Validation of information architecture: Cross-methodological comparison of tree testing variants and prototype user testing

Article *in* Information and Software Technology · April 2025

DOI: 10.1016/j.infsof.2025.107740

CITATIONS
0

READS
6

3 authors:

Eduard Kuric
Slovak University of Technology in Bratislava
18 PUBLICATIONS **116** CITATIONS
SEE PROFILE

Peter Demcak
UXtweak
8 PUBLICATIONS **24** CITATIONS
SEE PROFILE

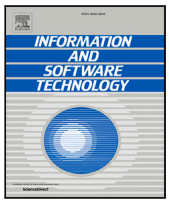
Matus Krajcovic
Slovak University of Technology in Bratislava
6 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Information and Software Technology 183 (2025) 107740

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



Validation of information architecture: Cross-methodological comparison of tree testing variants and prototype user testing

Eduard Kuric [a,b,*], Peter Demcak [b], Matus Krajcovic [b]

[a] Faculty of Informatics and Information Technologies, Slovak University of Technology, Ilkovicova 2, Bratislava, 84216, Slovakia
[b] UXtweak Research, Cajakova 18, Bratislava, 81105, Slovakia

ARTICLE INFO

Dataset link: https://github.com/treetest-research/information-architecture-validation

Keywords:
Information architecture
Navigation
User testing
Tree testing
Usability
Validation

ABSTRACT

Context: Tree testing is an established user testing method applied by software professionals to validate that an information architecture is logically navigable by users. We identify a methodological gap caused by previously unexamined non-uniformity between tree testing methods and software.
Objective: To reveal the role of the user interface representations in tree testing, this research compares the results of 3 commonly-used tree testing variants. To assess how indicative they are of the user's interaction with an information architecture implemented in an actual user interface, and to issue methodological recommendations, comparison with varied high-fidelity prototypes was performed.
Methods: Two between-subject studies were conducted to obtain a new dataset of users navigating an information architecture in tree testing and in interactive user interface prototypes. Data from 180 participants and 1800 task completions between 6 experimental conditions—3 tree testing and 3 prototype user interface variants—was evaluated quantitatively and qualitatively.
Results: Significant differences were found between results yielded by different tree testing method variants, and in how well they approximate user navigation in the same information architecture in high-fidelity prototypes. Implications for selection of the tree testing variant are proposed in the context of evaluated information architecture, with plausible broader applicability for tree testing methodology. Evidence supports the tree testing variant with highest visibility of previous navigation choices and direct controls over their reversal as the most accurate.
Conclusion: Presented findings can contribute to the design of software information architecture based on more accurate early validation, owing to tree testing that simulates less artificial user behavior more reflective of the user's navigation in the eventual user interface. We hope this will further the discussion and research leading to more holistic tree testing methodologies in the future.

1. Introduction

With most software being intended for human use, the fields of user experience (UX) and usability are important to software specification, design and testing. Empathizing with users is key for the success of any software system [1]. A variety of methods are dedicated to evaluating whether the users' mental state in a digital environment is up to par with user expectations [2,3] and whether users are capable of performing tasks in products of software engineering — effectively, efficiently, and to their satisfaction [4,5].

Information architecture (IA) is an aspect of the user experience that attributes logical structure to how the information in information systems is presented to users [6]. Tree testing is the most commonly performed usability research technique for evaluation of hierarchical (i.e., tree-like) information architecture models, which are the standard structure of menus or site maps [7]. Tree testing validates how content and functionality items within a prospective information architecture are labeled, categorized and organized [8]. Participants are presented with a model of an information architecture and tasked with finding a specific piece of information, typically with goals characteristic for the evaluated software solution. In today's age of the Web, a multitude of tree testing tools are available online.

Participants can provide their feedback by clicking through trees conveniently, remotely and – owing to the technique's high level of abstraction – early during the software design cycle.

The problem with tree testing, is that to the best of our knowledge, there is no validation of the various implementations of the method. Among tree testing software, detailed inspection reveals that despite

* Corresponding author at: Faculty of Informatics and Information Technologies, Slovak University of Technology, Ilkovicova 2, Bratislava, 84216, Slovakia.
E-mail address: eduard.kuric@stuba.sk (E. Kuric).

https://doi.org/10.1016/j.infsof.2025.107740
Received 15 March 2024; Received in revised form 10 January 2025; Accepted 1 April 2025
Available online 10 April 2025

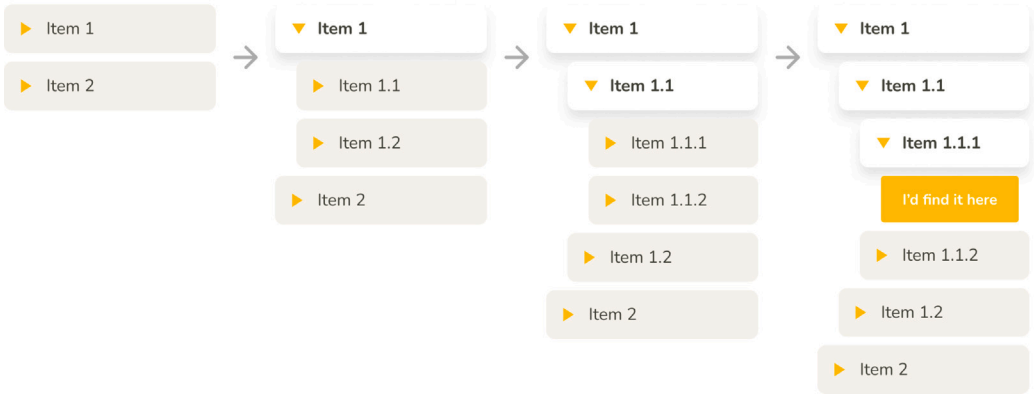


Fig. 1. Tree-visible variant of tree testing. The whole tree down to the current location in the tree is constantly visible (e.g., Item 2 as part of the first choice continues being presented in every step).

sharing the method's basic principles, there is a good amount of flexibility, allowing the tools to interpret what tree testing should look like. Some variants show more or less of the whole information architecture at a time, some bear more resemblance to a desktop or a mobile menu. Can all tree testing variants be treated as equal, or should there be any preferences for how they are applied?

To explore for answers to these questions, we introduce a taxonomy for three commonly used tree testing variants: Tree-visible, Path-visible and Compact. In practice, established UX research tools such as UXtweak,[1] Maze[2] and Optimal Workshop[3] implement a representation of these three variants. The above classification is designed to accurately represent and investigate typical yet distinct methods that are commonly used in practice. The inspiration for individual variants is left anonymous, to prevent conclusions from being drawn about specific commercial products from the outcomes of this study.

Tree-visible variant. See Fig. 1. In this variant, the user's previous choices and options remain continuously visible. After clicking items in the tree, thus moving to deeper levels, the clicked items and their siblings can still be seen (and interacted with, to undo previous actions). After traversing deeper down the tree, the whole tree of choices that lead to the selected solution becomes visible. This navigation pattern facilitates backtracking, most closely mirroring the appearance of cascading menus on desktop computers. As a disadvantage, this variant could lead to higher cognitive load, since the information from each previous choice remains visible.

Path-visible variant. See Fig. 2. Contrasting with the Tree-visible variant, previous choices remain visible, but their alternative options (siblings of previously selected nodes) are hidden from the user. The root of navigation starts from a single non-customizable root node labeled "Home". To backtrack, the user needs to click a previously selected node to see its children again.

Compact variant. See Fig. 3. Unlike the Tree-visible and Path-visible variants, the history of previous choices cannot be seen. Rather, the user can only see the last selected item and its direct children. After reaching the deepest level where no further children can be displayed, the parent remains visible so that two layers are always visible at once. The last clicked item functions as a "back" button. Backtracking is therefore possible at only one level per click.

The implications of the differences between the variants extend to the transferability of findings from tree testing to the prototypes created in ensuing steps of information system UX design process. Two prototypes can integrate the same information architecture, yet look and behave fundamentally differently (introducing their own visual identity, intrinsic usability issues, mobile vs. desktop design). Tree testing is supposed to validate an information architecture irrespective of how it will be implemented. Can information architecture be carried over from the various tree testing solutions into a prototype, with the expectation that its result will be similar if the prototype itself does not introduce new usability issues?

The contributions of this paper are:

- Identification of previously unexplored differences in the results of tree testing between representations of the method implemented by popular tree testing tools.
- Comprehensive investigation, quantitatively and qualitatively comparing users' interaction with information architecture between 3 tree testing variants and 3 high-fidelity navigation prototypes, employing a mix of standard and novel tree testing measures.
- For tree test design, identification of the best navigation pattern (closest to interaction in prototypes, without usability issues specific to the prototype).
- Two new between-subject studies, with 180 participants and 1800 task completions across 6 experimental conditions to obtain the required dataset.

Following sections of this paper provide a review of the necessary background 2, elaboration on the research questions 3, two studies as performed, evaluated and discussed (4, 5), a summary of implications 6, a recognition of threats to validity 7, consideration of future work 8 and a final conclusion 9.

2. Background and literature review

2.1. Information architecture (IA) and navigation

The definitions of information architecture (IA) have adapted diverse perspectives, yet are principally harmonious in meaning. To conjugate some of the popular definitions, IA is the organization of information in information environments, commonly also referring to the act/discipline/art/science of designing such organization to benefit human users [9–11]. The goal for well-designed IA is to facilitate user understanding of – and orientation in – information, which can be notoriously complex and thus difficult to cognitively grasp.

Menus, the typical application of a hierarchical information architecture, can be encountered anywhere, from the controls of a microwave to flight management systems used on airplanes. The usability of an information system can be critically hampered if its core information architecture is subpar. On the abstract, logical and semantic level, IA consists of information structure and labeling. The structure is typically hierarchical, a balance between breadth and depth being

[1] UXtweak - UX research software: https://uxtweak.com/.
[2] Maze: https://maze.co/.
[3] Optimal Workshop: https://optimalworkshop.com/.

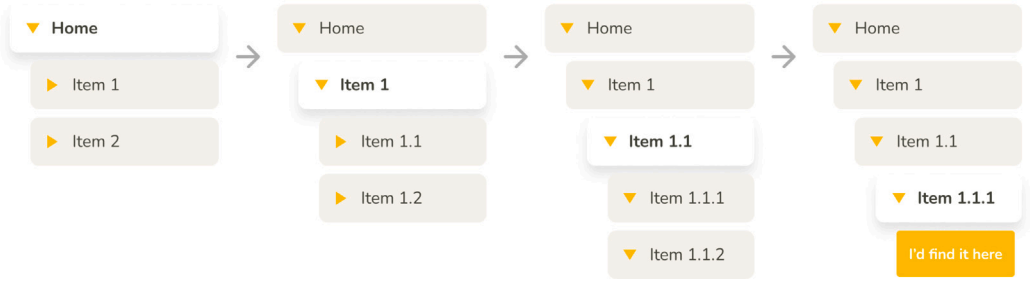


Fig. 2. Path-visible variant of tree testing. The whole path of selected choices down to the current location is constantly visible (e.g., Home and Item 1 are presented to the user at every step, but Item 2 is no longer visible after the first choice).

