Auditing E-Commerce Interfaces With Browser-Based Field Experiments: A Method to Assess Effects of Visual Elements on Consumer Behavior

ANONYMOUS AUTHOR(S)

Digital regulation is increasingly concerned with visual elements and their influence on consumer choices. However, regulators lack methods to independently, rigorously and realistically assess their effects. In this paper, we develop and apply a methodology to run browser-based experiments on a major online travel agency. Participants make consequential holiday accommodation choices from the lab, on a live version of the website. Using a browser extension, we manipulate the presence of marketing cues within-subject and the overall level of clutter between-subjects. We exert significant control over hotel choice sets, clustering offers into comparable groups and selecting which appear on screen. The results are causal effects on choices, attention and welfare. We show that a causal, credible measure of mere visual effects is affordable to regulators with a careful experimental design. We also discuss the main implementation challenges of this method and release all our artifacts to facilitate future studies.

CCS Concepts: • **Human-centered computing** → **Laboratory experiments**; *Field studies*; • **Applied computing** → **Online shopping**; **Economics**; Law.

Additional Key Words and Phrases: browser-based field experiment, browser extension, interface auditing, dark patterns, endorsement cue, visual clutter, online travel agency, consumer welfare, preregistration

ACM Reference Format:
Anonymous Author(s). 2026. Auditing E-Commerce Interfaces With Browser-Based Field Experiments: A Method to Assess Effects of Visual Elements on Consumer Behavior. 1, 1 (September 2026), 29 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Regulators have recently made interface design a matter of law. Enforcement against "dark patterns"[1] in California and in the EU under the Digital Services Act (DSA) both rest on the premise that a badge, banner, or default can push consumers toward choices they would not otherwise make — potentially harming them [26, 39, 42]. Evaluating whether an interface element harms consumers requires an answer to a causal question: what would consumers have done, had this element not been there? This answer must be obtained independently of the firm owning the website. It must be rigorous enough to survive legal challenges and realistic enough to convince skeptics. It should ideally be obtained in a limited time at a reasonable cost. Both the academic literature and real-world regulatory practices are still far from those three goals. Website audits and crawls document what interfaces contain, not what they do to consumers' choices [15, 41, 53]. Laboratory experiments on researcher-built emulations recover causality but often discard the brand, the stakes and the visual environment in which a cue actually competes for attention [39, 56]. Experiments run inside the platform are the most credible of all, but they require the platform's consent [3, 34].

[1]Dark patterns are defined in California law as a user interface "designed or manipulated with the substantial effect of subverting or impairing user autonomy, decisionmaking, or choice".

Manuscript submitted to ACM

Browser-based field experiments offer a promising alternative. The idea is that a browser extension rewrites a live commercial website on the participant's own screen, modifying only certain elements while keeping the rest of the website the same. This allows a regulator or a researcher to randomize an element of the interface without the firm's cooperation — but on real pages, at real prices, with real purchases. Because of the randomization, this delivers the same clear causal evidence as usual lab experiments, with a much higher level of realism. Another benefit is allowing exploration of a wider array of counterfactuals. An audit can only compare interfaces that already exist; an in-platform test is limited to those the firm agrees to run. An extension can remove a badge the platform cannot remove for contractual reasons, or strip a page of its ads, and observe the same consumer on both versions. This approach has been successfully applied to gauge the effects of hiding Amazon's own brands from search results [19, 22] inserting cookie-consent pop-ups [47], reranking social media feeds [49] or substituting posts into a live Reddit interface [40].

In this paper, we propose a method to run a browser-based experiment on Booking.com's *Preferred Partner* badge. Booking.com is a major online travel agency, a Very Large Online Platform under the EU DSA and hence subject to specific scrutiny by regulators. Booking.com features a *Preferred Partner* badge, which is a thumbs-up icon that gives properties higher visibility in exchange for paying a higher commission. Our main objective was to study the effect of this badge on consumer behavior. Additionally, we wanted to test whether the level of clutter on the page – in particular, the use of bold and multi-colored fonts and the presence of redundant information – moderates the badge's effect. To the best of our knowledge, this is the first attempt at a browser-based experiment on the online travel agent (OTA) industry and the first randomized experiment on Booking's *Preferred Partner* badge ever published[2].

We invited sixty participants with diverse demographics to a university laboratory to make 12 hotel choices on the live website. Choices had real consequences and monetary incentives. Participants were informed in advance that their chosen hotels might actually be booked: one of the sixty participants would receive a prize of €400 to spend in the city and the weekend of the reservation, from which we would subtract the cost of their reservation to make the booking in their name on the website. In addition to raffle entry, all participants were financially compensated for their participation in the one-hour experiment. To ensure participants had a genuine interest in the reserved weekend, they picked four weekends in October or November 2026 when they were free and discarded their least favorite of the four proposed destinations. Each participant made 12 hotel choices (4 weekends $\times$ 3 destinations).

Our browser extension varied the badge's visibility *within* participants — visible on two of the four weekends of each city, hidden on the other two. Meanwhile, the clutter of the interface varied *between* participants, half seeing the standard page and half a version where some twenty-five promotional and navigational elements had been removed or desaturated. Our analysis rests mostly on the comparison of choices made by the same subject for the same city, with and without visible badges. The extension also logged participants' choices, clicks, dwell times, and scrolls and hovers on pages. Participants answered a post-experiment questionnaire which explored, among other things, whether they noticed the badge, what they thought the badge meant, and how much they appreciated the associated listings. Hypotheses, outcomes, tests and exclusion rules were preregistered before any data were seen.

One important methodological innovation compared to past browser-based studies is the degree of control we exert over the choice sets. Instead of letting participants choose from every offer on the website, we selected in advance, for each choice, the nine properties that would appear. Properties were scraped three days before the sessions, clustered within each city by price, rating and geography. We then formed choice sets having three properties per cluster, exactly

[2]The closest studies on this industry are Ursu [58] on Expedia's ranking system, which was run by the platform itself, as well as Hunold et al. [30] and [31], which are purely observational studies on the supply-side effects of Booking Preferred Partner's badge

one of them with a badge. Although some online experiments are similarly sophisticated in their design, achieving such control over the choice set of a live website with a large offering is a real challenge that we are the first to address.

This additional control gives us leverage to study the potential harm caused by the badge without a structural demand model, the usual tool in the economics literature. For a given city, a participant's two badge-hidden choices reveal whether they consistently prefer one cluster of comparable offers; we then observe whether their two badge-visible choices stay in that cluster. This measure does not assume that participants share similar preferences, which makes it well suited to contexts where preferences are very heterogeneous.

This experiment's results should be read as indicative rather than definitive, given our limited sample size. For the badge, every preregistered point estimate goes in the expected direction: a visible badge raises the chance that a badge-eligible listing is clicked (significant at the 10% level), and also raises the time spent on badge-eligible listings' pages and the chance that a badge-eligible listing is chosen (not significant). Despite these effects, most participants did not know that properties paid for the badge to be displayed. When asked after the experiment, 40% described the badge as being awarded based on other customers' reviews, and 45% described it as a generic sign of quality. Only one participant mentioned that properties paid for the badge to be displayed. Despite the data showing that participants' choices were influenced by the badge, 77% of participants said that they did not use the badge in making booking decisions. A cue that consumers do not consciously recognize as influencing them, but nevertheless respond to subconsciously, is precisely the kind of element regulation is concerned with.

Our measure of preference consistency did not detect the existence of a positive or negative effect for consumers. Still, we can validate that our clusters are good measures of consumer preferences: in the no-badge condition, consumers' choices are much more consistent in cluster than what completely random choices would produce. Our visual clutter results are more puzzling. The de-cluttered interface was not perceived as simpler: if anything, self-reported visual complexity was slightly higher. Additionally, self-reported cognitive load moved in mixed directions, with only perceived time pressure significantly higher on the cluttered page. Clutter did not dilute the badge's effect and did not reduce its recognition; and, contrary to our preregistered heuristic-processing hypothesis, participants on the cluttered page opened *more* listings and took longer to decide, not fewer and faster.

We make available all the artifacts required to run the experiment as supplementary materials: the web browser extension, the scraping tool and the pipeline to build clusters and choice sets, the code of the oTree experiment and the analysis code.

2 Background and Related Work

Measuring interfaces from the outside. A large strand of the literature measures interfaces at scale without observing users. Crawls can document how widespread a design pattern is [15, 41]. Audits can infer steering and discrimination from the outside [13, 28, 43, 53]; for instance, Eslami et al. [17] showed that Booking.com's rating algorithm inflated hotel scores by up to 37%. This literature is predominantly descriptive, documenting occurrence rather than effect [42]. This is perfectly fine for auditing algorithms, since they can be deemed intrinsically problematic – because of discrimination or self-preferencing for instance – independently of their consequences. However, this approach cannot determine whether a specific interface element has "the substantial effect of subverting [...] choice".

Randomized evidence on interface elements in artificial choice tasks. A second tradition studies human behavior using randomization, but is thus far limited to relatively artificial environments. For instance, Luguri and Strahilevitz [39] found that aggressive dark patterns nearly quadrupled subscription uptake, and a growing HCI literature catalogs the

elements at stake [6, 8, 24, 27]. Similar instances can be found in experimental economics: de Haan and Linde [14] show that past exposure to good defaults reinforces default bias. Bogliacino et al. [5] show that dark patterns raise choice inconsistency relative to a plain design. Almost all of this evidence comes from interfaces the researchers built, which discards the brand, the stakes and, critically for a study of attention, the authentic visual environment in which a cue must compete. Even seminal studies on visual salience in neuroeconomics [44, 50] do not feature basic elements such as prices.

The visual environment in which choices are made is not just a source of measurement noise: this also affects search, recognition and decision-making in ways that are well characterized in vision science [45, 51, 52, 55]. Banner blindness shows that salient elements can be ignored precisely because they look promotional [4, 7] and aesthetic impressions form in milliseconds and shape later judgments [38, 56, 57]. This is precisely why the ability of large e-commerce websites to fine-tune their visual environment at scale through A/B testing is concerning. Our experiment echoes the methodological debates on the external validity of these experiments run in artificial choice environments by comparing the effect size on the original Booking interface to that obtained in a low clutter condition — closer to the visual interfaces usually tested in the lab.

Experiments inside and outside the platform. Running the experiment inside the platform is the surest way to ensure both realism and causality at once [3, 34, 35]. However, studies requiring the firm's cooperation pose ethical questions [36] and may not be adapted to support regulatory intervention. Without firm cooperation, learning about consumer behavior on a real website becomes very difficult: the most recent CHI analysis of Amazon's cancellation flow had to reconstruct the interface from screenshots disclosed in FTC litigation [25].

Browser-based field experiments offer an alternative. The WebMunk platform instruments and modifies participants' own browsing on Amazon, including hiding Amazon's house brands from search results to estimate the effect of platforms preferencing their own products [19–22]. Extensions have also injected synthetic consent pop-ups onto real sites [47], reranked social feeds [49], and substituted researcher-authored posts into a live Reddit interface [40]. Like the WebMunk studies, we suppress the platform's authentic elements rather than injecting fabricated ones. What we add is control over the choice set itself. We carefully select ex ante, from the whole set of offers on the website, which offers will appear at which stage of the experiment. In comparison, similar studies like one on Amazon's Choice badge [37] just limited results to the first search page, which is technically much easier, but brings no specific benefit to the analysis. This places our study closer to a *field simulation* than to either a laboratory experiment or a field study, a class of designs HCI has long argued can recover much of what field deployment offers at a fraction of its cost [9, 32, 33].

