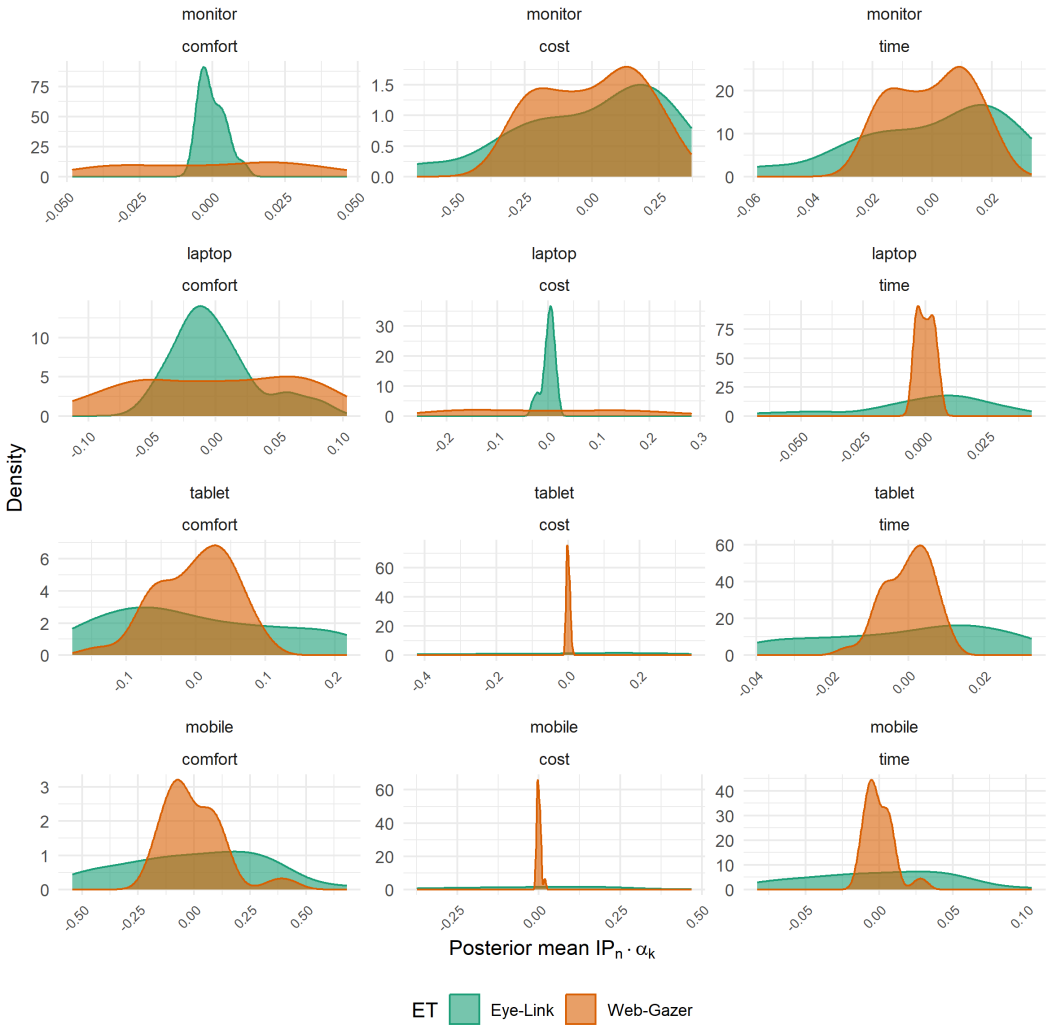


Figure A7

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

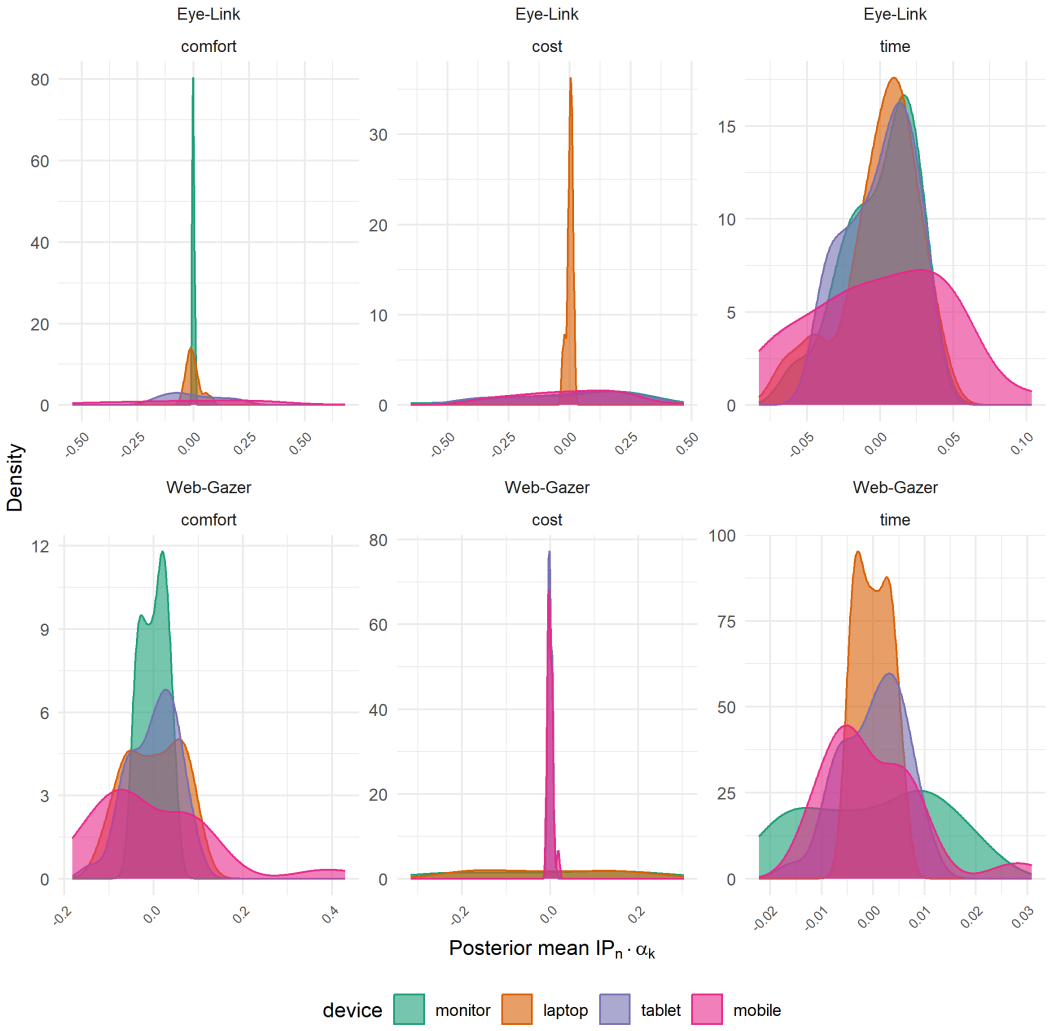


Figure A8

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