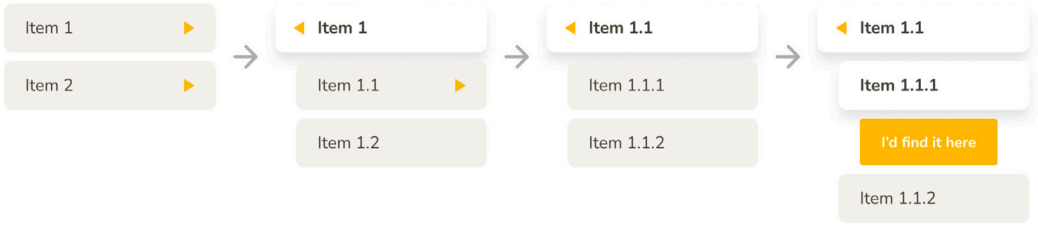


Fig. 3. Compact variant of tree testing. Only the options available in the current location in the tree are visible (e.g., the first choice Item 1 is no longer displayed once its child has been selected). Expansion indicators are not displayed on the leaf nodes.

a sought-after quality. Other structural schemes may be employed for different purposes, such as alphabetical, chronological or task-oriented [9]. In some tasks, a tag-based structure can yield higher efficiency than a hierarchy [12].

Navigation is a concept intrinsically linked with IA, whether considered as part of IA itself, or as a separate facet. Navigation is the actual physical manifestation of an information architecture, utilizing user interface design elements such as menus, contextual buttons, filters, footers and other controls [9]. Multiple navigation schemes may be employed at once to portray the same IA, fulfilling different purposes.

2.2. Tree testing and adjacent IA methods

When designing an information architecture (IA), usability research is conducted to leverage the knowledge of users in of making informed design decisions. Card sorting and Tree testing, as established methods of usability research, are counterparts that clearly represent the distinction between generative and evaluative research. Generative card sorting is used to lay the groundwork for the design of information hierarchy. By having research participants place functions and content into categories of their own choosing, insight is gained of mental models that users have of relevant concepts [10]. Evaluative tree testing is used once information architecture is designed. Tree testing [13,14] evaluates how users perform tasks within the hierarchy, in a style similar to a usability test, but independently from the navigational aspects that have yet to be designed.

Schmettow and Sommer [15] have shown that the mismatch between the user's mental models obtained from card sorting and the IA of a system does not axiomatically result in decreased user performance. This has been presented by Schmettow and Sommer [15] as a failure to validate card sorting, but can be interpreted more as a validation of tree testing. The ostensible discrepancy can be ascribed to the users' ability to adapt to the conceptual models different from their own, or to the "errors in translation" between card sorting results and actual information architecture. As a generative exercise, card sorting can show what is intuitive to users in a still vaguely defined and thus flexible conceptual space. Tree testing as an evaluative method can assess the intuitiveness (or lack thereof) regarding a specifically defined information architecture.

Analysis of tree testing reports includes metrics that indicate performance and behavior during a task, such as success ("How many participants find the right solution?"), directness ("How many participants avoid backtracking and go straight to an answer?"), node visit frequencies, first click frequencies, destination frequencies and task completion time [16–18]. The nature of usability issues can be qualitatively analyzed by investigating paths that participants traversed within the tree.

Tree testing's primary advantage is that information architecture and tasks are the sole requirements. This allows for quick and early validation that is easy to conduct prior to further design and development [7,14]. As such, tree testing has been deployed for information architecture evaluation in a number of domains, like government websites [19], education [20] and health [8]. Procedural improvements and frameworks revolving around tree testing have been proposed [7,21,22].

In the fields of computing and optimization, mathematical models have been investigated for automated creation of menu structures. Dayama et al. [23] designed an algorithm with an evaluation function that models the food foraging behavior of animals. Authors claim pioneership by endeavoring to mathematically model user behavior in menus. From the perspective of usability research, such a goal raises questions about how a foraging algorithm (where randomness is a key element in exploration) could realistically model real-world factors that influence user behavior, such as information organization, semantics and logic of labeling, domain, user mental state, knowledge, previous experiences, etc. While authors show improvement compared to baseline information architectures, the experiment task involved asking participants to locate an exact label in a menu. Without understanding further context, the participants' search behavior could truly be random (a warning sign about menu usability if it were discovered in an actual user task).

Troiano et al. [24] generate menu structures based on user preferences such as "which items should have a predecessor–follower or parent–child relationship", or "how many items should be on a single level". Even if preferences of this kind were obtained from real users, an information architecture that benefits users based on their behavior may contradict what users say about their preferences. Given the complexity of information architectures, automated methods could be valuable when assisting information architects with IA design, such as

by analyzing clusters obtained from card sorting [25], potentially with the help of large language models. Categorically, they should not be thought of as substitutes for usability research.

2.3. Isolating the impact of navigation elements

As a method, tree testing can be implemented in a variety of ways. Theoretically, inconsistency between implemented variants of the method may be reflected as inconsistency between their results. There is a gap in knowledge surveying commonly used methods and investigating how their individual differences may imprint on the results they obtain.

A number of remote tree testing tools are currently commercially available. As a core principle, tree testing tools evaluate austere versions of hierarchical information architectures. Closer inspection of individual tree testing tools reveals visual and functional irregularities between navigation elements employed to portray information architecture (see Figs. 1, 2 and 3).

Key differences between tree testing navigation elements stem from how the history of previously visited items–and backtracking within it–is treated. To design the variants, only the differences that have been identified in surveyed tree testing tools as having the potential to alter user behavior were integrated. Negligible cosmetic brand identity distinctions between the original online tools that implement these variants were normalized as part of variant abstraction and anonymization (e.g., font type, since each tool utilizes a different, but well-readable and aesthetically pleasant font).

Despite the premise of tree testing being to evaluate information architectures devoid of navigation, when representing an abstract information architecture for testing, some navigational aspects cannot be fully avoided. Even a bare text-only representation can be considered a type of simple navigation baseline. There is a lack of validation on how navigational representation of tree testing may impact the results of a tree testing.

In the closest related work, Le et al. [8] did not find differences between the task accuracy recorded in tree testing and an in-person usability test in a full-fledged application, but found differences in path length. Difference in path length could imply mismatch in other qualitative properties of the traversed paths attributable to the controlled variable of navigation, though the paper does not elaborate on further details. Between individual tree testing variants, similar or other kinds of differences could be found in users' interaction with information architecture.

Plurality of research supports the notion that changes to navigation design can impact user interaction with an information architecture. In an eye tracking study conducted by Tang et al. [26], performance load, user experience and satisfaction were altered by whether menu or flow diagram navigation is used. Jiang and Chen [27] presented guidelines for menus on mobile phones, including navigational aspects such as cascading menus and font sizes. Chhetri et al. [28] presented alternative hierarchical navigation designed for mobile device screens. Leuthold et al. [29] highlighted higher task effectiveness and gaze behavior efficiency when vertical menus are fully visible rather than dynamic. According to Mari Carmen Puerta Melguizo and van Oostendorp [30], users with lower spatial skills perform worse in sequential navigation. Placement of menus vertically on the left side is favorable in terms of efficiency and user satisfaction [31,32]; performance further improves by separation of submenus from the primary menu [33].

Navigation elements of hierarchical menus can vary (e.g., navigation placement [33,34], click vs. hover to uncollapse nested items [35], the presence of ads or other distractors). To validate tree testing as a method, it would have merit to evaluate the ability of tree testing to test the quality of information structure and labels in isolation from usability issues caused by bad navigation. Depending on whether differences in tree testing methods are substantial, specific tree testing methods may be also preferable in some instances (e.g., when designing primarily for menus on desktop vs. mobile devices). Differences should be investigated on a range of task complexities, with expectation that differences manifest as more pronounced in more complex tasks, as corroborated by Bodrunova and Yakunin [36].

3. Study aim

To investigate the relationship between information architecture (IA) validation techniques and the role filled by navigation when techniques are assessed in parallel, we posed the following research questions:

RQ1: *Do differences in the navigation design of tree testing variants (Tree-visible, Path-visible, Compact) cause differences in their results?*

No research has yet investigated the impact of the design aspects of tree testing. Multiple online testing tools implement their own versions where the hierarchical tree is displayed and designed to behave differently. Our main goal is to establish whether these differences are significant enough to alter tree testing results.

RQ2: *When an information architecture is tested via tree testing (Tree-visible, Path-visible, Compact), how does user interaction compare to the interaction with the same architecture implemented in a standard cascading menu prototype?*

The premise of tree testing is that it can validate an information architecture hierarchy independently from a digital environment where it will be implemented in later stages of the UX design process. If tree testing variants yield inconsistent results, a reasonable question arises: which variant best reflects user behavior in a prototype of higher fidelity? If the prototype's menu introduces no additional usability issues of its own, comparing their results could help with identifying the tree testing variant that provides the most accurate evaluation.

RQ3: *Do cosmetic changes to a menu's navigation design in a prototype (e.g., color, font weight, animations) affect how users interact with its information architecture compared to tree testing variants?*

A menu in a prototype can properly implement an IA validated by tree testing, yet the cosmetic design of the menu's navigation elements may introduce additional usability problems unrelated to the IA itself. By comparing tree testing results to user interactions with menu prototypes – both with and without cosmetic changes – the impact of these changes can be assessed. This insight can provide further context for selecting a tree testing variant that offers an accurate evaluation of the IA.

RQ4: *Between prototypes of menu navigations designed for mobile and desktop devices yet implementing the same information architecture, how does interaction in IA compare with results yielded by tree testing variants?*

Modern websites commonly adapt their menu navigation depending on whether they are being viewed on a desktop or mobile device. Can tree testing results be considered a generally acceptable baseline for the validation of IA for both types of navigation? Or are there preferences for which variant of tree testing to utilize when developing primarily desktop-first or mobile-first?

4. Study 1: Comparison of tree testing navigational variants

The goal of the first study was the assessment of the distinguishing factors of tree testing variants (Tree-visible, Path-visible, Compact) and whether they can lead to significant divergence in results due to differences in navigational representation. This goal corresponds with RQ1.

4.1. Method

Given the research goal of comparing the results of tree testing variants, the following dependent and independent variables [37] could be defined for our experiment:

- Independent variable: The three tree testing variants—Tree-visible, Path-visible and Compact.

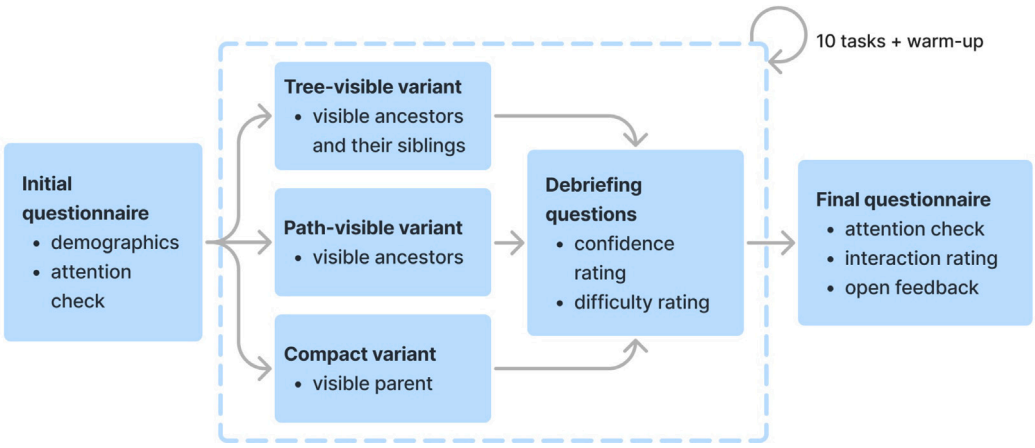


Fig. 4. Experiment procedure of Study 1, administering between-subject three tree testing variants to the same information architecture.