Assessing harm from choice data. It is generally not sufficient to observe user behavior to derive conclusions about the harm caused by an interface element. Such an assessment requires some understanding of consumer preferences, which may be extremely heterogeneous across individuals in e-commerce. Economists analyze this question in terms of utility and typically follow one of two approaches. The first estimates a structural model of demand on observational or experimental choice data in a non-real-website reference situation — in which consumers' choices are assumed to be aligned with their actual preferences — and then uses these data to evaluate the utility derived from choices made on the real website [22, 23, 58]. This approach can be quite data-intensive and makes numerous modeling assumptions in the specification of the structural model. The second approach asks consumers about their decision-making directly, for example asking about their willingness to accept deactivation of a service [1]. This second approach is quite dependent on consumer sophistication, i.e., consumers being aware of their tendency to deviate from their own preferences. It is more suited to studying addiction to social media, a behavior consumers are very conscious about [2], than to

quantifying the harm caused by a specific interface element that operates below the threshold of consciousness. Our design permits a third, reduced-form answer: because the choice set is fixed and organized into clusters of comparable offers, repeated choices in the absence of the cue reveal a preferred cluster, and the cue's welfare effect is read from whether choices made in its presence leave that cluster (Section 3.3). Our approach is also perfectly compatible with the estimation of structural demand models — and actually attractive, because choice sets are well-balanced — but we choose to focus on the more novel and direct approach in this paper.

Preregistration in HCI.. Finally, our experimental design takes seriously the concern that flexible analysis inflates false positives [48, 54], a concern HCI has addressed through calls for preregistration and fairer statistical reporting [11, 12, 16, 46]. Every hypothesis, outcome, test direction and exclusion rule below was fixed on OSF before data collection, and we mark in the paper the analyses that were not.

3 Methods

Browser-based field experiments have two main components: an experimental design, which ensures that effects are causally, precisely and realistically measured, and a web browser extension implementing it. Both must be tailored to the interface under study. After introducing these two components which determine the structure and nature of our data, we present the statistical approach we followed to estimate causal effects.

3.1 Experimental design

The experiment was a 2×2 factorial design in which each participant makes 12 consequential accommodation choices on a live but altered version of the Booking.com website. The 12 choices correspond to three destination cities and four weekends. The manipulated factors are the visibility of the platform's endorsement cue, varied within subject and city across weekends, and the visual clutter of the interface, varied between subjects. The design was preregistered before any data were observed.

Treatments. The first treatment is *endorsement cue visibility*. Booking.com marks properties enrolled in its Preferred Partner Program with a yellow thumbs-up badge, an arrangement in which the property pays a higher commission for greater visibility. By construction, exactly three of the nine listings in each choice set are cue-eligible, meaning they carried the badge on the unmodified site on the date of the experiment. In the cue-hidden condition, the extension removes the badge from every listing in the set; in the cue-visible condition, the property's badge status is the same as Booking.com originally displayed it. We never display a badge on a property that did not have it originally. Cue visibility varies within subject and within city: for each of a participant's three cities, the badge is visible for two of the four weekends. Figure 1 illustrates the resulting design for three example participants.

The second treatment is *visual clutter*, varied between subjects and constant across all twelve of a participant's tasks. The cluttered condition leaves the property cards as they appear on the standard interface. The low-clutter condition removes or desaturates roughly twenty-five types of supplemental text from the search page's property cards. For instance, whether taxes are included, cancellation policy and number of rooms available are irrelevant in our experimental design. Price, review score, address, availability, room configuration, are untouched and left on property cards. The objective is to manipulate how hard the page is to read without changing the information participants have access to.

Crossing the two treatments gives four possible configurations, illustrated in Figure 2.

Notes: Cue visibility varies within subject and city — exactly two of each city's four weekends show the badge, in a random permutation. Clutter varies between subjects — one condition for all twelve of a subject's tasks, alternated by order of arrival. CS1–CS4 label a city's four choice sets, randomly assigned to that city's four weekends, independently for each subject.

Fig. 1. Illustration of the design of the experiment

oTree application and randomization. Experimental sessions are orchestrated by the laboratory management application oTree [10]. Sessions unfold as follows: instructions with comprehension checks, consent, a pretask in which the participant chooses cities and weekends, twelve choice-task rounds, and a post-experiment questionnaire. oTree owns the protocol, the randomization and the questionnaire data; the extension manipulates the interface and tracks behavior. The two communicate instantly using short codes on screen that the participant cannot transcribe. Some AI tools were involved in writing the oTree code.

Three things are randomized at session creation. Cue visibility is drawn per participant and per city as a random permutation of two visible and two hidden weekends. Within each city, choice-set order is a random permutation, so position in the session is orthogonal to weekend and choice set. Clutter is assigned to participants by alternating conditions in order of arrival. The three retained cities are presented in a fixed order (La Ciotat, Arcachon, Le Tréport, Sète, minus the excluded one), and the four weekends of a city in chronological order.

Subject pool and recruitment. Participants were recruited from the subject pool of [MASKED FOR ANONYMIZATION] and took part in one of three sessions run in July 2026 on two successive days. The study was approved by the institutional review board of [MASKED FOR ANONYMIZATION]. Eligibility was screened before invitation on two criteria: prior experience booking holiday accommodation online, since the task presumes familiarity with this class of interface, and availability for at least four of the nine proposed weekends in October/November. Comprehension checks were

administered after the instructions, but we deliberately did not preregister any exclusion based on them; they served to identify participants who needed the instructions repeated and to perform robustness checks a posteriori.

Real incentives. Choices made in the experiment had a positive probability of materializing. Participants were told in the instructions that they each had a 1% to 2% chance that one of the twelve choices they made in the experiment would be implemented[3]. The prize had a total value of €400, and any part of the €400 not absorbed by booked lodging could be reimbursed for documented travel, meal or activity expenses incurred during that weekend in that city. We made sure that the reservation could be implemented by running the sessions and building the choice sets a few days after our last scraping. We also ensured that participants could be genuinely interested in the reservation by letting them remove one destination city out of four and having them confirm in advance that they would be available to benefit from the reservation in at least four of the nine possible weekends.

Pretask. The pretask ensured that the experiment engages actual consumer preferences. Participants saw nine candidate weekends across October and November 2026 and selected exactly four for which they are genuinely available. They needed to keep three of the four coastal cities — La Ciotat, Arcachon, Le Tréport and Sète — and were asked to remove the one in which they would have an alternative place to stay, if any. Their twelve choice tasks were the retained cities crossed with the chosen weekends. They were then prompted to reflect on what they look for when choosing accommodation, and given information about the cities in the form of links to relevant websites. The websites were tourist information offices, the Wikipedia and OpenStreetMap pages of the city and a Google Maps itinerary by train from Paris to estimate the duration. Participants were free to navigate on these websites, could use Google Search and the website of the main railroad company in France, but could not access other domains throughout the experiment. They were told so explicitly on this page. They were also asked not to try to obtain information about specific accommodation offers during that phase and encouraged to learn more about the destinations. This stage has an incompressible duration of ten minutes, which left participants time to form preferences — even if they had none initially — before the choice tasks start.

Choice Task. During the choice task, participants were reminded on the oTree page of the city and weekend for which they are about to make a choice. They had to click on a link opening a new tab with the corresponding page of Booking.com. The page initially displayed a loading panel while the extension filtered out properties that are not part of the choice set. The oTree page also featured the same links to external websites as the pretask page, but only for the choice-relevant destination. Once the Booking page had loaded, participants could explore the property pages and had to click one of the reservation buttons on screen to confirm their choice. Instead of taking them to the room selection or payment part of the website, the extension saved a record of their behavior, closed all the opened Booking tabs and brought them back to the oTree page. They then needed to click a button to move to the next choice.

Post-experiment questionnaire. The questionnaire collected four groups of variables. First, participants were given two manipulation checks: their rating of the perceived visual complexity of the interface, on a seven-point scale from "very simple" to "very complex", and cue recognition. For cue recognition, three icons were displayed — the thumbs-up badge and two that never appeared on the site — and participants were asked if they noticed each. Second, participants were asked to rate the experiment along the six subscales of the NASA Task Load Index [29] — mental demand, physical demand, temporal demand, performance, effort and frustration. Each item was rated on a 0–100 slider in steps of five,

[3]In practice, the research team made the reservation on Booking.com for a single subject drawn at random out of the 62 who came to the lab, which gives a probability of 1.6% of being chosen.

with the standard anchors ("low" to "very high", and "failure" to "very successful" for performance). Third, we collected data on participants' beliefs on what the badge represented and self-reported use of it in their decision-making process. We only revealed at this point that the thumbs-up was the only real icon, and asked participants to rate their agreement on a four-point scale with the statement "in general, accommodations with a yellow thumbs-up correspond more to my preferences than the others", used in Section 4.4. We also asked them for their open-ended answer on what the thumb icon meant and whether/how they used it in making decisions. Fourth, we also collected participants' demographics and background: gender, age, student status, household structure, residence in the Paris region, and familiarity with Booking.com on a four-point scale from "never used" to "use it regularly". The questionnaire variables used in the analysis are defined in Table 5, Appendix A.2.

Choice sets. Each choice set contained exactly nine listings, arranged as three clusters of three, with exactly one cue-eligible property per cluster. Clusters were groups of comparable offers defined at the city level that we introduce to facilitate the subsequent welfare analysis.

Clusters were built from scraped listing data in two stages. First, we used k-means to partition a city's properties on price, review score and geography. Properties above €400 for two nights were excluded, as this was beyond the maximum possible payout for lottery winners.[4] We used quantiles of latitude and longitude to make the geographic component robust to outliers. Second, because a k-means partition will not hold enough cue-eligible properties in every cluster to supply four disjoint choice sets, a mixed-integer program reassigned properties across clusters, minimizing the total increase in distance to centroids subject to each cluster holding enough cue-eligible and non-cue-eligible properties to fill all four sets. The clusters are therefore both feasible and close to the unconstrained solution. Choice sets were then drawn without replacement, using two non-cue-eligible and one cue-eligible property for each cluster, four times per city. In Appendix A.3, Figure 8 illustrates the outcome of the clustering for La Ciotat and Table 6 compares the characteristics of the properties in the original scraped data in each cluster and in the choice sets. The properties from a cluster that do not appear in the choice set serve as substitutes in case a property is not found on the website during the experiment. We explain this mechanism in greater detail in the next section.

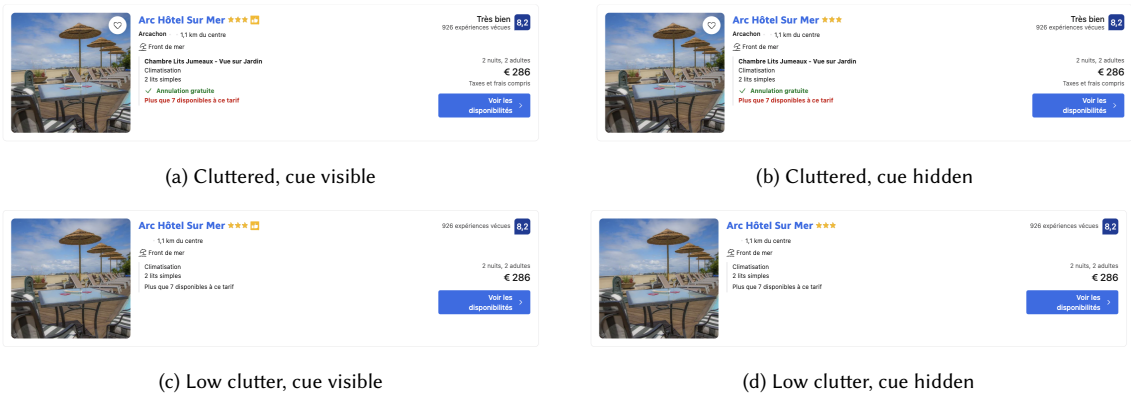
3.2 Web browser extension

We developed a Manifest V3 Chrome extension injected as a content script into the live Booking.com page. The extension was installed on the machine of the laboratory before participants entered. It does mostly two things on the Booking.com website: it rewrites what the participant sees and it records what the participant does. We detail in a third section how it controlled the choice sets that appear on screen. Note that the extension also tracked which URLs were visited and blocked access to domains outside a safelist[5]. Some AI tools were involved in the development of the extension.

Manipulating the Booking.com interface. Every manipulation is subtractive: the extension hides, restyles or disables DOM nodes from Booking.com, but never injects new content. The extension performs two types of changes on the interface.

First, the extension permanently removes or deactivates some elements of the interface for all participants. It removes the map from both the search page and the property pages, it removes filters and carousels from the search page, and

[4] At extension runtime, we displayed all properties below or equal to €450 for two nights to allow for minor price fluctuations.
[5] The safelist corresponds to the list of websites mentioned earlier in the paragraph on the pretask.



(a) Cluttered, cue visible

(b) Cluttered, cue hidden

(c) Low clutter, cue visible

(d) Low clutter, cue hidden

Notes: Each panel shows the same property container as rendered by the extension under one of the four treatment cells of the 2×2 design (clutter × cue visibility). The cluttered/cue-visible cell (a) is the one found on the original Booking.com page; the extension subtracts the cue badge and/or desaturates clutter elements to reach the other three cells. Nine such property listings appear on the page during each choice task, stacked vertically.

Fig. 2. The four possible conditions as rendered by the extension

the types of available bedrooms from the property page[6]. Most of the search buttons of the top panels are deactivated. In particular, participants cannot change the weekend or the city. Appendix A.1 shows the search and property pages before and after these changes.

Removing access to the map view illustrates the general difficulty of isolating one element without disturbing the rest of a page you do not own. Hiding obvious entry points on the search page that subjects are supposed to view stops most attempts. However, by default, Booking keeps the full-screen map overlay mounted underneath its default search results list, so a participant who reaches this default search results page somehow (a missed link, the browser's back button) suddenly has access to the map feature and can navigate away from the experiment. Once participants navigate away, they cannot be easily returned to the search page that they are supposed to be on – only going back to oTree and clicking the link again returns them to the search page. Thus, in addition to disabling the map, we also disabled the page's header, search bar and breadcrumb during our experiment, closing off other paths for a participant to silently navigate off the assigned city and weekend. Besides, allowing a more open form of browsing on the website would be much less practical to analyze quantitatively in terms of attention allocation: for instance, a user can be on the page of one property, but compare the price of others on the map at the top of the page.

Second, the extension implemented changes corresponding to our two treatments. These changes affected mostly the property cards on the search page, but also what is displayed on each listing's specific pages. Because Booking is a client-rendered application that continuously re-renders and lazy-loads cards as the participant scrolls, this is driven by a page observer rather than applied once, so a card that reappears after a re-render is re-classified the instant it lands. The badge also appears on the property page and has to be removed.

Finding a stable hook for each element was rather difficult. Booking's CSS classes are machine-generated and change without notice, so the extension anchors on semantic markers — test identifiers, accessibility roles, visible text — and

[6]In the experiment, the reservation can only be made for the cheapest bedroom offered in the property, without additional payment for breakfast, which corresponds to the price displayed on the search page.

treats none of them as durable: a handful of elements (a location-pin icon, a "view on map" link with no distinguishing attribute) are only identifiable by matching the shape of what they render.

Tracking user behavior. The extension records what participants saw and did. On Booking.com, the extension recorded several families of events covering navigation, timing, interaction, exposure, choice and quality control. Events are buffered on the page and flushed in batches to the extension's background process, which only clears its buffer once a batch is confirmed written. As such, a participant closing a tab mid-session loses at most a few seconds of data rather than the session. The event-log variables from the extension used in the analysis are listed and defined in Table 5, Appendix A.2. The extension also records which URLs were visited outside of the Booking.com website, but not the corresponding events.

Controlling the choice sets. The nine listings assigned to a choice set are fixed before the session. The extension must make sure that these nine listings do appear on screen and the others do not. Running the experiment on the live website, we faced two main challenges.

First, Booking paginates: a search page renders roughly two dozen cards and loads more only as the visitor scrolls. Since the cities considered had hundreds of listings, an assigned listing ranked outside that window is invisible until the extension has driven enough scrolling to surface it. Our extension addresses this point by introducing a loading page hiding the actual search results during the time it takes for the extension to scroll down many times until it finds the nine listings of the choice set. Note that this process takes time: up to two minutes. This time is not counted in the decision time of the subject, since they cannot see anything during that time.

Second, this live inventory of the properties of the page sometimes fails, leading to properties that should be displayed failing to appear for a participant. It could be that a property sold out between our preparatory scraping and the session, but our analyses suggest instead that Booking randomizes some elements affecting property display. To ensure that participants face choice sets satisfying our requirement, we use the pool of substitutes — a byproduct of the choice set creation process described earlier — to replace a missing listing by another from the same cluster and with the same cue status. This replacement takes place either when the extension has finished scrolling down to display the full property inventory, or when a threshold of four minutes of loading has been reached and some of the nine properties are still to be found.

This mechanism worked well during our sessions. There was only one incomplete choice set across our whole experiment.[7] The substitutes proved necessary: of the 739 choice sets delivered, 81 needed one substitute, 0 needed two and 0 needed three or more, with the largest number used in a single set being 1. The one missing property was caused by a choice set with an insufficient number of substitutes. When creating the clusters, the constraint of having a sufficient number of properties with and without the Preferred Partner badge was often binding. Requiring more of each would have artificially distorted the clusters for some city, so some substitute pools were empty, and our sessions happened to feature a failure to load on one of the clusters with empty substitute pools.

None of this is testable in the abstract. We observed that failures to detect a listing in the inventory seem to relate to network conditions: how long a card takes to render, whether a scroll event fires before Booking's own lazy-load, whether a request times out. We had to develop automated tests to be run on the live site to verify that the page elements that the manipulation depends on still matched Booking's current markup and that every one of the sixteen choice

[7] As preregistered, data for the incomplete city's four weekends were removed. However, data for the eight other weekends in the other two cities for the participant who had an incomplete choice set were kept.

sets across all four cities could still be detected. The tests were re-run in the days before, and at points during, data collection, since Booking's availability, markup and inventory could all drift on their own schedule.

3.3 Empirical strategy

This section provides some background on the statistical analyses performed in the Results section.

Hypotheses and preregistration. Our preregistration (which can be found on Open Science Framework) organizes our analysis around three main hypotheses:

- (H1) Badge visibility affects how consumers act with cue-eligible listings.
- (H2) Clutter moderates the effect of the badge, either by making it less salient or by affecting consumers' decision-making mode.
- (H3) Badge visibility and clutter affect consumer welfare, measured by choice consistency across clusters of properties.

The preregistration states each hypothesis, outcome, test direction and exclusion rule. Some AI tools were involved in the implementation of the analysis code from the preregistration.

Causality. In observational data, a listing that carries the badge differs from one that does not in every way that made the platform enroll it, so the correlation between badge and choice mixes the cue's effect with the quality it signals. In our randomized experiment, the design removes this confound by drawing at random, for each participant and city, which two of the four weekends show the badge. Note that the listings differ across weekends for a given participant, but the same choice set appears with the cue for some participants and without for others. In the same way, clutter is randomized at the participant level and the comparison of outcomes between the two clutter treatments has a clear causal interpretation.

Some features of our design may affect the magnitude of the estimates but not their causal interpretation. For instance, the extension randomizes the order in which listings appear on screen. This is orthogonal to cue visibility by construction, so the estimated effect remains a treatment effect; however, the estimated effect is an average over the particular pages participants actually saw, weighted by how often each configuration occurred, and may take a different value under a different ordering or composition.

Choice-set-level outcomes. Outcomes that vary within participant — whether a cue-eligible listing was chosen or clicked, the time spent on cue-eligible detail pages, decision time, page visits — are observed once per choice set s, that is once per participant i, city c and weekend. For a binary outcome y_{is} we estimate the conditional logit

$$\Pr(y_{is} = 1) = \Lambda\big(\beta\,\mathrm{Cue}_{is} + \alpha_i + \gamma_{c(s)}\big), \tag{1}$$

and for a continuous outcome — typically, decision times or time on page — we use the linear analogue with $\log(1 + y_{is})$ as the dependent variable. The participant effect α_i absorbs everything stable about the person, including their clutter condition, and the city effect γ_c absorbs the pool of listings, so β is identified from the contrast between a participant's cue-visible and cue-hidden weekends within a city. Since cue visibility is balanced two-to-two within every participant $\times$ city, this contrast is exact rather than in expectation. The preregistration sets a Type I error rate of 0.05 on the predicted tail for its one-sided tests; the tables report the one-sided p-value and, given the sample size, also mark estimates significant at the 10% level (†), which we read as suggestive.

Subject-level outcomes. Cue recognition and the NASA-TLX subscales are observed once per participant and vary only with clutter, so they are estimated by a logit or OLS of the outcome on the clutter indicator with no fixed effects, one observation per participant. Choice-set-level outcomes whose hypothesis concerns clutter (decision time, page visits) keep the city effect but drop α_i, which would absorb the treatment.

Moderation analysis. To test whether clutter moderates the cue effect, we augment Equation (1) with an interaction:

$$\Pr(y_{is} = 1) = \Lambda\big(\beta\,\mathrm{Cue}_{is} + \delta\,\mathrm{Cue}_{is} \times \mathrm{Clutter}_i + \alpha_i + \gamma_{c(s)}\big)\,. \tag{2}$$

The clutter main effect is absorbed by α_i; β is then the cue effect in the low-clutter condition and δ the difference between the cluttered and low-clutter cue effects. The preregistration is agnostic on the sign of δ, so the test is two-sided.

Welfare. The welfare analysis rests on the assumption that consumers have preferences over clusters. Assume that a consumer prefers Cluster 1 over Cluster 2 over Cluster 3, resulting in a probability $\pi_1 > \pi_2 > \pi_3$ of choosing from each of them in the absence of the badge (with $\pi_1 + \pi_2 + \pi_3 = 1$). Assuming the two choices of the consumer are independent draws from the previous distribution, our ideal measure of preference consistency — defined as the probability that the two choices are from the same cluster — is $\pi_1^2 + \pi_2^2 + \pi_3^2$.