- Dependent variables: Behavioral and attitudinal metrics that describe users' interaction with an information architecture, which are typically obtained as the result of tree testing (see 4.1.5 Measures).

To organize the experimental conditions under which the effect of the independent variables on the dependent variables could be observed and evaluated, the experiment design was carefully considered. Between-subjects design (i.e., a separate participant group for each condition) was selected over within-subjects design (i.e., each participant exposed to each condition) to address threats to internal validity such as the carryover effect and fatigue [38,39]. As a drawback, between-subjects design requires more participants and group balancing procedures between the conditions to eliminate potential confounding variables. See 4.1.8 Recruitment and participants for the recruitment strategy implemented to obtain comparable groups.

Today, tree testing and other usability research is often conducted in the wild: remotely, online, and in an uncontrolled environment. Therefore, we also conducted our study online, in an uncontrolled environment, since a controlled study could raise concerns about its ecological validity. Only the variables connected to the participant's physical environment were uncontrolled (e.g., the user's device, configuration and physical surroundings), while variables of the digital environment were controlled (e.g., user interface, information architecture). Participants were randomly assigned into groups (see 4.1.8 Recruitment and participants) to mitigate the validity threat of selection bias.

4.1.1. Procedure

The experiment procedure (see Fig. 4) modeled the standard tree testing activity, as it can be performed in usability research to improve the information architecture of a website. The experiment was conducted remotely via the Tree Testing[4] tool in the usability research platform UXtweak. UXtweak is a web application, with a participant side that guides participants through a study, and a researcher side that provides analytical features to UX researchers.

The experiment procedure started with an initial questionnaire. Its questions comprised demographic profiling and quality assurance of responses via an attention check (see 4.1.6 Questionnaires). The experiment then branched into one of three tree testing activities that were identical procedurally, but each implemented distinct tree testing variant methods (see 4.1.4 Tree testing variant design). These variants formed the basis of the experiment's three conditions. To isolate the effects of variable navigational representation of the tree test from the structural aspects of an information architecture, an identical information hierarchy and labeling system was used in each variant (see 4.1.2 Information architecture design). Participants received a single warm-up task, then 10 proper tasks in randomized order. Tasks were designed for a multifaceted examination of navigation in a complex information architecture. They were followed by debriefing questionnaires for explicit feedback. The procedure concluded with a final questionnaire (second attention check and summative attitudinal evaluation of the tree testing activity).

4.1.2. Information architecture design

An online magazine was chosen as an everyday domain where the general audience can be expected to be familiar with the majority of presented concepts. Within this domain, a singular information architecture (IA) was created as the subject of tree testing comparable between its multiple variants. To design an IA of a realistic online magazine, we reviewed the IA of several online magazines as the basis for information items and categories. The designed IA is large so that is represents a complex system where IA design and tree testing are more relevant than in a simplistic system. It contains 4 layers with 53 branch nodes and 279 leaf nodes. The size of the information architecture does not inherently impact the result of the users' navigation through it, depending on how well its structure reflects user expectations. Excess depth increases the number of selections in which users can make mistakes, while breadth presents more options at once, increasing the cognitive complexity of choosing.

To increase the likelihood of usability issues surfacing in tree testing results (as they realistically do during information architecture design), flaws in structure and labeling were intentionally introduced into the IA. Example: The top level includes logically adjacent concepts like "Hobbies" and "Entertainment", which can make it difficult for users to choose between. The complete information architecture is available in our online repository (see Data availability statement).

4.1.3. Information architecture testing tasks

For the purposes of tree testing, a set of ten (10) tasks was designed to be given to participants. An eleventh preliminary task was used as warm-up to avoid inconsistency caused by participants' acclimatization to the experiment process. The warm-up task was excluded from the evaluation.

The tasks were designed to have varied degrees of difficulty and a broad coverage of topics linked to the online magazine's business goals (see Table 1). To mitigate researcher bias in achieving these criteria, tasks were designed by one researcher, then validated in a collaborative and iterative review process involving two other researchers. The initial difficulty of the tasks was determined by an informed estimate by a researcher based on the number of choices that users need to make, the number of options estimated as potentially misleading, and validation in a pilot experiment. Through collaborative review, we prepared 5

[4] UXtweak Tree Testing tool: https://www.uxtweak.com/tree-testing-tool.

Table 1
Tasks for information architecture (IA) traversal that participants engaged with during experiments. T0 is the warm-up task. Tasks have an expected correct answer and difficulty, designed for variability and complexity characteristic for realistic tree testing. Flaws relevant to the tasks were introduced to the information architecture prior to the experiment, either by deliberate design, or emerged naturally during pilot testing and the collaborative task review process.

Task number	Task wording	Correct answer	Difficulty	Flaws
T0	You would like to find articles that a lot of people are reading right now. Where would you look for them?	Trending->Popular articles	Easy	–
T1	You are having friends over for a movie night and you would like to watch something with a lot of jokes. Where would you look for the right movie recommendation?	Entertainment->Films->Genres->Comedy	Easy	–
T2	Planning a movie date, you want to check out what movies are currently playing in movie theaters. The movie theater should not be further than a thirty-minute drive from your current location. Where would you look for relevant information?	Entertainment->Cinema->Cinemas near you	Moderate	Also fits the "Entertainment->Films" or "Culture" category.
T3	You have bet on a boxing match and would like to see the results. Where would you look for them?	Hobbies->Sports->Individual sports->Boxing	Easy	Also fits the "Entertainment" category.
T4	You are in the mood to try making a spicy mesoamerican dish. Where would you look for recipe ideas?	Hobbies->Food->Regional cuisine->Mexican cuisine	Easy	–
T5	You'd like to update your wardrobe with some fresh articles to wear this season. Where would you look for inspiration?	Culture->Fashion->New collections	Easy	Also fits the "Hobbies->Beauty->Clothes" category.
T6	You would like to educate yourself a bit, and one topic that interests you is how life used to be back during the middle ages. Where would you look for articles about this?	Culture->History->Medieval	Moderate	Also fits the "Science" or "Society" categories.
T7	Your old laptop is out of commission, so you are interested in buying a new one. Where would you look for pointers on what brand and model may be the best for you?	Science->Technology->Computers and internet	Hard	Also fits "Lifestyle", "Hobbies" or "Entertainment" categories.
T8	You would like to learn about bird species–such as how they migrate or care for their young. Where would you look for articles like this?	Science->Nature->Animals	Easy	Also fits the "Environment", "Science->Earth and Space" or "Science->Biology->Zoology" category.
T9	Your health is important to you, so you like to keep abreast of the latest information about various illnesses, their symptoms and how best to avoid them. Where would you look for such information?	Society->Health->Health issues	Hard	Also fits the "Science->Health and medicine" category.
T10	Imagine you are soon about to become a mom or a dad for the first time. Where would you look for advice on how to raise and take care of a child?	Society->Family and relationships->Parenting	Moderate	Also fits the "Lifestyle" category.

easy, 3 moderate, and 2 hard tasks (number decreasing by estimated difficulty to maintain participant engagement).

The estimated difficulty of the tasks is relative — even from easier tasks, participants can be expected to encounter usability issues in navigation, so their behavioral results can be compared for the purposes of this study. The function of the estimated difficulty is to represent tasks of varied difficulty from the perspective of the usability researcher performing tree testing in practice, based on characteristics of the tasks known prior to its evaluation. As is common in usability research due to its inherent purpose, the actual difficulty of the tasks as established during the usability research (e.g., tree testing) may match or diverge from the researcher's expectations.

The order of all tasks (except for the warm-up) was randomized to avoid order bias. The tasks can have multiple valid answers within the information architecture tree as this can happen in realistic IAs to support users with varied goals and cognitive models. Not all answers that could be perceived as valid were selected as the expected correct answers (where the designer would place the content/functionality in the designed system) for the calculation of the success ratio (see 4.1.5 Measures). This was to invoke realistic cases of misalignment between the IA designer's expectations for user behavior and usability research results that provide indicators of usability issues. This represents a labeling issue where users perceive labels with different purposes as the solution to the task.

Tasks were phrased in such a way as to not contain verbatim any labels present in the information architecture itself. This was done to avoid priming participants towards a specific solution achieved by the simple matching of text [14]. Tasks were presented as scenarios that required participants to engage their own reasoning skills.

4.1.4. Tree testing variant design

For evaluation and comparison of the three tree testing variants presented in 2.3 Isolating the impact of navigation elements (Tree-visible, Path-visible, Compact), we implemented all variants as alternative methods in a unified digital environment for conducting user testing and online research experiments. We integrated the tree testing variant implementations in the infrastructure of the UXtweak usability research platform, as alterations of the platform's Tree Testing tool. UXtweak is a web application developed with Vue.js and provides all the other features required for the completion of the experiment (configurable guidance of participants through experimental procedure, collection of tree testing data, questionnaires)

4.1.5. Measures

During the experiment, the UXtweak Tree Testing tool captured a log of the user's clicks by which they navigated through the information architecture tree. The clicks contain the identification of the clicked

item within the tree, and a timestamp. This data was used to calculate behavioral measures for evaluating the information architecture.

A list of commonly used tree testing metrics was selected from literature [16–18] as the measures of comparison between tree testing variants. Due to the differences between variants being rooted in the method by which users are allowed to see and to trace back their steps, three novel metrics were introduced specifically as keys for measuring reversion of previously-made choices: backclicks, backtracks and backsteps.

The analyzed metrics originated either from implicit or explicit feedback. The source of implicit feedback is typically user behavior, such as mouse movements in a user interface [40] or actions taken in an information system (e.g., purchase) [41]. In tree testing, implicit feedback metrics describe the participant's performance while solving tasks in an information architecture:

- *Success ratio*: the percentage of participants who select the correct solution to the task.
- *Directness ratio*: the percentage of participants who solve the task by traversing a direct route through the information architecture hierarchy, without backtracking. It indicates the participant's certainty with their solution, regardless of whether their solution is a correct one.
- *Direct success ratio*: the percentage of participants who select the right solution directly, without backtracking.
- *Completion time*: the number of seconds taken by participants to complete the task.
- *Backtracks, Backclicks and Backsteps*: three novel inter-related metrics for comparison of backtracking behavior. A plurality of perspectives was adopted for a nuanced, balanced and holistic examination of the modes in which users can skip steps when they backtrack. Skipping refers to the option of reversing to a previously visited item in the hierarchy (see Fig. 2), or to a sibling of a previously visited item (see Fig. 1). No skipping means that a backtracking user who aims to go further back needs to trace back their steps one by one (see Fig. 3). Definition of the metrics:
  - *Backclicks*: the number of individual backtracking actions taken by participants during the task, regardless of their length. In navigations where a backclick can skip steps by leaping back multiple levels, even longer leaps are only counted once so that clicks are counted consistently and participants' use of the skipping option can be analyzed. Used to examine the frequency of backtracking action inputs.
  - *Backtracks*: the number of continuous backtracking subsequences in the traversed path. If multiple backclicks are performed in succession, they are accumulated into a single backtrack. Compared to backclicks, they are consistent for the same navigation in the tree regardless of whether skipping is available. Used to examine the frequency of decisions to backtrack.
  - *Backsteps*: the number of hierarchy levels traced backwards in participants' task completion paths. In navigations that allow skipping steps, multi-step backtracking actions add their length to the number of backsteps. Used to examine the overall length of backtracking actions.
- *Path length*: the number of items visited by participants during the completion of the task. This starts with the first and ends with the final clicked item. If the participant backtracks to the root category of the tree, it is also counted as an item visited in the path, although the root location is not explicitly labeled as an item in the tree itself (except for the Home item in the Path-visible variant).