Why would preference consistency be a good indicator of consumer welfare? Consider any alternative way of choosing (with probabilities π_1', π_2' and π_3') that increases the probability of choosing from one cluster and reduces the probability of choosing from another by the same amount. This transformation increases preference consistency if and only if the cluster whose probability was increased is preferred over the other. This is the motivation for comparing preference consistency across the clutter condition. Note that our experiment does not observe this ideal measure of preference consistency, but only a single Bernoulli realization whose parameter is that measure. Moreover, we pool these measures across cities and individuals, even though the underlying choice probabilities may vary.

To analyze the effect of the badge, we note that a participant choosing twice from the same cluster gives a stronger signal about their preferences than them choosing only once from the same cluster. Thus, when a user chooses twice from the same cluster without a badge, we examine their likelihood of choosing from that cluster once a badge is present. Formalizing the properties of the corresponding estimator and using the Delta method leads to the upper and lower thresholds mentioned in the preregistration and used in the analysis to test whether the value is significantly above or below what the null hypothesis ($\pi_1 = \pi_1'$, $\pi_2 = \pi_2'$ and $\pi_3 = \pi_3'$) would predict. This approach is subject to the same limits as reported above.

It should also be mentioned that participants' preferences on clusters may only partially reflect actual preferences, and a mixing of choices from the best and second best preferred clusters may actually hide the fact that some properties from the second best are actually preferred to some of the first best. The main advantage of this approach is that it is extremely simple to compute, it requires few modeling assumptions and it does not assume at any point that preferences over cluster are similar across subjects.

4 Results

Among the 62 participants who undertook the experiment, 2 subjects had to leave before the end of their twelve choices, generating 739 of the 744 theoretically possible choices. The extension displayed all nine listings in 738 of these 739 choice tasks: one task contained eight listings. As preregistered, we completely excluded the data from two participants who could not complete the whole experiment and we removed only the four choices for the city in which one choice

set had failed, leaving 716 complete choice sets for analysis. Unless stated otherwise, every subject-level statistic below is computed on these 60 participants and every choice-set-level one on these 716 choices.

Participants averaged 46.0 years old (SD 17.1, median 48.0); 51.7% identified as women and 48.3% as men. 18.3% were students, and 96.7% lived in the Île-de-France region. Mean self-reported familiarity with Booking.com was 2.97 on the 1–4 scale (SD 0.96). 43.5% failed at least one of the three comprehension checks (whether they would receive €400 on their bank account: 29.0%; whether they will be able, if they win, to choose the city and weekend of the reservation: 40.3%; whether the reservation can be canceled: 16.1%); per the preregistration, no exclusion was applied on this basis, but they could not move to the next page until they entered the right answer, so they raised their hand and we re-explained the corresponding part.

In the pretask, each participant had to select three of the four candidate cities and four of the nine candidate weekends. Le Tréport, the only destination in the north of France, was the most often excluded (38.7% of subjects), followed by La Ciotat (25.8%) and Sète (24.2%), with Arcachon the most attractive (11.3%). The All Saints' weekend (corresponding to the end of the fall school holidays) and the last two proposed weekends were chosen less often than the others.

The choice-set page took on average 126s to load the nine listings on screen (median 104s, SD 56s; $N = 739$ subject × weekend observations). Decision time — from the choice set becoming visible to the reservation click — fell over the course of a session. Since each participant went through their three cities in order, the twelve choices form three blocks of four: the median decision time was 80s over the first four choices, 59s over the next four and 46s over the last four.

The full session lasted 74.4 minutes on average (median 72.2, SD 10.8), of which 46.7 minutes (median 44.4, SD 10.8) were spent on the twelve choice tasks themselves. We adjusted the participation fee from €10 to €12 between the first session and the next two when realizing that the session took longer than the expected one hour. We also warned participants about the duration when inviting them to the lab for the last two sessions, which did not result in any dropout participants.

We report below the main pre-registered analyses of the study: first the effect of the badge, then the potential role of clutter as a moderator and the associated mechanisms and finally the welfare effects. We conclude with some exploratory analysis of consumers' perception of the badge and the task in general, to see how it relates to the causal effects we measure.

4.1 H1 — Cue effects on choice and attention

We first examine the effect of the badge on the probability of choosing a property, clicking to open its property page and spending time on the corresponding page. Table 1 reports fixed-effects (subject and city) models of choice and attention as a function of cue visibility: logit for whether the cue-eligible listing was chosen or clicked, OLS for log time on its detail page. All tests are one-sided ($H_1 : \beta > 0$), as preregistered.

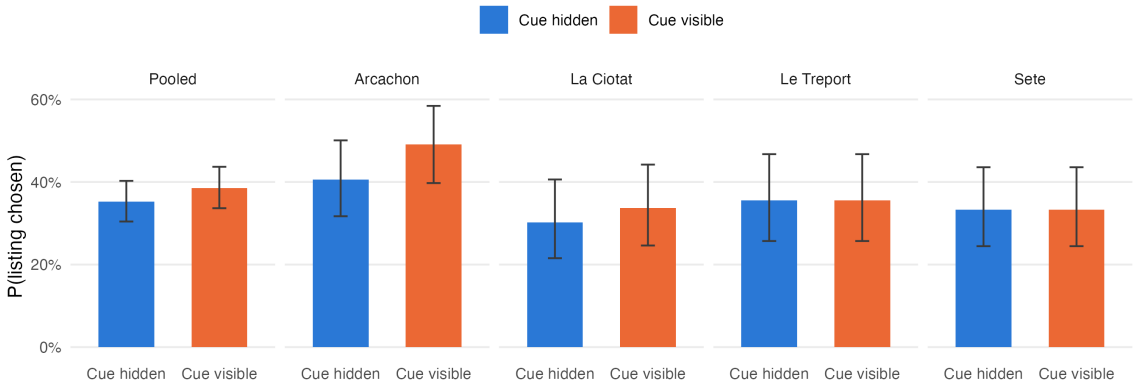
All three point estimates are positive. The cue raises the probability that at least one cue-eligible listing's detail page is opened, significant at the 10% level; the estimates for time on that page and for the final choice are smaller relative to their standard errors and not significant. Figure 3 illustrates the heterogeneity of these results across cities. The effect of the badge seems larger in Arcachon and La Ciotat than elsewhere. This could be because the effect is stronger when consumers are more engaged, since Arcachon is the city that was the most selected in the pretask. This could also be driven by the fact that the composition of cue-eligible listings varies across cities and that these characteristics interact with the effect of the badge.

Figure 4 gives the same decomposition for the click outcome, distinguishing between the cue-eligible and the non-eligible listings of each set. Unlike the choice outcome, for which the probability of choosing a non-cue-eligible

	Chosen (logit)	Clicked (logit)	Time on page (log) (OLS)
Cue visible	0.158	0.265†	0.082
	(0.170)	(0.171)	(0.091)
p (one-sided)	0.176	0.060	0.183
Subject FE	Yes	Yes	Yes
City FE	Yes	Yes	Yes
N	716	704	716

Notes: Logit for Chosen and Clicked, OLS for log time on page. Standard errors in parentheses; preregistered one-sided tests ($H_1 : \beta > 0$). N counts subject × weekend choice sets; the Clicked column drops the one subject whose outcome never varies. $^{\dagger}p < 0.1$, $^{*}p < 0.05$.

Table 1. Cue effects on choice and attention (H1).



Notes: Share of choice sets in which the chosen listing was cue-eligible – i.e. carried the Preferred Partner badge on the unmodified site – the Chosen outcome of Table 1, with 95% Wilson score confidence intervals over sets, split by whether the badge was actually visible that weekend, pooled across cities and separately for each. Sample sizes (subject × weekend sets): Pooled 716; Arcachon 212; La Ciotat 172; Le Tréport 152; Sète 180.

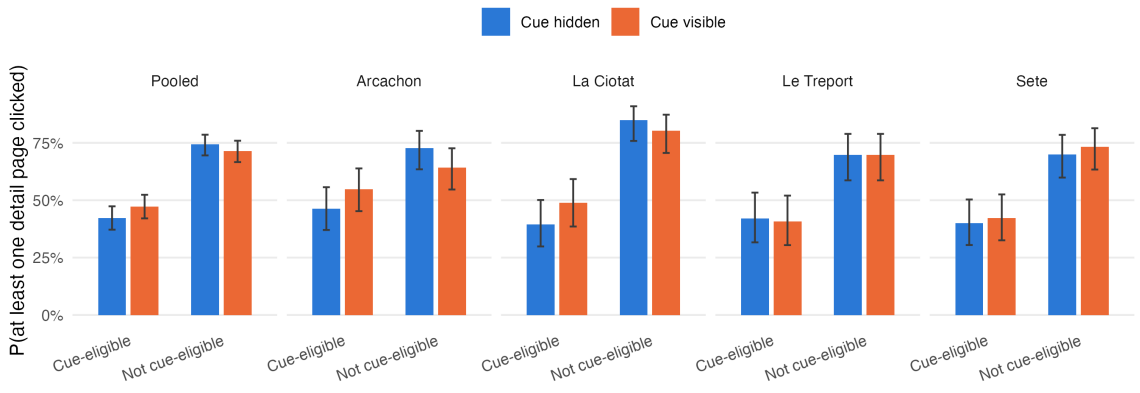
Fig. 3. Choice probability of cue-eligible listings by cue visibility

listing is just the complement to that of choosing a cue-eligible listing, the non-eligible group is informative here. The figure shows whether the badge redirects attention from non-eligible to eligible listings or simply adds to eligible listings' attention. In terms of clicks, it seems that the increase for cue-eligible listings caused by the badge induced a corresponding decrease for non-cue-eligible listings.

4.2 H2 — Clutter as moderator and associated mechanism

We now analyze the effect of the second treatment, the level of clutter, and its potential moderating effect on the badge.

Internal validity check. The first thing to report is that our internal validity check on the clutter treatment failed. Self-reported visual complexity did not differ significantly between the Cluttered ($M = 2.59$) and Low Clutter ($M = 2.90$) groups ($p = 0.516$), which means the clutter manipulation was not perceived as intended, and if anything, clutter was



Notes: Share of choice sets in which at least one listing of the group had its detail page opened – for cue-eligible listings, the Clicked outcome of Table 1 – with 95% Wilson score confidence intervals over sets, by whether the listings carried the Preferred Partner badge on the unmodified site ("cue-eligible") and whether the badge was actually visible that weekend, pooled across cities and separately for each. Sample sizes (subject × weekend sets): Pooled 716; Arcachon 212; La Ciotat 172; Le Tréport 152; Sète 180.

Fig. 4. Click probability by cue-eligibility and cue visibility

increased rather than decreased by our treatment. This surprising result seems to be driven by younger subjects and by those answering fastest. Splitting the sample at the median age, participants aged 48 or younger rated the cluttered interface as *less* complex than the low-clutter one (2.16 vs. 3.07), others rated it as more complex (3.40 vs. 2.76). Likewise, the quartile of participants with the shortest median decision time rated the cluttered interface as less complex (2.12 vs. 3.14). It could be that the statement of the question in the internal validity check rather than the treatment itself was faulty, but we do not have an alternative measure of visual complexity to compare it to. All this suggests that the following results should be read with caution.

Moderation. Table 2 reports the estimates of Equation (2), which augments each H1 model with the cue × clutter interaction (Section 3.3). None of the interaction terms are statistically significant. The point estimates suggest that clutter may have increased the effect of the badge on choice ($\delta = 0.277$, $p = 0.417$) and reduced its effect on time spent on the detail page ($\delta = -0.106$, $p = 0.561$), whereas the null interaction on clicks is sharp ($\delta = 0.022$, $p = 0.948$).

Overall, this suggests that the effect of the badge is mostly unaffected by the presence of clutter. If this conclusion were confirmed with a larger sample size, it would be a good signal regarding the external validity of artificial visual environments. Indeed, this would suggest that an additional degree of realism in the appearance of the choice interface is not needed to assess the magnitude of the effect of the Preferred Partner badge of Booking.com. However, note that an experiment can only test whether a specific variation in the visual environment moderates the effect (the precise changes made in our low clutter condition), it cannot rule out that some other changes would moderate the effect.

Clutter effect on subject-level outcomes. We analyze some subject-level variables to understand the mechanisms behind the ambiguous results on the moderation effect. Table 3 reports the effect of clutter on cue recognition, decision time and search intensity, with the preregistered one-sided tests.

	Chosen (logit)	Clicked (logit)	Time on page (log) (OLS)
Cue visible	0.025	0.254	0.133
	(0.236)	(0.237)	(0.126)
Cue × Clutter	0.277	0.022	-0.106
	(0.341)	(0.342)	(0.182)
p (two-sided, interaction)	0.417	0.948	0.561
Subject FE	Yes	Yes	Yes
City FE	Yes	Yes	Yes
N	716	704	716

Notes: Table 1 models augmented with a cue × clutter interaction; subject and city fixed effects absorb the clutter main effect. Standard errors in parentheses; preregistered two-sided tests on the interaction. N as in Table 1. $^{\dagger}p < 0.1$, $^{*}p < 0.05$.

Table 2. Moderation analysis: Cue × clutter interaction.

One could think that clutter makes the environment so complex that the badge becomes harder to notice. Results show that clutter does not seem to reduce recognition of the badge (62.1% of High Clutter subjects reported noticing it versus 61.3% of Low Clutter subjects, one-sided $p = 0.525$), so we find no support for this dilution mechanism.

We had also hypothesized that clutter could increase the complexity of the task and trigger a more heuristic-based mode of decision involving less effort and a lower expected performance. The decision-mode hypothesis predicted faster decisions and fewer page visits under clutter; the estimates run the other way. Participants on the cluttered interface opened more listings per set (two-sided $p < 0.001$) and took longer to decide (two-sided $p = 0.070$), which is why the one-sided p-values in the table are close to one. We explore this aspect further using the NASA-TLX subscales. Table 7 in Appendix A.4 reports the same clutter effect on the six NASA-TLX subscales collected in the post-experiment questionnaire. Overall, it seems difficult to argue that the clutter condition significantly changed the decision mode of the subjects.

	Cue recognized (logit)	Decision time (log) (OLS)	Listing page visits (OLS)
Clutter	0.033	0.100	0.421
	(0.531)	(0.055)	(0.118)
Constant	0.460	4.153	1.627
	(0.369)	(0.057)	(0.122)
p (one-sided)	0.525	0.965	1.000
City FE	No	Yes	Yes
N	60	716	716

Notes: Logit for cue recognition (one observation per subject), OLS with city fixed effects for log decision time and page visits. The constant is the low-clutter level (reference city: Arcachon). Standard errors in parentheses; preregistered one-sided tests ($H_1 : \beta < 0$). $^{\dagger}p < 0.1$, $^{*}p < 0.05$.

Table 3. Clutter effects on badge recognition and search patterns.

4.3 H3 — Welfare effects

In this last preregistered series of analyses, we use our clusters of properties to investigate whether the badge or its combination with the clutter seems to harm consumers.

Internal validity check. Our welfare measure rests on preference consistency — whether a participant's two no-cue choices in a city fall in the same cluster. Before turning to the hypotheses, we must check that our clusters meaningfully capture some consumer preferences. Preference consistency across all subject × city pairs was 41.9% ($N = 179$), significantly above the one-third chance benchmark ($p = 0.011$). As preregistered, we consider that our clusters indeed relate to consumers' preferences.

Clutter effect. We now examine whether clutter affects the ability of the subject to be consistent in their choices. Preference consistency does not differ significantly between clutter conditions (logit, $\beta = 0.178$, two-sided $p = 0.562$); the point estimate is positive, i.e. slightly more consistent choices under clutter, the opposite of the preregistered expectation.

Badge effect. Turning to the effect of the badge, among the 75 subject × city pairs with consistent preferences, 40.0% of cued choices matched the preferred cluster, against an overall consistency rate of 41.9%. This is significantly below the preregistered threshold for detecting an increase in welfare (0.647, $p < 0.001$) and not significantly different from the preregistered threshold for a decrease (0.374, $p = 0.635$). Therefore, we consider the welfare result as ambiguous under the preregistered rule. Given that we can confidently reject the possibility that the consistency rate is above the pre-registered threshold, increasing the sample size should help settle between a sharp zero and a slightly negative effect in terms of welfare.

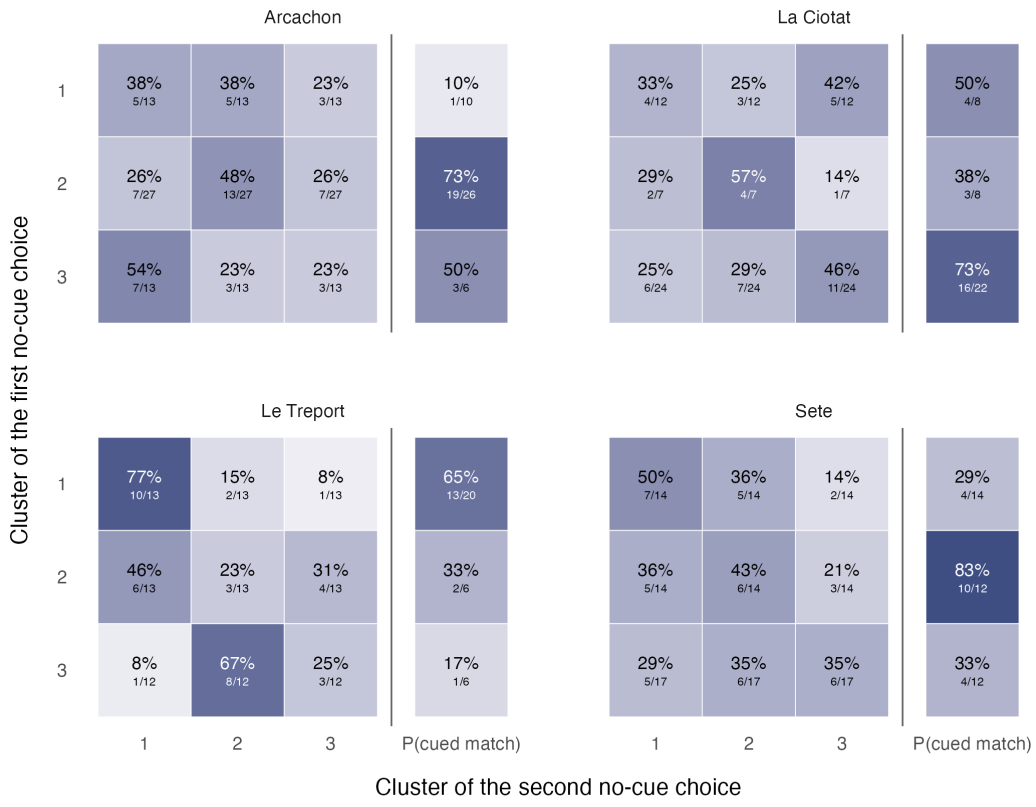
Heterogeneity across cities. Figure 5 illustrates the heterogeneity of the results across cities. The 3×3 matrix corresponds to choices when the badge is not visible: the row indicates the cluster of the first choice and the second splits this population depending on the cluster in their second choice. The diagonal relates to our consistency measure. The fact that the clusters are meaningful is reflected in the high diagonal values for a given row: they indicate that those choosing from one cluster once were indeed more likely to choose a second time from it.

The vertical column to the right of that matrix reports the share of choices made when the badge was visible that were consistent with the choices made when it was not, only for the subjects who were consistent in the absence of the cue — those on the diagonal. This value can be compared to that of 0.33 which would be obtained when choosing at random and to the value of the diagonal which was obtained when the badge was not visible and which conditioned on a single choice from that cluster — rather than two. When the value in the column is lower than that of the corresponding diagonal, it suggests that the badge had a negative effect on consumer welfare — at least, analyzed at the cluster level. For instance, this is the case for two of the three clusters in Le Tréport.

4.4 Consumer beliefs and perceptions

We analyze below how consumers' perceptions relate to the causal effects measured above. Note that the analyses below were not preregistered, in part because they rely on the open questions of the questionnaire.

Open answers. Table 4 summarizes the answers to the four open questions of the questionnaire. They were coded with keyword rules into overlapping categories — some AI tools were used to design the categories and classify the text.



Notes: One panel per city (N = subject × city pairs: Arcachon 53; La Ciotat 43; Le Tréport 38; Sète 45). Rows: cluster of a participant's first no-cue choice in the city; columns 1–3: cluster of their second no-cue choice. Each cell gives the share of participants with that first choice whose second choice fell in the column's cluster, and below it the count X/Y: X participants with that pair of choices out of Y with that first choice; rows sum to 100% and the diagonal is preference consistency. The rightmost column, set apart by a vertical rule, covers participants whose two no-cue choices agreed on the row's cluster: it gives the share of their cue-visible choices that landed in that same cluster, with the matching choices out of all their cue-visible choices below. In this column, the denominator is twice the number of subjects in the numerator of the corresponding diagonal cell of the matrix, because there are two choices with visible badge for each of these subjects.

Fig. 5. Cluster of the second no-cue choice given the first, and cued-choice match, by city

Badge recognition. 37 participants reported having noticed the thumbs-up badge. The recognition item also listed two decoy icons that never appeared on the site: 39 participants reported noticing at least one of them. We note that only 8 (13.3%) answered the item entirely correctly, reporting the badge and neither decoy. The binary badge recognition measure should therefore be read as noisy. We also report contrasts on the stricter, fully-correct recognition dummy.

Self-reported belief about the badge and the individual-level cue effect. For each subject, we define the magnitude of the cue effect on choice as the within-subject difference, across their twelve weekends, in the share of cue-eligible listings

Category	*N*	%
What does the thumbs-up icon mean?		
Other customers' reviews or ratings	24	40
Generic quality or recommendation, no source	27	45
Endorsement by the platform	6	10
of which paid placement or promotion	1	2
Value for money or a promotion	4	7
Does not know	3	5
Other	1	2
Did you use the icon, and how?		
Did not use it (any of the three below)	46	77
Had not noticed it	13	22
Noticed it but did not use it	16	27
Bare "no"	17	28
Used it	11	18
as a quality signal	11	18
as a prompt to open the listing or its reviews	6	10
Did not understand it	3	5
Distrusts such icons	2	3
How did you choose? (criteria mentioned)		
Location	46	77
Price or budget	30	50
Ratings or reviews	21	35
Photos or look	24	40
Amenities, size or comfort	27	45
The badge or any icon	1	2
Comments or suggestions?		
Pages loaded slowly or session too long	9	15
Positive about the experience	26	43

Notes: Keyword coding of the open answers of the 60 participants; categories within a question may overlap, so shares need not sum to 100. Indented rows are subsets of the row above.