Explicit feedback metrics captured the participant's attitudes towards the interaction they have just experienced, collected after each task via a 7-point Likert scale in the debriefing questionnaire:

- *Confidence*: "I felt sure about my answer". (1 = strongly disagree, 7 = strongly agree)
- *Difficulty*: "It was easy to complete this task". (1 = strongly disagree, 7 = strongly agree)

Qualitative analysis was performed with the paths traversed by participants during the tree test, which are the sequences of items in the order in which they were clicked. Aside from complete paths, item clicks were extracted as indicators of the users' decision-making process in individual locations within the information architecture hierarchy. First clicks are the users' initial clicked items on the tree's highest level. Destinations enable the comparison of endpoint items (known as leaf nodes in tree testing) that participants submitted as task solutions.

4.1.6. Questionnaires

Questionnaires were included at both the beginning and the end of the experiment (see Data availability statement for their full version in the online repository). To ensure even distributions of demographic attributes between participant groups, the initial questionnaire contained profiling questions about age, gender identity, education, annual income, frequency of using the web and frequency of reading online magazines (see 4.1.8 Recruitment and participants for descriptive data results).

Since the experiment was conducted remotely, as additional quality assurance, attention check questions were included both at the beginning and at the end of the experiment. The attention check questions were designed to verify effort, concentration, and to prevent speeding [3,42]. This was achieved by questions that are easy to answer after fully reading the text of the question, yet easy to misinterpret when not reading carefully.

Aiming for an explicit measure of the participants' satisfaction with the tree testing variant, an interaction rating Likert scale question concluded the final questionnaire. The question asked the participants to rate their perceived difficulty of clicking through the menu. The question explicitly asked the participants to disregard the structure of the menu in their rating, a phrasing aimed at decreasing the effect of the perceived complexity of the information architecture (which is identical between variants). Due to no significant differences between variants found in its answers, interaction rating was not analyzed further, reaching the conclusion that this question could be too abstract to answer properly.

4.1.7. Analysis

For assessing the differences between ordinal variables (success, directness and direct success ratio metric), the Chi-squared statistical test with Cramer's V effect size was used with two-stage Benjamini–Krieger–Yekutieli post-hoc p-value correction method. Where Chi-squared test could not be utilized due to its prerequisites not being met, Fisher's exact test was used. Non-ordinal metrics were assessed via the Kruskal–Wallis non-parametric test with Eta-squared effect size and Dunn's test post-hoc method. For correlation, Spearman's rank correlation coefficient was utilized.

The variable of task completion contained a small number of outliers where the time to complete the task was over 200 s. This can be attributed to the uncontrolled conditions of the experiment in which the participants completed the experiment in the wild, potentially allowing them to take a short break from the activity. This reflects the conditions in online tree testing tools for which the methods were evaluated. Since the data from these tasks was valid and without further deviations from other task completions (e.g., the traversed paths in the tree did not indicate the participants spent the whole time solving the task, explicit feedback did not indicate that these tasks were perceived as more difficult), the data from these tasks was preserved in the dataset as natural completion of tree testing in naturalistic conditions. As a treatment for outlier time completion values, mean imputation was used (stratified by task and variant), which is a simple, standard and valid approach when the replaced data points are few (5 out of 900 task completions, 0.56%) [43]. This imputation did not affect the analysis of other measures.

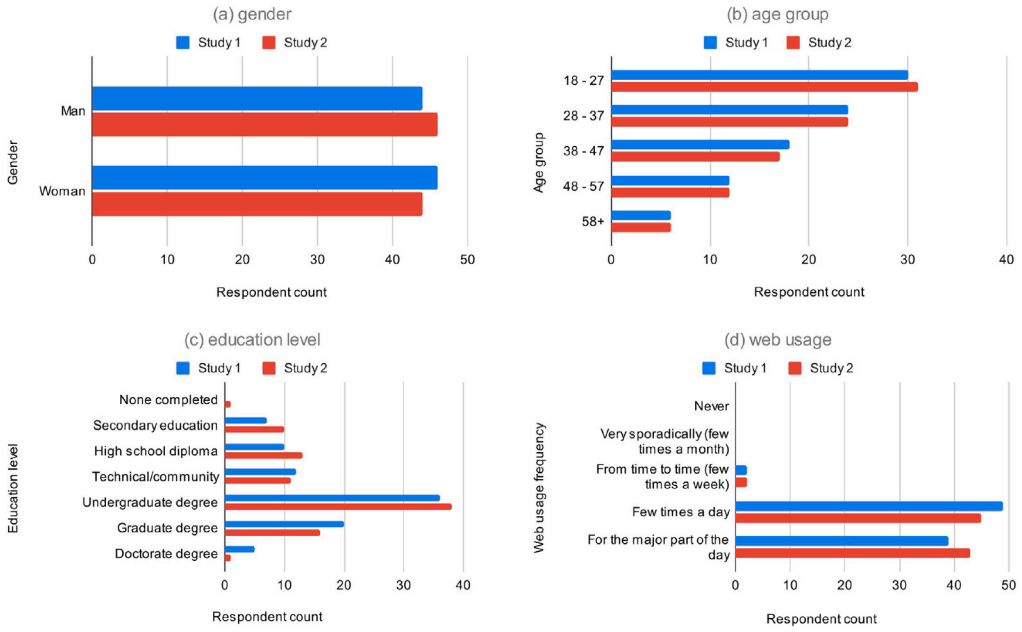


Fig. 5. A comparative visualization of participant demographics in Study 1 and Study 2 (see 5 Comparison of tree testing method variants and usability testing), showcasing gender proportions (a), age group breakdowns (b), education level (c) and web usage (d).

4.1.8. Recruitment and participants

In total, 90 participants were recruited for the experiment in the United Kingdom via the UXtweak User Panel (44 men, 46 women). Participants were members of the general population, the intended audience of the online magazine that is the domain of the information architecture employed in the study. Stratified sampling was used to ensure equal representation of genders and an age distribution aligned with the distribution of internet users.[5] Within the groups, the sampling was random. The sample was randomly divided into groups by 30 people to complete the experiment with three tree testing variants.

The sample consists of 30 participants aged 18–27 (33%), 24 participants aged 28–37 (27%), 18 participants aged 38–47 (20%), 12 participants aged 48–57 (13%), and 6 participants over the age of 58 (7%). See detailed sample characteristics in Fig. 5. To access the experiment, participants were required to use a desktop computer, so that the tree testing navigation would be viewed under equal conditions. No statistically significant differences were present between the conditions in terms of participant gender ($p = .96$), age ($p = .99$), frequency of using the Web, with the majority reporting answers "few times a day" (54%) and "for the major part of the day" (43%) ($p = .37$) and frequency of reading online magazines, where "daily" (40%) and "from time to time" (40%) were reported the most ($p = .42$). A pilot experiment with 15 participants was conducted beforehand, discovering no flaws with the experiment setup.

4.2. Quantitative results

Across all observed conditions, the overall task success ratio was 58.6%, direct success ratio was 23.8% and directness ratio was 37.8%. The mean task completion time was 35.89 s ($SD = 26.76$). The average self-evaluated measures were 5.7 ($SD = 1.45$) for confidence in answers and 5.24 ($SD = 1.65$) for task difficulty. A success ratio below 60% means that even tasks in the Easy category (50% of the tasks) contained some usability issues suitable for further evaluation. In particular, analysis of the success ratio demonstrates the importance of a mixed-methods approach for the analysis of tree testing results (and for comparison of their results in our study). For example, task T10, which was estimated with moderate difficulty, achieved the highest success ratio (87.78%), but a low direct success ratio (24.44%), indicating a significant usability issue.

Statistical evaluation of the similarities/differences between the measures (see 4.1.5 Measures) observed in the three variants involved a null and alternative hypothesis for every measure. For example, for the Success ratio evaluated for all tasks:

- Null hypothesis (H_0): Success ratio across all tasks is similar between variants Tree-visible, Path-visible and Compact,
- Alternative hypothesis (H_1): Success ratio across all tasks is different between at least two conditions: Tree-visible, Path-visible and Compact.

When taking a surface look at the average values of metrics aggregated together for all the tree testing tasks (see Table 2), the following measures contained no statistically significant difference between tree testing variants: success $\chi^2(2, n = 900) = 4.29, p = .12, v = .1$, directness, $\chi^2(2, n = 900) = 4.24, p = .12, v = .1$, direct success ratio $\chi^2(2, n = 900) = 4.57, p = .1, v = .11$, completion time $H(2, n = 900) = 0.34, p = .84, \eta^2 = .003$, traversed path length $H(2, n = 900) = 0.34, p = .84, \eta^2 < .002$, self-evaluated confidence in answers $H(2, n = 900) = 2.62, p = .27, \eta^2 = .001$, self-evaluated task difficulty $H(2, n = 900) = 0.72, p = .7, \eta^2 = .001$. Note that due to the task-driven nature of human interaction with information architecture, a lack of significant difference in just the global averages does not in itself provide evidence that the behavior described by the measures does not change in association with the conditions.

By comparison, global differences in two of the proposed measures – backclicks and backtracks – highlight how the variants' different representation of previous choices in display of ancestor items can significantly impact backtracking behavior during tree testing. Backclicks – $H(2, n = 900) = 6.06, p = .048, \eta^2 = .005$ – and backtracks – $H(2, n = 900) = 6.61, p = .037, \eta^2 = .005$ – reject the null hypothesis of similarity between their average values in tree testing variants. Backclicks and backtracks strongly correlated with one another, $r(898) = .99, p < .001$, as did backclicks with backsteps, $r(898) = .99, p < .001$, and backtracks with backsteps, $r(898) = .99, p < .001$. The relationship of correlation was maintained when evaluated within individual variants. Post-hoc

[5] Internet user age distribution by Statista: https://www.statista.com/statistics/272365/age-distribution-of-internet-users-worldwide/.

Table 2
Aggregated tree testing measures calculated for the sum of all evaluated user tasks. Statistically significant difference in backclicks and backtracks points to tree testing variants not being interchangeable as methods for IA evaluation.