Table 4. Coded open answers from the post-experiment questionnaire.

chosen when the badge was visible versus hidden. The post-experiment belief item ("properties with the thumbs-up badge correspond more to my preferences than others," 1–4 scale) correlates weakly with this individual-level magnitude of the cue effect (Spearman $\rho = 0.177$, $p = 0.176$).

How participants describe their choice. When participants were asked to respond open-endedly about how they chose, which was asked before the badge was ever mentioned in the questionnaire, 77% mention location (distance to the center, the beach, the station or shops), 50% the price or budget, 35% ratings or reviews, 40% photos or the look of the property, and 45% amenities, size or comfort. The influence of the badge, or of any icon, is only mentioned by a single participant.

Subjective meaning of the badge. The badge is awarded to properties that pay Booking.com a higher commission under its Preferred Partner Program. Almost no participants knew this: a single subject mentioned paid placement or promotion, and only 10% attribute the badge to the platform at all ("recommended by the site").[8] By contrast, 40% read it as an aggregate of other customers' reviews or ratings ("approved by previous travelers", "positive reviews above 8/10"), and a further 45% describe it as a generic quality or recommendation label with no source ("a good property", "trustworthy"). Attributing the badge to the platform is the one reading that goes with answering the recognition item entirely correctly (50% of those who do so, against 9% of the others, $p = 0.027^*$). The social-proof reading is unrelated to the belief item, familiarity with Booking.com, badge recognition, the clutter condition and the individual cue effect (all $p > 0.4$). The generic-quality reading is likewise unrelated to these variables (all $p \geq 0.10$), except for an association with the individual cue effect ($p = 0.029^*$).

Self-reported use of the badge. Asked whether they used the icon, 77% say they did not, split about evenly between those who had not noticed it, those who noticed it but did not use it, and a bare "no"; 18% say they used it, mostly as a quality signal ("I prefer properties with a thumb and usually book them"), some as a prompt to open the listing or its reviews, in line with the significant effect of the badge on clicks. Self-reported use goes with a higher belief score (3.09 versus 2.43, $p = 0.042^*$) and with recognizing the badge (91% versus 55%, $p = 0.039^*$), but not with familiarity with Booking.com, the clutter condition or the individual cue effect.

5 Discussion and conclusion

We set out to provide a complete methodological pipeline for regulators to learn about the effects that an interface element has on consumers in the OTA industry using a browser-based field experiment. We applied our approach to Booking.com and tested it in an experiment on the live site, with real prices and a real booking at stake. We created clusters of properties that turned out to be related to consumer preferences. We imposed more control on the choice sets than earlier live client-side experiments. We made use of that additional control to investigate questions about consumer welfare without the use of a structural model. For these reasons, we consider that the experiment was successful from a methodological viewpoint.

The findings about the interface are subject to caution, given our sample size, and should be read as indicative. This is particularly true for the clutter manipulation, which did not pass our pre-registered internal validity check of increasing the self-reported visual complexity of the website. Given our sample size, these null results on clutter effects may not be sharp. The findings are also specific to the counterfactuals we chose — our notions of low clutter, the badge, and its removal. The data suggests that participants respond, at least in clicks, to a cue whose meaning they almost uniformly misread as a summary of other customers' opinions, and that they do so whether or not they report having used it. The badge having a causal effect without consumers having a clear understanding of its purpose calls for further investigation by regulators.

The experimental method that we propose has its own limits. To avoid deceiving participants, we never display the badge on a property that did not carry it on the original website, so we only measure the effect of the badge on properties that had it on the website. Choices are forced: participants cannot decline to book, and there is no opt-out option. Changing slightly the design, for instance offering €200 with no obligation to travel as an alternative, may have cost observed choices and weakened an identification that rests on two badge-visible and two badge-hidden weekends

[8] The badge's explanatory tooltip, which spells out the Preferred Partner meaning, was opened at least once by 2 participants; of the 2 who answered the meaning question, 2 described a generic quality label and 0 mentioned payment.

per city. We cannot really capture learning effects and adaptation to the visual environment in a single session of twelve choices. The fact that our experiment took place on a computer in the lab, rather than at home or on a phone, limits its external validity. Asking consumers to activate the extension on their own, however, would likely be too complex. Finally, we had to remove some elements of the interface which proved overly difficult to manipulate and to track using the extension, such as the map of the listed properties, even though they may be quite attractive to users in that industry.

Regulators are sometimes reluctant to undertake audits based on experiments because of the time and cost of their implementation. In the case of our experiment, finding a suitable experimental design and developing the browser extension indeed took some time, with the choice-set machinery and its automated tests taking an especially large amount of time. Outcomes are also specific to the website under study. However, this effort should be seen as an investment rather than a cost, since both the design and the extension can be reused – after minimal changes – to test other counterfactuals on the same website in the future. Maintenance of the extension as the website evolves may also take time. However, this takes much less effort than the initial development of an extension, and both durations have shrunk dramatically with the advent of coding agents. A session of twenty participants takes approximately two and a half hours for experimenters to run. The total cost of the experiment was around €30 per participant, including the participation fee of €12, the probability of winning the €400 prize and some fees charged by the laboratory. This cost would be closer to €20 per participant if the experiment were scaled up, with a 1% probability of earning the prize and saving on fixed lab costs. Reusing the same design could help anticipate effect sizes and better adjust the sample size to the question addressed.

Enforcement of existing consumer law is sometimes the one point of consensus between consumer organizations, national authorities, and e-commerce firms [18]. Browser-based experiments appear well-suited to study the interface of the very large online platforms (like Booking.com) placed under specific scrutiny by the EU's Digital Services Act. Maintaining one such design — an extension, a recruitment and incentive protocol, a method for building clusters and choice sets — for some of the twenty-three designated services would be an affordable way for the authorities enforcing the Act to level the playing field with firms that run A/B tests at scale on their own sites every day.

6 Acknowledgments

[MASKED FOR ANONYMIZATION]

Ethics and Privacy Statement

The study was approved by the institutional review board of [MASKED FOR ANONYMIZATION]. Participants gave informed consent to join the experiment. They agreed that their personal data be handled by the research team in accordance with the EU General Data Protection Regulation, without direct identifiers after the lottery has been run, and were informed about their rights. The experiment does not involve deception: everything the subjects were told was correct. The extension never added false claims about a property. A reservation on Booking.com was made for one of the participants drawn at random as planned in the experimental design. Participants who could not complete the whole experiment still took part in the lottery for the reservation, weighted by the number of choices they had the time to make, as preregistered and agreed in advance with the lab manager. The scraping of Booking.com was done by a research organization after the IRB approval. Since the experiment on the live interface was performed client-side and the participants were informed that they took part in an experiment, it is unlikely to have harmed the website under study.

References

[1] Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The Welfare Effects of Social Media. *American Economic Review* 110, 3 (March 2020), 629–676. doi:10.1257/aer.20190658

[2] Hunt Allcott, Matthew Gentzkow, and Lena Song. 2022. Digital Addiction. *American Economic Review* 112, 7 (July 2022), 2424–2463. doi:10.1257/aer.20210867

[3] Eytan Bakshy, Dean Eckles, and Michael S. Bernstein. 2014. Designing and deploying online field experiments. In *Proceedings of the 23rd international conference on World wide web*. ACM, Seoul, Republic of Korea, 283–292. doi:10.1145/2566486.2567967

[4] Jan Panero Benway. 1998. Banner Blindness: The Irony of Attention Grabbing on the World Wide Web. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 42, 5 (1998), 463–467. doi:10.1177/154193129804200504

[5] Francesco Bogliacino, Leonardo Pejsachowicz, Giovanni Liva, and Francisco Lupiáñez-Villanueva. 2026. The transaction test: An experimental method for assessing online interfaces. *Journal of Economic Behavior & Organization* 244 (April 2026), 107469. Previously SSRN 10.2139/ssrn.5217551 (2025). doi:10.1016/j.jebo.2026.107469

[6] Kerstin Bongard-Blanchy, Arianna Rossi, Salvador Rivas, Sophie Doublet, Vincent Koenig, and Gabriele Lenzini. 2021. "I am Definitely Manipulated, Even When I am Aware of it. It's Ridiculous!" - Dark Patterns from the End-User Perspective. In *Designing Interactive Systems Conference 2021*. Association for Computing Machinery, New York, NY, USA, 763–776. doi:10.1145/3461778.3462086

[7] Moira Burke, Nicholas Gorman, Erik Nilsen, and Anthony Hornof. 2004. Banner ads hinder visual search and are forgotten. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1139–1142. doi:10.1145/985921.986008

[8] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3290605.3300733

[9] Scott Carter, Jennifer Mankoff, Scott R. Klemmer, and Tara Matthews. 2008. Exiting the Cleanroom: On Ecological Validity and Ubiquitous Computing. *Human–Computer Interaction* 23, 1 (2008), 47–99. doi:10.1080/07370020701851086

[10] Daniel L. Chen, Martin Schonger, and Chris Wickens. 2016. oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* 9 (2016), 88–97. doi:10.1016/j.jbef.2015.12.001

[11] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a replication crisis in empirical computer science. *Commun. ACM* 63, 8 (2020), 70–79. doi:10.1145/3360311

[12] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal, QC, Canada, 1–12. doi:10.1145/3173574.3173715

[13] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112. doi:10.1515/popets-2015-0007

[14] Thomas de Haan and Jona Linde. 2018. 'Good Nudge Lullaby': Choice Architecture and Default Bias Reinforcement. *The Economic Journal* 128, 610 (May 2018), 1180–1206. doi:10.1111/ecoj.12440

[15] Linda Di Geronimo, Larissa Braz, Enrico Fregnan, Fabio Palomba, and Alberto Bacchelli. 2020. UI Dark Patterns and Where to Find Them. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu, HI, USA, 1–14. doi:10.1145/3313831.3376600

[16] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Human–Computer Interaction Series*. Springer International Publishing, Cham, Switzerland, 291–330. doi:10.1007/978-3-319-26633-6_13

[17] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior Around Them in Rating Platforms. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (2017), 62–71. doi:10.1609/icwsm.v11i1.14898

[18] European Commission. 2025. *European Consumer Summit 2025 Report*. Technical Report. European Commission, Directorate-General for Justice and Consumers, Brussels. https://commission.europa.eu/document/download/6a842fae-e0f5-4d4e-8d91-c506ffa1ce38_en?filename=2025+Consumer+Summit+report+-+FINAL+11.07+-+15.04.pdf

[19] Chiara Farronato, Andrey Fradkin, and Chris Karr. 2024. *Webmunk: A New Tool for Studying Online Behavior and Digital Platforms*. NBER Working Paper 32694. National Bureau of Economic Research, Cambridge, MA. Working paper; no published version as of September 2026. doi:10.3386/w32694

[20] Chiara Farronato, Andrey Fradkin, and Tesary Lin. 2025. *Designing Consent: Choice Architecture and Consumer Welfare in Data Sharing*. NBER Working Paper 34025. National Bureau of Economic Research, Cambridge, MA. Working paper; no published version as of September 2026. A one-page abstract appeared at ACM EC '25. doi:10.3386/w34025