Tree testing metric	Tree-visible		Path-visible		Compact		P-value*
Success ratio	58.7%		54.3%		62.7%		.117
Directness ratio	33.3%		38.7%		41,3%		.102
Direct success ratio	21%		22.3%		28%		.12
	M_1	SD_1	M_2	SD_2	M_3	SD_3	
Completion time	37.57	54.34	36.85	30.31	39.11	33.45	.107
Path length	7.53	6.54	6.87	4.93	7.83	6.98	.843
Backclicks	2.68	3.95	1.75	2.32	2.33	3.49	**.048**
Backtracks	2.37	3.38	1.6	2.09	1.94	2.87	**.037**
Backsteps	2.78	4.09	1.98	2.67	2.37	3.57	.12
Confidence	5.59	1.5	5.72	1.49	5.78	1.37	.269
Difficulty	5.22	1.57	5.24	1.72	5.26	1.67	.7

* P-value for corresponding statistical test (Chi-squared or Kruskal–Wallis) comparing tree testing variants.

Table 3
Fisher Exact Test for contingency tables larger than 2 × 2 provides evidence of significant differences in results between tree testing variants in tasks with intensive backtracking (as measured by backclicks).

Task number	Backclicks (mean)	Fisher Exact Test p-value of measures		
		Success	*Direct success*	*Directness*
T3	3.56	.73	.94	.83
T4	2.10	**.009**	1	.42
T5	2.12	.95	.15	.34
T7	5.66	.39	**.011**	**.044**
T8	2.61	.17	.59	.47
T10	2.03	**.037**	.75	.87

tests revealed that the Tree-visible variant with its fastest access to higher levels of the tree differs from other variants most significantly ($p = .016$ with Path-visible variant, $p = .045$ with Compact variant). Backsteps showed no significant differences between variants, $H(2, n = 900) = 4.25, p = .12, \eta^2 = .003$.

Given that task complexity should be considered in analysis [36], it was illustrated how tree testing results can change in task context. Table 3 shows a breakdown of six individual tasks where backtracking was the most intensive as an indicator of complexity. The differences in the average measure of success became statistically significant when evaluated specifically for these six tasks, $\chi^2(2, n = 540) = 6.54, p = .038, \upsilon = .2$ (Tree-visible: 56.7%, Path-visible: 50.6%, Compact: 63.9%). Impacted tree testing result metrics (success, directness, direct success) were associated closely with the tree testing variant in three of the individual tasks. See Fig. 6 for the results of tasks T4, T7 and T10 which were most significantly associated with the variant.

4.3. Qualitative path analysis

To understand the nature of differences between tree testing method variants beyond the resulting quantitative indicators of task performance, we performed an in-depth investigation of tree testing results—participant behavior while navigating through the tree. Each participant's path within the tree starts from the first clicked item and ends with the submitted solution. Focus was given to tasks where the differences in results were the most significant: T4, T7 and T10 (see Data availability statement to review the tasks and the information architecture). Although not all tasks demonstrated the same degree of disparities, the presence of variance itself could have broader implications for the methodological robustness of tree testing methods. The discovered patterns have potential for future exploration and generalization. Patterns were also found in other tasks, though to a smaller degree than the examples provided to highlight them.

4.3.1. Judicious vs. spontaneous exploration

Explorative analysis revealed differences between strategies adopted by participants between the individual tree testing variants. Participants interacting with the Compact tree testing variant (a single level of the tree visible at a time) repeatedly exhibited a more prudent and meticulous strategy to solving tasks, achieving higher success as the result. It was more common for participants to systematically search the tree, looking for solutions that matched the task better. This could be interpreted as them expending more effort to figure out which path was supposed to be correct, even as participants in other variants were more likely to be misled. In the Tree-visible and Path-visible variants (the path to the current position is visible at all times), participants were more likely to adopt strategies that were spontaneous and trial-and-error, both in traversing the tree and nominating their final solutions.

In task T4 (search for Mesoamerican dishes), 7 participants out of 30 found the correct solution, as opposed to only 2 participants in the Tree-visible variant and 0 participants in the Path-visible variant. From the start, they proceeded more analytically and with better consideration for all the options, with the item "Culture" being the most prevalent first click in this variant, as opposed to "Hobbies" and "Lifestyle" which received more first clicks in the other variants. The higher success ratio can be directly attributed to fewer participants settling on the earlier intuitive solution "Recipes" and more of them managing to find the designer's intended solution to the task: "Regional cuisine", which is located further down in the same category.

Task T7 (search for computer buying advice) as a complex task reinforces that participants in the Compact variant engaged with the options that tree testing presents to them with greater degree of attention and analytical thinking. On the highest level in the hierarchy, the correct solution to the task is placed in the category "Science", which has a logical link to the concept of technology compared to the other options, but can also be seen as counter-intuitive for this specific task (shopping advice). In the Compact tree test, it was still the most picked option, with 11 participants clicking it first. In the other variants, "Science" was only the fourth most clicked highest-level item, after "Lifestyle", "Hobbies" and "Entertainment").

Task T10 (search for parenting advice) demonstrates participants in the Compact variant as less likely to settle for suboptimal task solutions. All 30 participants in it completed the task on the intended correct solution (item "Parenting" in the super-category "Society"). Meanwhile, in the Tree-visible and Path-visible variants, 6 and 5 participants respectively submittted alternative solutions, predominantly "Self-growth" and "Self-care" in the "Personal development" category, super-category "Lifestyle". Although 10 participants in the Compact variant also visited "Personal development", none of them submitted any of the items within it as their answer.

4.3.2. Recovery aversion in the path-visible variant

Among the tree testing variants, the Path-visible variant is the one where participants were the most likely to reach an apparently suboptimal solution and not attempt a recovery by backtracking. These task solutions appeared as outliers that none or few other participants have submitted. This cannot be attributed to low effort by participants, since such solutions were also submitted by participants who solved other tasks properly and even engaged in backtracking in other tasks (hence they were aware of backtracking being an option). More frequent outlier solutions commonly had no apparent logical link with the task from a certain point in their path, implying that participants chose to nominate them to proceed with the test rather than backtrack.

4.4. Discussion

The experiment provides sufficient evidence that distinctions between tree testing variants can be meaningful enough to affect the

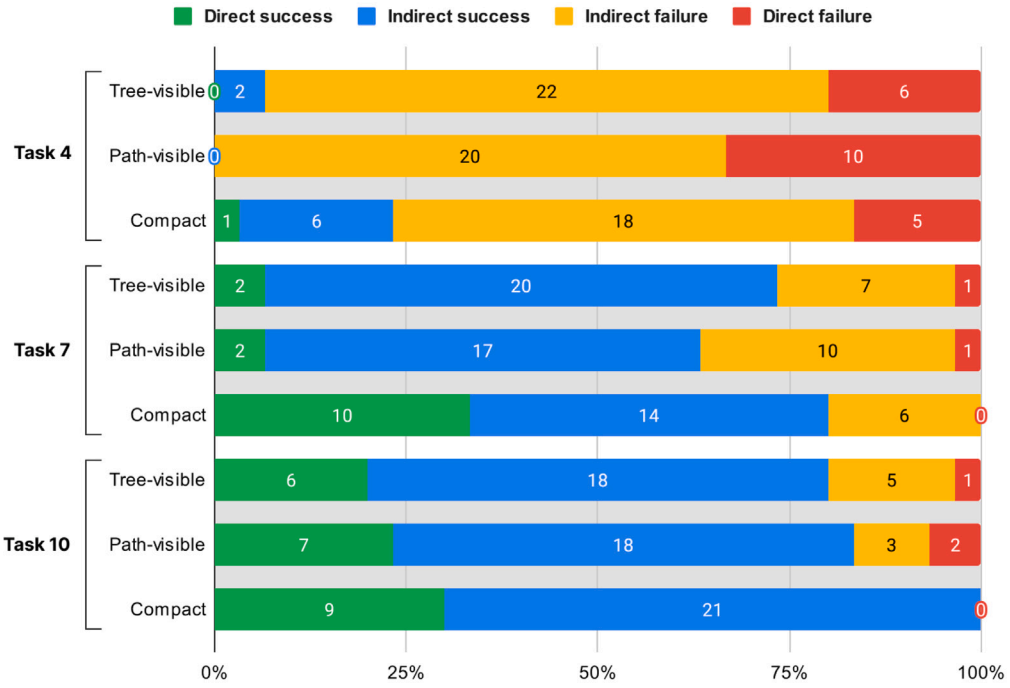


Fig. 6. Identical tree testing tasks individually manifest significantly different results between variants of the tree testing method.

method's results in a significant way. Corresponding with expectations, distinct methods of displaying the user's location in the information architecture and how this changes the interaction involved with backtracking, can be cited as the cause. These differences are substantial, given that they are encountered between otherwise identical instances of users completing the same tasks in the same information architecture.

The Tree-visible variant, where the user can effectively undo any perceived mistake by reselecting a choice they made previously with a single click, encouraged participants to backtrack the most. We posit that since users are able to backtrack in real navigations that implement information architecture in the digital environments they are designed for, this is a positive for the Tree-visible variant. Relevant implications are further discussed after the comparison with user behavior during a usability test of prototypes in Study 2.

It can be viewed as a paradox that the Path-visible variant encouraged less backtracking than the Compact variant. It is more similar to the Tree-visible variant, sharing features such as a fully visible path to the user's current location and granting users the ability to backtrack multiple steps at once. The seemingly random emergence of backtracking aversion in the Path-visible navigation implies that it is not specific to individual participants or tasks, but rather a situational usability issue with the path-visible navigation schema that discourages users from backtracking. We posit two potential causes through a juxtaposition with the Tree-visible variant, which does not have the same issue:

- Invisibility of ancestor siblings means that previous alternative choices are not immediately available. To effectively backtrack, the user needs to choose the correct step in the path they took. This may be difficult, since human working memory has limited capacity for the amount of novel information it can process concurrently [44]. Traditionally, this amounts to 7 ±2 items (Miller's Law [45]) Even a single category within an information architecture can easily contain more items.
- The Home root node as the starting point does not represent a specific concept that the user may intuitively like to backtrack to in order to solve the task. This can prevent users from recovering in cases where their first click was incorrect.

By comparison, in the Compact variant, the user needs to click more times to backtrack further. But the backtracking itself is a simple action with predictable results, avoiding potential decision paralysis.

Participants' more meticulous strategy in the Compact variant raises further questions about consistency of tree testing and how user behavior specifically in this variant compares to user behavior in actual user interfaces, reinforcing the motivation for Study 2.

At the time of the user's first click during a task, differences between tree testing variants are not visible to them yet. Users are located in the root of the information architecture, with no previously visited ancestors. Despite this, first clicks were also different between tree testing variants. Interaction with tree testing variants in multiple tasks can thus be posited to induce in participants a tendency to adapt in their navigation strategy.

5. Study 2: Comparison of tree testing method variants and usability testing

The second study was aimed at following up on the discovery of inconsistencies between the results of tree testing variants (see Study 1). Which tree testing variants yield results that most closely reflect how users interact with an information architecture through a standard navigation menu? How do the results vary between desktop and mobile navigation design? Are there potential meaningful differences with non-standard navigation? Usability testing with high-fidelity prototypes was adopted as the source of truth on more realistic user behavior.

5.1. Method

To allow for comparison with the results of the previous study, we employed an equivalent experiment design to Study 1, but instead of tree testing, participants interacted with a prototype.

5.1.1. Procedure

The core aim of the experiment was to support comparative analysis between tree testing variants and usability testing in actual user interfaces that implement the same information architecture. Experiment design thus had to place participants under equivalent conditions as

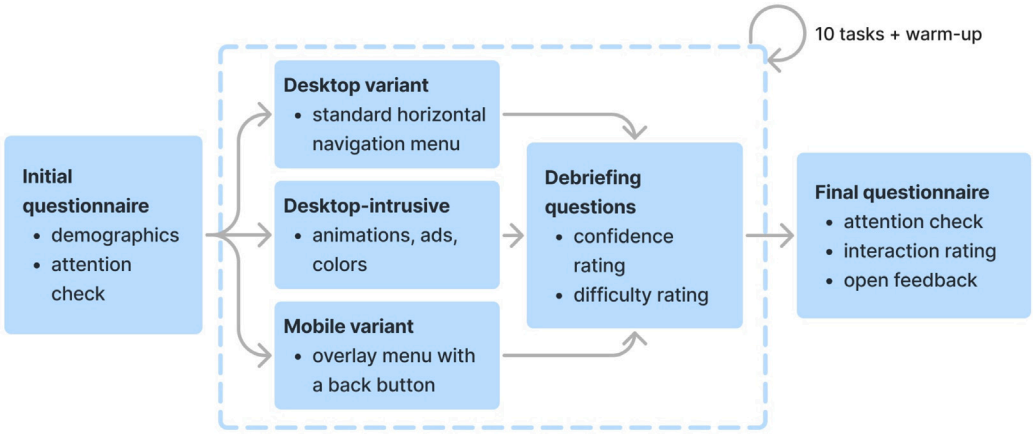


Fig. 7. Experiment procedure of Study 2, between-subject usability testing evaluation of three navigational implementations of an identical information architecture.

during the experiment of Study 1. The study scenario (see Fig. 7) was identical except for the controlled condition. Instead of completing tasks in a dedicated research environment of a tree testing tool (therefore abstracting other navigation aspects of the website), participants were tasked with finding information in the menu embedded directly in one of three interactive visual prototypes of a newly designed website (Desktop, Desktop-intrusive, Mobile).

The study was conducted remotely online on the participant's own device. To evoke authentic interactions with prototypes designed specifically for desktop and mobile devices (see 5.1.2 Navigation design in website prototypes), participants were required to use the corresponding type of device to complete the study. The UXtweak usability research platform infrastructure was used to carry out the study via the Website Testing tool dedicated to online usability testing.

In the Website Testing tool, participants were guided through the research process within the UXtweak application. During the task, the website prototype opened in a new tab. The UXtweak JavaScript snippet was inserted into the prototype source code, responsible for displaying the task information overlay at the bottom of the page and collecting data for analysis (screen recordings, low-level interaction events such as clicks and DOM changes). To complete a task after finding a solution in the prototype, participants clicked a button in the task information overlay.

Participants were asked the same questions as in Study 1 (see 4.1.6 Questionnaires). To enable comparative analysis with tree testing, the same measures were obtained as in Study 1 (see 4.1.5 Measures).

5.1.2. Navigation design in website prototypes

The three prototypes subjected to usability testing all implement the same information architecture designed in Study 1 (see 4.1.2 Information architecture design). The tasks completed by the participants were consistent with the tree test (see 4.1.3 Information architecture testing tasks. Multiple prototypes were designed to represent the disparity in navigation elements employed in real user interfaces, which is tied to answering research questions RQ2, RQ3 and RQ4. These prototypes represent potential implementations of the information architecture evaluated by variants of tree testing in Study 1, with which they are compared.

Web prototypes of an online magazine website were created with Bootstrap and HTML code and hosted online (see Data availability statement). Some alterations were made to native Bootstrap functionalities using JavaScript or CSS to make custom changes to the navigation menu behavior and appearance.

The Desktop prototype variant (see Fig. 8) was designed as a web page, navigable with a standard horizontal cascading dropdown menu. Users opened categories in the menu by clicking them. The menu had no immediately apparent usability issues concerning its navigation elements (horizontal bar, dropdown categories, clickable items, nonintrusive page content). Effectively, it was a straightforward application of the information architecture tested in tree tests to a user interface. The purpose of this prototype was to validate the inter-relational dynamics between user behavior in tree testing variants and a state when the same information architecture has been physically implemented (RQ2). Since the navigation path continued to be visible while descending down the menu, among the tree testing variants, this prototype bore the most resemblance to Tree-visible.

The Desktop-intrusive prototype variant (Fig. 9) was an adaptation of the Desktop prototype created to contain multiple examples of potential usability issues that may intrude on the user's interaction with the information architecture — distracting ad animations in website content, menu hover animations, excessive color clutter and small font weight. Intrusive factors were chosen to mimic the appearance of menus of real established business websites (see Data availability statement for screenshots), which they are directly inspired by. None of these design aspects should be interpreted as inherently harmful for user experience — real websites may have various well-supported justifications for their incorporation (e.g., color coding of product lines). The purpose of this prototype was to validate whether common cosmetic design variables in user interfaces (colors, fonts, animations) can have impact on inter-relational connections between navigation behavior on the website and in tree testing variants (RQ3).

The Mobile prototype variant (see Fig. 10) implemented a drill-down menu – a space-efficient navigation schema deployed typically on responsive websites displayed on mobile devices (inspired by a real banking website – see Data availability statement for screenshots). The user opened the menu by the press of a "hamburger button". The contents of a single menu category were visible at a time. The user could revert their actions by pressing the Back button. Participants involved in the experiment with this variant of the prototype completed the research activity in its entirety on their own mobile phone. The purpose of this prototype was to explore patterns between usability testing of prototypes and tree testing variants, and whether they were affected by the type of device that the user interface is designed for (RQ4). With a single level of the menu visible at a time, Compact was the tree testing variant that this prototype was similar to most.

5.1.3. Analysis

Statistical methods applied for analysis of results were inherited from Study 1 (see 4.1.7 Analysis). Due to technical malfunction, a small number of collected task completions (6 out of 900) were missing certain data (confidence and difficulty rating, task completion time), likely caused by participants closing the experiment prematurely. In these instances, mean imputation (stratified per task and prototype

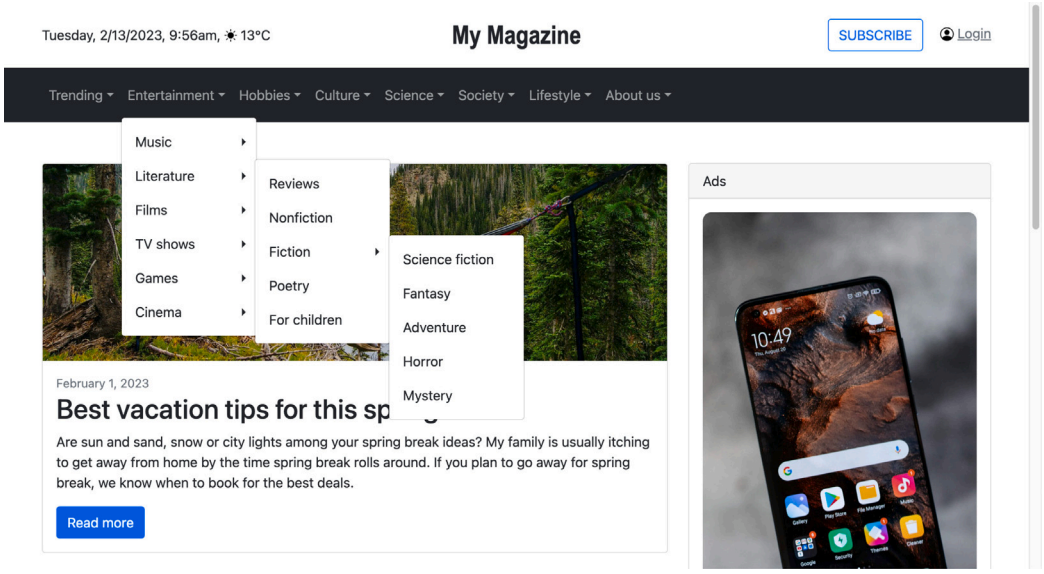


Fig. 8. Desktop high-fidelity prototype, the first condition in Study 2. A standard cascading horizontal menu used on desktop.

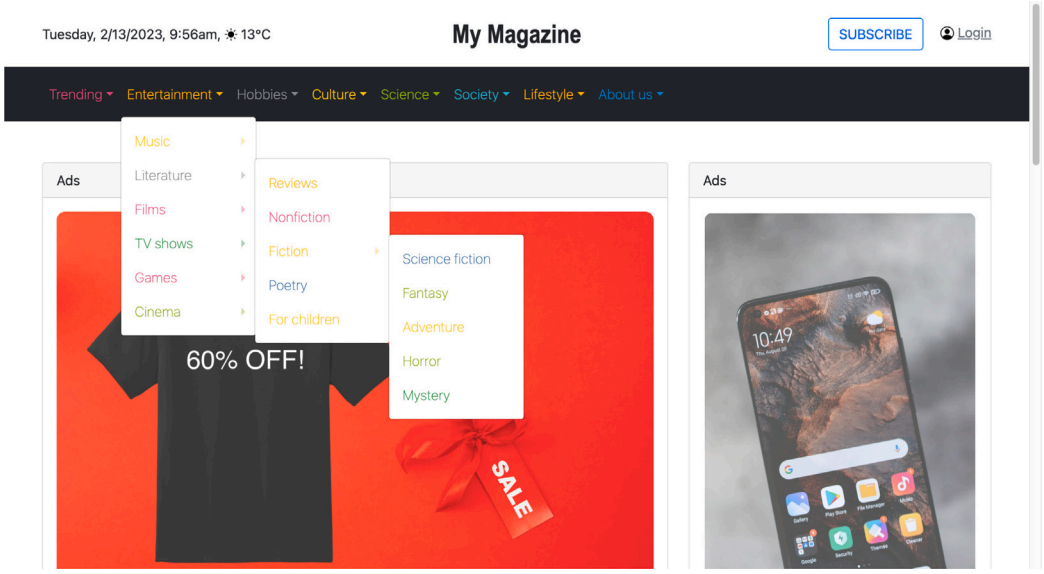


Fig. 9. Desktop-intrusive high-fidelity prototype, the second condition in Study 2. The ads are animated.