[21] Chiara Farronato, Andrey Fradkin, and Alexander MacKay. 2023. Self-Preferencing at Amazon: Evidence from Search Rankings. *AEA Papers and Proceedings* 113 (May 2023), 239–243. doi:10.1257/pandp.20231068

[22] Chiara Farronato, Andrey Fradkin, and Alexander MacKay. 2025. *Vertical Integration and Consumer Choice: Evidence from a Field Experiment*. NBER Working Paper 34135. National Bureau of Economic Research, Cambridge, MA. Working paper; no published version as of September 2026. doi:10.3386/w34135

[23] Anindya Ghose, Panagiotis G. Ipeirotis, and Beibei Li. 2014. Examining the Impact of Ranking on Consumer Behavior and Search Engine Revenue. *Management Science* 60, 7 (2014), 1632–1654. doi:10.1287/mnsc.2013.1828

[24] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/

3173574.3174108

[25] Colin M. Gray, Thomas Mildner, and Ritika Gairola. 2025. Getting Trapped in Amazon's "Iliad Flow": A Foundation for the Temporal Analysis of Dark Patterns. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama, Japan, 1–10. arXiv:2309.09635v2. doi:10.1145/3706598.3713828

[26] Colin M. Gray, Cristiana Santos, Nataliia Bielova, Michael Toth, and Damian Clifford. 2021. Dark Patterns and the Legal Requirements of Consent Banners: An Interaction Criticism Perspective. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3411764.3445779

[27] Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, and Thomas Mildner. 2024. An Ontology of Dark Patterns Knowledge: Foundations, Definitions, and a Pathway for Shared Knowledge-Building. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–22. arXiv:2309.09640v1. doi:10.1145/3613904.3642436

[28] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*. ACM, Vancouver, BC, Canada, 305–318. doi:10.1145/2663716.2663744

[29] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Elsevier, Amsterdam, 139–183. doi:10.1016/s0166-4115(08)62386-9

[30] Matthias Hunold, Reinhold Kesler, and Ulrich Laitenberger. 2020. Rankings of Online Travel Agents, Channel Pricing, and Consumer Protection. *Marketing Science* 39, 1 (2020), 92–116. doi:10.1287/mksc.2019.1167

[31] Bright Isaac Ikhenaode. 2025. Prominence and Pricing: Evidence from Booking.com Preferred Partner Program. (2025). Working paper; no published version as of September 2026; the SSRN posting was withdrawn by September 2026. doi:10.2139/ssrn.5286661

[32] Jesper Kjeldskov and Mikael B. Skov. 2007. Studying Usability In Sitro: Simulating Real World Phenomena in Controlled Environments. *International Journal of Human-Computer Interaction* 22, 1-2 (2007), 7–36. doi:10.1207/s15327590ijhc2201-02_2

[33] Jesper Kjeldskov and Mikael B. Skov. 2014. Was it worth the hassle?. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. ACM, Toronto, ON, Canada, 43–52. doi:10.1145/2628363.2628398

[34] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Chicago, IL, USA, 1168–1176. doi:10.1145/2487575.2488217

[35] Ron Kohavi, Alex Deng, Roger Longbotham, and Ya Xu. 2014. Seven rules of thumb for web site experimenters. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. Association for Computing Machinery, New York, NY, USA, 1857–1866. doi:10.1145/2623330.2623341

[36] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (2014), 8788–8790. doi:10.1073/pnas.1320040111

[37] Markus Lill, Nastasia Gallitz, Lucas Stich, and Martin Spann. 2026. *How Platform Endorsement Shapes Consumer Search and Choice in Online Retail*. Rationality and Competition Discussion Paper Series 569. CRC TRR 190 Rationality and Competition. Working paper; no published version as of September 2026. https://rationality-and-competition.de/wp-content/uploads/discussion_paper/569.pdf

[38] Gitte Lindgaard, Gary Fernandes, Cathy Dudek, and J. Brown. 2006. Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & Information Technology* 25, 2 (2006), 115–126. doi:10.1080/01449290500330448

[39] Jamie Luguri and Lior Jacob Strahilevitz. 2021. Shining a Light on Dark Patterns. *Journal of Legal Analysis* 13, 1 (2021), 43–109. doi:10.1093/jla/laaa006

[40] Chenchen Mao, Hanjing Shi, Haiyan Jia, Daniel Unguryan, Eric Baumer, and Dominic DiFranzo. 2026. Outer Limits: An Experimental Approach to Controlled Content Manipulation within the Reddit Interface. Working paper; no published version as of September 2026. doi:10.48550/arxiv.2608.10115

[41] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Eli Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark Patterns at Scale. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–32. arXiv:1907.07032v2. doi:10.1145/3359183

[42] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama, Japan, 1–18. arXiv:2101.04843v1. doi:10.1145/3411764.3445610

[43] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human–Computer Interaction* 14, 4 (2021), 272–344. doi:10.1561/1100000083

[44] Milica Milosavljevic, Vidhya Navalpakkam, Christof Koch, and Antonio Rangel. 2012. Relative visual saliency differences induce sizable bias in consumer choice. *Journal of Consumer Psychology* 22, 1 (Jan. 2012), 67–74. doi:10.1016/j.jcps.2011.10.002

[45] Aliaksei Miniukovich and Antonella De Angeli. 2014. Quantification of interface visual complexity. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. Association for Computing Machinery, New York, NY, USA, 153–160. doi:10.1145/2598153.2598173

[46] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2600–2606. doi:10.1073/pnas.1708274114

[47] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu, HI, USA, 1–13. arXiv:2001.02479v1. doi:10.1145/3313831.3376321

[48] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716. doi:10.1126/science.aac4716

[49] Tiziano Piccardi, Martin Saveski, Chenyan Jia, Jeffrey T. Hancock, Jeanne Tsai, and Michael Bernstein. 2026. Reranking Social Media Feeds: A Practical Guide for Field Experiments. *ACM Transactions on Social Computing* 9, 1 (2026), 1–17. doi:10.1145/3800557

[50] Elena Reutskaja, Rosemarie Nagel, Colin F. Camerer, and Antonio Rangel. 2011. Search dynamics in consumer choice under time pressure: An eye-tracking study. *American Economic Review* 101, 2 (2011), 900–926.

[51] Ruth Rosenholtz, Yuanzhen Li, Jonathan Mansfield, and Zhenlan Jin. 2005. Feature congestion. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 761–770. doi:10.1145/1054972.1055078

[52] Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. 2007. Measuring visual clutter. *Journal of Vision* 7, 2 (2007), 17. doi:10.1167/7.2.17

[53] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry, 64th Annual Meeting of the International Communication Association*. International Communication Association, Seattle, WA, USA.

[54] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-Positive Psychology. *Psychological Science* 22, 11 (2011), 1359–1366. doi:10.1177/0956797611417632

[55] Anne M. Treisman and Garry Gelade. 1980. A feature-integration theory of attention. *Cognitive Psychology* 12, 1 (1980), 97–136. doi:10.1016/0010-0285(80)90005-5

[56] Alexandre N. Tuch, Javier A. Bargas-Avila, Klaus Opwis, and Frank H. Wilhelm. 2009. Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *International Journal of Human-Computer Studies* 67, 9 (2009), 703–715. doi:10.1016/j.ijhcs.2009.04.002

[57] Alexandre N. Tuch, Eva E. Presslaber, Markus Stöcklin, Klaus Opwis, and Javier A. Bargas-Avila. 2012. The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International Journal of Human-Computer Studies* 70, 11 (2012), 794–811. doi:10.1016/j.ijhcs.2012.06.003

[58] Raluca M. Ursu. 2018. The Power of Rankings: Quantifying the Effect of Rankings on Online Consumer Search and Purchase Decisions. *Marketing Science* 37, 4 (2018), 530–552. doi:10.1287/mksc.2017.1072

A Research Methods

A.1 The Booking.com interface as it appeared in the experiment

Figures 6 and 7 show, for the search page and for a single property page, the unmodified Booking.com interface next to the version participants saw. On the search page, the extension removes the map, the left-hand filter and budget-slider sidebar, the sort-by dropdown, the commission-ranking notice, and the list/mosaic view toggle; on the property page, it removes the "Infos et tarifs" ("Infos and price") tab, the "voir la carte" ("show on map") location link, and the map thumbnail, and folds the price directly into the reservation button. Participants cannot re-sort, re-filter, or otherwise reconstruct a ranking or map view of the choice set once it is assigned to them.

A.2 Variables

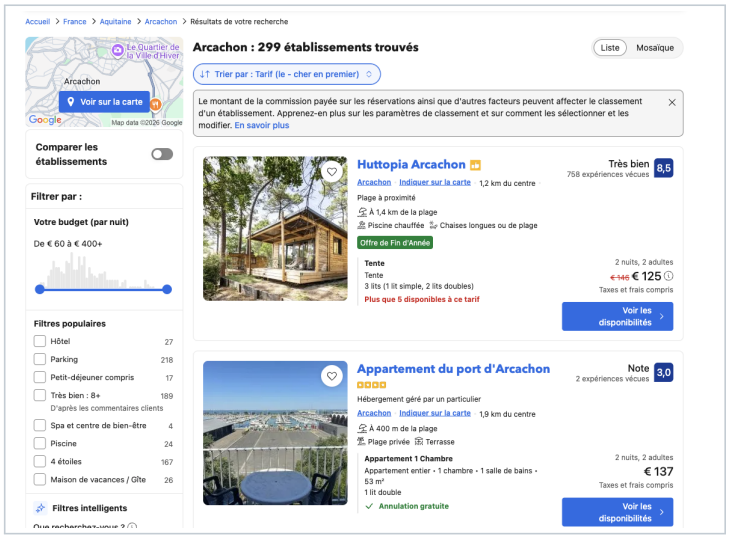
The analysis dataset is assembled from three sources: the extension's event log, the oTree database (assignment, pretask and questionnaire), and the scrape from which the choice sets were built. Table 5 defines every variable used in Section 4 and the level at which it is observed.

A.3 Scraping and choice set construction for the city of La Ciotat

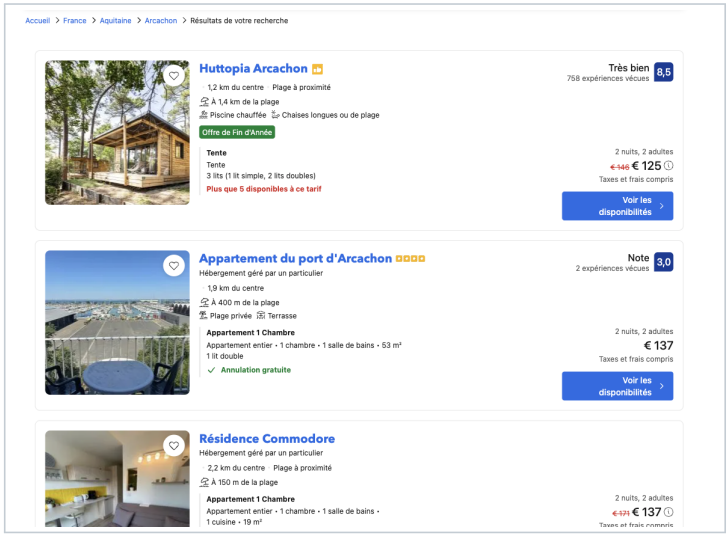
This appendix illustrates the process through which choice sets are created from scraped data described in Section 3. Focusing on La Ciotat, one of the four destination cities in the experiment, we provide summary statistics about the properties in the scrape, cluster and choice set in Table 6 and a visualization on a map in Figure 8.

A.4 Clutter effects on NASA-TLX subscales

Table 7 reports the clutter effect on each of the six NASA-TLX subscales discussed in Section 4.



(a) Original



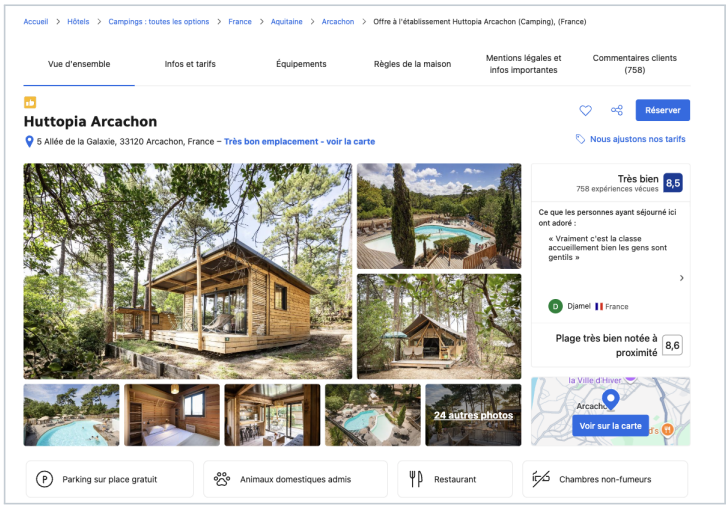
(b) Experimental

Notes: Map, sidebar filters, sort control, commission notice and view toggle are removed in (b). The topmost part of the interface with the search tool, invisible here, was the same in both cases, except that clicking on it had no effect during the experiment.

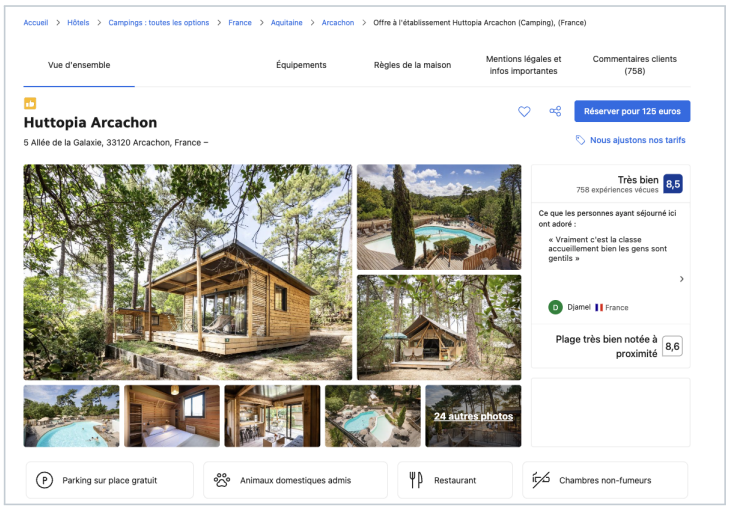
Fig. 6. Search page, original vs. experiment interface

B Online Resources

The supplementary materials contain four folders: the source code of the web browser extension, the oTree application (including the exact instructions, comprehension checks, questionnaire wording and the oral instructions script read to



(a) Original



(b) Experiment

Notes: Infos et tarifs tab, map link and map thumbnail are removed in (b); the price — which normally appears at the bottom of the page — moves into the reservation button.

Fig. 7. Property page, original vs. experiment interface

participants at the start of each session), the scraping and choice-set construction pipeline, and the replication package for the analyses. [PUBLIC REPOSITORIES WILL BE LINKED IN THE CAMERA READY VERSION]

Variable	Unit	Range	Definition
Treatments			
Cue visible	participant × choice set	0–1	Badge shown on the choice set's page (two of the four weekends per city)
Clutter	participant	0–1	Two levels: standard interface (cluttered) vs. interface with promotional and navigational elements removed or desaturated
Cue-eligible	listing	0–1	Carried the Preferred Partner badge on the unmodified site at scraping (three per choice set, one per cluster)
Preregistered outcomes			
Cue-eligible listing chosen	participant × choice set	0–1	The reserved listing is cue-eligible
Cue-eligible listing clicked	participant × choice set	0–1	At least one cue-eligible listing's detail page opened
Time on cue-eligible pages	participant × choice set	≥ 0 s	Sum, over cue-eligible listings, of seconds from the detail page opening to the next navigation away from it; $\log(1+x)$
Decision time	participant × choice set	≥ 0 s	Seconds from the nine listings being in place to the reservation click; $\log(1+x)$
Property page visits	participant × choice set	≥ 0	Number of detail-page views
Cue recognition	participant	0–1	Reports having noticed the thumbs-up icon
NASA-TLX subscales [29]	participant	0–100	Mental, physical and temporal demand, performance, effort, frustration; recorded in steps of 5
Preference consistency	participant × city	0–1	The two cue-hidden choices fall in the same cluster
Cued-choice consistency	participant × city	0–1	Given preference consistency, whether a cue-visible choice falls in that cluster (one observation per cue-visible weekend)
Perceived visual complexity	participant	1–7	"How visually complex did you find the site's interface?", 1 (very simple) to 7 (very complex)
Other questionnaire variables			
Decoy recognition	participant	0–1	Reports having noticed a check-mark icon / a badge icon that never appeared
Belief about the badge	participant	1–4	Agreement that badged properties "correspond more to my preferences than the others"
Familiarity with Booking.com	participant	1–4	1 (never used) to 4 (use it regularly)
Gender	participant	3 levels	Self-reported gender: woman, man, other
Age	participant	18–100	Age in years
Student status	participant	2 levels	Student vs. non-student
Household structure	participant	4 levels	Lives alone, with a partner, with family, or other
Residence in the Paris region	participant	0–1	Lives in the Île-de-France region
Comprehension failures	participant	0–3	Wrong first answer to the three instruction checks (prize, city/weekend choice, no cancellation)
Open answers	participant	text	Criteria for choosing accommodation (pretask); how the participant chose; what the icon means; whether and how it was used; free comments
Other tracked and derived variables (exploratory)			
Loading time	participant × choice set	≥ 0 s	Seconds the extension needed to assemble the nine listings behind the loading screen
Trial order	participant × choice set	1–12	Position of the choice set in the participant's session
Individual cue effect	participant	-1 to 1	Share of choice sets with a cue-eligible choice, cue-visible minus cue-hidden weekends
Hover dwell	participant	≥ 0 s	Mean hover duration on a listing card
Scroll depth	participant	0–100%	Mean of the deepest scroll position reached per choice set
Viewport dwell	participant	≥ 0 s	Mean total time listing cards were in the viewport per choice set
Tooltip opened	participant	0–1	Opened the badge's explanatory tooltip at least once

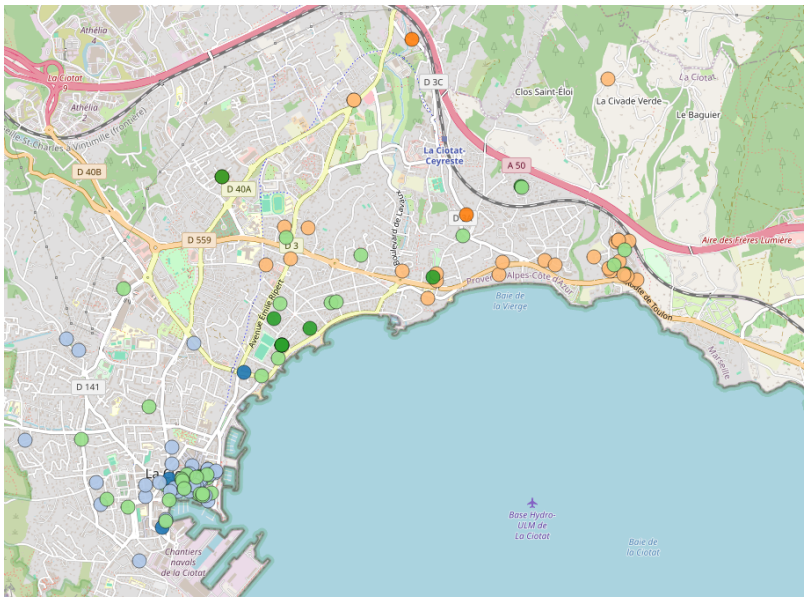
Notes: A choice set is one participant × city × weekend, and contains nine listings; each participant faced twelve. Ranges give the values a variable can take before any transformation noted in the definition. All questionnaire items were administered in French; wordings are translated.

Table 5. Variables used in the analysis.

	Total		Cluster 1		Cluster 2		Cluster 3	
	Full	In set	Full	In set	Full	In set	Full	In set
N properties	97	36	33	12	26	12	38	12
Preferred Partner (%)	16.5	33.3	24.2	33.3	15.4	33.3	10.5	33.3
Price, total stay (€)	254	256	332	332	220	223	210	212
Review score (0–10)	7.0	7.0	6.5	6.5	8.0	6.9	6.9	7.8
Distance to center (km)	2.8	3.1	2.4	3.0	3.4	3.3	1.1	1.4
Hotel or aparthotel (%)	4.1	5.6	3.0	0.0	3.8	8.3	5.3	8.3

Notes: Properties from La Ciotat scraped on Booking.com on 7 July 2026, the snapshot used to build the choice sets used in the paper. The table reports only on properties available across the nine candidate weekends and priced below the €400 experimental cap — those fed into the clustering pipeline described in Section 3. Preferred means that the property hold the Preferred Partner badge on the website. Price is the total stay price (2 adults, 2 nights); review score is Booking's own 0–10 scale, our mean excludes newly listed properties with no reviews yet. Distance to center is Booking's own assessment as displayed on the property card.

Table 6. Scraped properties, clusters and choice sets — La Ciotat



Notes: Map of the clusters obtained for La Ciotat. Same snapshot as Table 6. Cluster 1 is in blue, Cluster 2 in orange, Cluster 3 in green. Preferred properties appear darker. Map background: © OpenStreetMap contributors, CC-BY-SA

Fig. 8. Clusters of properties for La Ciotat

	Mental (OLS)	Physical (OLS)	Temporal (OLS)	Performance (OLS)	Effort (OLS)	Frustration (OLS)
Clutter	4.18	-1.58	12.7*	-3.52	-1.92	1.08
	(7.76)	(6.18)	(7.37)	(4.67)	(7.36)	(6.1)
Constant	45.2	27.5	37.7	80.5	44.1	26.7
	(5.39)	(4.3)	(5.12)	(3.25)	(5.12)	(4.24)
N	60	60	60	60	60	60

Notes: OLS, one observation per subject; subscales are 0–100 self-reports. The constant is the low-clutter mean. Standard errors in parentheses; preregistered one-sided tests ($H_1 : \beta > 0$). $^{\dagger}p < 0.1$, $^{*}p < 0.05$.

Table 7. Clutter effects on NASA-TLX subscales.