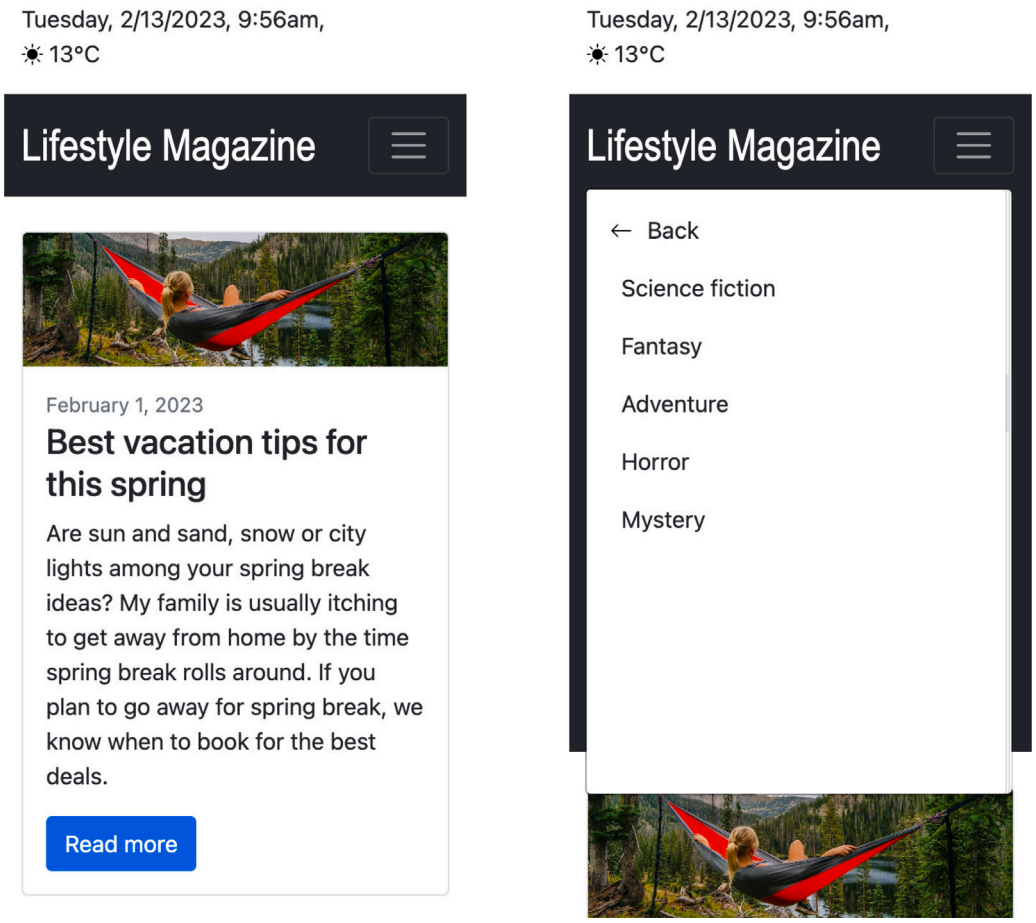
variant) was used to substitute the missing data points, a standard solution [43] given the small number of missing values (6 out of 900, 0.67%) of select measures, which also did not affect the further analysis.

5.1.4. Recruitment and participants

Ninety (90) participants were recruited for Study 2. A requirement for participation was that those involved could not have also been present in the tree testing Study 1 (avoiding familiarity with the evaluated information architecture and tasks). To allow for comparison with the results of Study 1, the population and the method of sectioning the sample were the same as in Study 1 (general population, stratified sampling for gender and age, see 4.1.8 Recruitment and participants) All included participants were residents of the United Kingdom, recruited via the UXtweak User Panel, 46 men and 44 women divided equally into three groups of 30 as per prototype variant. Age group representation was consistent with Study 1–30 participants aged 18–27 (33%), 24 participants aged 28–37 (27%), 18 participants aged 38–47 (20%), 12 participants aged 48–57 (13%), and 6 participants over the age of 58 (7%). The majority of participants use the Web "few times a day" (54%) or "for the major part of the day" (43%), while reported frequency of reading online magazines was "daily" (40%) or "from time to time" (40%). Between the groups, there was no evidence of differences in gender ($p = .96$), age ($p = 1$), the frequency of Web use ($p = .89$) or reading of online magazines ($p = .14$), see more detailed sample characteristics in Fig. 5.

5.2. Quantitative results

As a summary of key metrics across the sum of interactions in all website prototypes, the results were similar to the sum of all tree tests. Participants performed slightly worse in prototypes and were also faster, though only the difference in success was significant. The success ratio was 65.2%, the direct success ratio was 19.2% and directness ratio was 34%. Mean task completion time was 30.75 s ($SD = 22.37$), confidence 5.63 ($SD = 1.4$) and task difficulty 5.43 ($SD = 1.48$). No significant differences were found against tree testing in respect to correlations between observed measures.



(a) Collapsed drill-down menu. (b) Uncollapsed drill-down menu.

Fig. 10. Mobile high-fidelity prototype, the third condition in Study 2. States with the menu collapsed and uncollapsed are pictured.

A manipulation check was performed by assessing the associations between the prototype variants to evaluate whether the user behavior between the prototypes was distinctive enough to merit comparing prototypes to tree testing methods one-by-one (RQ3, RQ4). Statistical differences were the most evident in success, $\chi^2(2, n = 900) = 6.12, p = .047, v = .14$, path length, $H(2, n = 900) = 13.79, p = .001, \eta^2 = .013$, and confidence ratings $H(2, n = 900) = 11.44, p = .003, \eta^2 = .011$. Post-hoc analysis related this to the Mobile variant, where users completed tasks less successfully, traversed longer paths and were less confident in their answers (see Table 4). Other measures did not provide strong enough evidence of differences between variants, suggesting that at least on the global scale, cosmetic differences in navigation elements (as seen between the Desktop and Desktop-intrusive variants) may not have been significant enough to affect interaction with the information architecture.

Statistical tests for the association between conditions and measures were performed for all 6 information architecture interaction datasets (three tree tests variants from Study 1, three website prototype usability tests from Study 2):

- Success ratio – $\chi^2(2, n = 1800) = 18.79, p = .002, v = .2$
- Direct success ratio – $\chi^2(2, n = 1800) = 11.25, p = .047, v = .12$
- Directness ratio – $\chi^2(2, n = 1800) = 12.19, p = .032, v = .13$
- Completion time – $H(2, n = 1800) = 27.3, p < .001, \eta^2 = .01$
- Path length – $H(2, n = 1800) = 14.99, p = .01, \eta^2 = .006$
- Backclicks – $H(2, n = 1800) = 18.21, p = .003, \eta^2 = .007$
- Backtracks – $H(2, n = 1800) = 21.28, p < .001, \eta^2 = .009$
- Backsteps – $H(2, n = 1800) = 20.02, p = .001, \eta^2 = .008$

Table 4
IA evaluation measures aggregated across all tasks performed using the three evaluated prototype variants. Measures with statistically significant differences are highlighted.

Web testing metric	Desktop		Desktop-intrusive		Mobile		P-value*
Success ratio	68%		68%		59.70%		**.047**
Directness ratio	29%		31.60%		31,3%		.646
Direct success ratio	19.30%		20.70%		17.70%		.819
	M_1	SD_1	M_2	SD_2	M_3	SD_3	
Completion time	30.76	24.47	31.39	22.04	30.11	20.47	.606
Path length	7.17	5.92	6.49	4.55	8.74	6.95	**.001**
Backclicks	3.09	4.88	2.61	3.69	2.66	3.32	.669
Backtracks	2.91	4.5	2.48	3.5	2.28	2.8	.676
Backsteps	3.36	5.15	2.77	3.87	3.01	3.68	.467
Confidence	5.68	1.46	5.74	1.39	5.48	1.33	**.003**
Difficulty	5.51	1.53	5.44	1.51	5.33	1.39	.062

* P-value for corresponding statistical test (Chi-squared or Kruskal–Wallis) comparing web testing variants.

- Confidence – $H(2, n = 1800) = 16.72, p = .005, \eta^2 = .007$
- Difficulty – $H(2, n = 1800) = 8.83, p = .12, \eta^2 = .002$

Beyond what was discussed in Study 1 (navigation design impacting information architecture interaction), differences observed in these overarching tests (sans task difficulty rating) did not contribute to the objective of this analysis per se. Rather, their post-hoc analysis served as the cornerstone for delving deeper into connections between prototype and tree testing variants.

Table 5
Post-hoc analysis of Chi-squared and Kruskal–Wallis omnibus tests. P-values are corrected to mitigate the multiple comparisons issue.

	Path-visible	Compact	Desktop	Desktop-intrusive	Mobile
Success ratio					
Tree-visible	.401	.407	.076	.076	.846
Path-visible		.091	**.006**	**.006**	.295
Compact			.295	.295	.528
Desktop				.910	.091
Desktop-intrusive					.091
Path length					
Tree-visible	.645	.979	.934	.436	**.008**
Path-visible		.626	.587	.750	**.002**
Compact			.955	.421	**.009**
Desktop				.389	**.010**
Desktop-intrusive					**<.001**
Backclicks					
Tree-visible	**.014**	.116	.365	.926	.431
Path-visible		.380	**<.001**	**.011**	**.001**
Compact			**.013**	.096	**.018**
Desktop				.416	.905
Desktop-intrusive					.488
Backtracks					
Tree-visible	**.016**	**.049**	.235	.718	.612
Path-visible		.653	**<.001**	**.005**	**.003**
Compact			**.002**	**.020**	**.013**
Desktop				.408	.496
Desktop-intrusive					.883
Backsteps					
Tree-visible	.053	.117	.187	.754	.158
Path-visible		.713	**.001**	**.024**	**<.001**
Compact			**.004**	.060	**.003**
Desktop				.314	.928
Desktop-intrusive					.272
Completion time					
Tree-visible	.099	**.047**	.053	.348	.114
Path-visible		.737	**<.001**	**.010**	**.001**
Compact			**<.001**	**.004**	**<.001**
Desktop				.321	.723
Desktop-intrusive					.523
Confidence					
Tree-visible	.160	.149	.457	.292	**.050**
Path-visible		.969	.509	.726	**<.001**
Compact			.484	.697	**<.001**
Desktop				.756	**.007**
Desktop-intrusive					**.003**

Reviewing the post-hoc analysis (see Table 5) revealed that among the tree testing variants, Tree-visible bore the most similarity to the Desktop as well as Desktop-intrusive website prototypes implementing the same information architecture. None of the observed metrics supported the alternative hypothesis of the statistical test that users would perform navigation in the Tree-visible tree test differently from a straightforward implementation in desktop menu. On the other hand, post-hoc tests revealed significant differences when comparing the Path-visible and Compact variant to desktop prototypes, mainly in terms of backtracking behavior and completion time. On a global, task-insensitive scale, it made no notable difference if cosmetic adjustments related to colors, fonts or animations are made in the Desktop-intrusive prototype.

The mobile variant prototype was similar to the tree tests in terms of success, direct success and directness, but contrasted tree testing methods with longer path length and lower confidence. Aside from the tree-visible variant, the mobile prototype also differed from tree testing by involving more reverse actions–backclicks, backtracks and backsteps.

Given that in tree and usability testing, usability issues are identified in the context of user tasks, to understand the nature of similarities and differences between conditions, per-task post-hoc tests were performed for all the measures. Building upon the finding that applying the Tree-visible variant yielded results that were the most similar in results to genuine user interactions in a website navigations, focus was given to comparing the Tree-visible variant to website conditions in pairs:

Tree-visible and Desktop. Examined to validate with respect to task context that the Tree-visible tree-test yields similar data to interaction with navigation on a website user interface on desktop computers (RQ2). T1 was the sole task to contain statistically significant differences. Not in the participants' submitted solutions, but in the amount of effort expended–path length ($p = .0014$), the number of backclicks ($p = .002$), backtracks ($p = .002$) and backsteps ($p = .001$).

Tree-visible and Desktop-intrusive. Examination of the potential for cosmetic changes in menus to result in different user interaction with information architecture (RQ3). Task T8 provideed evidence of differences in backclicks ($p = .044$), backtracks ($p = .027$) and backsteps ($p = .039$). The lack of a difference in T1 is notable, since Desktop was different from the tree test in the same task. Desktop and Desktop-intrusive differed in path-length ($p = .010$), and equally in backclicks, backtracks and backsteps ($p < .001$).

Tree-visible and Mobile. Among the prototypes, the Mobile variant differed the most from tree testing in terms of user activity (RQ4). The nature of differences to the most similar tree testing method – Tree-visible – bears further investigation. Differences were related to the complexity of participant activity. In task T5, significant difference was found in path length ($p = .010$). In task T8, differences manifested in backsteps ($p = .049$) and confidence ($p = .002$).

5.3. Qualitative path analysis

Analysis of tree testing results typically involves identification of usability issues based on decisions made by users as they navigate through an information hierarchy. To understand what the discovered differences in task results between tree testing and prototypes can mean for the applicability of tree testing methods for accurate assessment of information architectures, we explored the paths traversed by participants at a higher degree of granularity. Note that these are illustrative examples that pertain specifically to the information architecture and user tasks evaluated in this study. Analysis was performed by researchers based on usability research expertise, and should not be used to draw generalized conclusions (see 7 Threats to validity). This extended analysis is presented to suggest important questions about tree testing that future research should aim to elucidate.

Task T1 being the only one to differ between the Tree-visible tree testing and Desktop prototype, merited a look at potential explanations. The analysis of participant's traversed paths revealed that in the prototype, participants performed comparatively more complex interactions within the Entertainment category. This category was characterized by containing multiple subcategories that could be logically tied to the topic of the task (finding something funny to watch at home): Films, TV shows, Cinema. It can be hypothesized that the horizontal positioning of the menu in the prototype is the contributing factor of a higher number of user actions within the category. Uncertain users picking between multiple similar subcategories could quickly swap between them in the horizontal menu. In Tree-visible tree testing, the tree displayed as a vertical list can require scrolling once multiple categories have been opened, potentially encouraging more careful exploration.

Inspection of tasks where results deviated in the Desktop-intrusive variant, revealed that users displayed a different preference for the options they were choosing. It can be proposed that this is due to variance in colors granting labels inconsistent visual weight. In T1, participants did not explore the category Entertainment as much as in the Desktop prototype, clicking the correct option Film directly in the majority of cases. The Film category in the prototype had a red label, while the categories that competed with it for attention in the Desktop prototype were cold green in the Desktop-intrusive prototype.

In T7, where the difference between Desktop and Desktop-intrusive was statistically significant, only a single participant first clicked on the option Hobbies. Hobbies was neutral gray, unlike the other more saturated labels in the root of the tree. In both Tree-visible and Desktop prototype, 7 participants click Hobbies.

The Mobile prototype's difference from the Tree-visible tree testing in task T5 can be explained by the differences in navigation. Seeing only one category in the Mobile menu, participants sometimes navigated multiple steps, then traced them all backwards. Occasionally, they clicked the same item that they just backtraced from, only to leave it again. This can be interpreted as the user no longer remembering the labels of their previous locations, a behavior not seen in the task within the Tree-visible variant where the current location is constantly visible.

For comparison, in Compact tree testing, less varied responses were submitted as the tree was explored more cautiously (see Study 1). The top submitted solution was in a different supercategory than top solutions in Tree-visible and Mobile prototype, where it was only the third.

5.4. Discussion

Juxtaposition of the results of commonly used tree testing methods against usability testing of prototypes reveals that despite overarching similarities, tree testing methods should not be treated as completely equivalent. While users navigate the Tree-visible tree testing and Desktop prototypes very similarly, this is not true for tree tests and prototypes in general. Tree testing frames an information architecture within a system defined by visual and interactive navigation design aspects of its user interface. While it can aim to approximate the evaluation of an abstract information architecture, it should be understood that the tree test is in itself a model. Different models may be more or less suitable in different contexts.

As expected, between the prototypes themselves, users can perform navigation tasks significantly differently even as it involves the same information architecture. Even cosmetic differences – particularly those that modify the visual weight to navigation items, as seen with colors – can change likelihoods of user navigation choices. Users may also associate concepts with specific colors. The differences are more significant when contrasting a prototype with a cascading menu on desktop and a mobile prototype with a drill-down menu.

The demonstration of the highest similarity being found between the navigation in the Tree-visible variant and Desktop/Desktop-intru sive prototypes is within reasonable expectations. The Tree-visible tree test bears most visual and interactive resemblance to cascading desktop menus, where previously taken choices remain visible. Compounded with findings about Path-visible and Compact variants from Study 1, the Tree-visible variant can be interpreted as the most accurate option for testing an information architecture intended primarily for desktop computers. Users can be reasonably expected to navigate an information architecture in similar manner, were it to be implemented as a standard cascading desktop menu. The need for re-testing menu hierarchy can potentially be reduced in standard menus if tree testing was previously conducted, allowing for usability testing to focus more on other aspects aside from navigation.

Usability testing would still be recommended for more peculiar menus or other IA implementations to gauge the impact of navigation element design. Even minor cosmetic changes may be susceptible to introducing their own usability issues (as seen in the Desktop-intrusive variant).

The lack of resemblance between Compact tree testing and the Mobile prototype testing results could be seen as surprising. Among the tree testing variants, the Compact tree test is the closest to a drill-down menu adapted for a mobile screen, with only the contents of a single category visible at a time. Significantly better results and more methodical navigation behavior in certain tasks could be reasonably attributed to the Compact variant's overt visual simplicity. Although the Mobile prototype navigation is just as simple, the prototype's higher level of realism (the appearance of a website user interface) could be seen as the reason for user behavior to be closer to the Tree-visible variant, where user behavior is more spontaneous.

6. Implications

The discussed findings can be used to draw potential implications for the evaluation of information architectures in usability research practice:

- The choice of a tree testing variant can be relevant for evaluation of information architecture (IA), since the method can affect the accuracy of results: the presence and severity of usability issues tied to IA. Different variants deployed in parallel can yield inconsistent results.
- Tree testing in which participants are allowed to efficiently undo their previously taken actions (Tree-visible variant) may be the most similar to the user interaction with an information architecture in prototypes. User behavior in this variant can be similar to user behavior in standard cascading menus made for desktop.
- For hierarchical information structures that implement non-standard navigation elements (e.g., menu with color differences between item labels), usability testing should likely be recommended to validate navigation separately, with tree testing serving as a benchmark. Navigation element design can alter which items users select in an information architecture.
- Between desktop and mobile, the same Tree-visible tree test can potentially serve as as the most accurate baseline for information architecture evaluation, before the implementation of prototypes. Task success can remain consistent between desktop and mobile. Usability testing with mobile prototypes can be used for the assessment of effort (path length, confidence) required to complete navigation tasks, which can differ from tree testing.

7. Threats to validity

As part of our conducted studies, we continually assessed the threats to the validity of our findings [37] to implement strategies that address them adequately. For the understanding of the validity of our results in the context of our methodology, we discuss the threats in this section.

External validity pertains to the generalizability of our findings. Since tree testing is a usability research method commonly employed in practice, threats to ecological validity (generalization to its real world use conditions) can be seen as of particular interest. Since in the wild, tree testing is commonly conducted online, we also conducted our experiment in an online uncontrolled environment so that our findings would not be applicable only to in-the-lab settings. As a trade-off, threats to internal validity that originate from this decision are discussed below. To maintain a manageable scope of our studies, a single IA was utilized. Further research of tree testing variants is important to assess the generalization of our findings to other IAs. However, our IA, the prototypes and the tasks for their evaluation were meticulously created to be realistic. Thus, it is feasible that our findings – such as the benefits of the Tree-visible variant – could be generalized to similar IAs. The sample was sourced from the general population so that the results are not restricted to specific groups. Nonetheless, to generalize to tree testing studies aimed at specific audiences, further experiments would be valuable.

Internal validity in our studies is concerned with whether the effects on the dependent variables (information architecture interaction measures) are caused by the independent variables (tree testing and prototype variants) rather than confounding factors. Participant factors were carefully managed to be consistent between conditions as illustrated in Sections 4.1.8 and 5.1.4. Between-subject experiment design prevented carryover bias. The threat of uncontrolled environment

factors (e.g., different screen sizes, distractions in surroundings) was managed with stratified random sampling of groups of sufficiently large size to mitigate selection bias. Additional steps to address internal validity threats included task order randomization and a pilot experiment to ensure the understanding of the study and its tasks.

Construct validity concerns the constructs we examined (measures, behavioral models), and whether they adequately correspond with the concepts involved in the research questions (user interaction with the information architecture during tree testing or usability testing of a prototype). The measures used in quantitative analysis and the tree traversal paths used for qualitative analysis are standard constructs utilized in tree testing [16–18] to represent user behavior. Being abstractions of the navigation behavior on a conceptual level, they are independent from specific IAs and their representations (prototypes or tree testing variants). The usability issues in the IA structure and labeling, and their depiction by the constructs was ensured through a collaborative review process and a pilot experiment. Since a usability issue as a misalignment between the expectation of users and IA designers can appear in leaf nodes of the tree, this was also reflected in the intended difference between designated correct solutions and the solutions submitted by participants in certain tasks. Thus, we could investigate a measure such as Success ratio, for which the expected correct solution is a parameter.

Reliability is an aspect that focuses on minimizing the researcher bias in the conclusions extracted from the study results. To assert a position of neutrality and manage the influence of personal beliefs, researchers reviewed the results of our study independently, then collaboratively reconciled them to reach a consensus.

8. Future research

Future research could focus on specific patterns discussed here (e.g., dissimilarity between information architecture interaction in compact tree testing and mobile prototype) on a variety of information architectures. Linking the discovered patterns to characteristics of an information architecture may lay grounds for formulating methodologies for selecting appropriate tree testing methods according to the circumstances under which IA will be applied.

The prototypes of navigation compared with tree testing in this study (cascading desktop menu, drill-down mobile menu) are not the sole straightforward representation of an information architecture in user interfaces. Other navigation schemes (menus with dropdown on hover, accordion menus on mobile devices, site maps, breadcrumbs, file management directories) may benefit from similar evaluation.

The relationship between user characteristics (such as age, domain familiarity, knowledge or cognitive abilities) and their navigation of tree testing variants is another aspect where more research may provide useful answers. In between-subject studies conducted here, these variables are controlled to allow for inter-group comparison. However, behavioral patterns may emerge differently for different target audiences.

9. Conclusion

Findings from our between-subjects experiments show that depending on which commonly used tree test variant is employed, the observed results of a tree test can significantly differ. Furthermore, some tree tests can resemble the user interaction with standard menus more or less than others, demonstrating that the tree test itself can be affected by its own navigation design. Within the context of the information architecture aimed at and evaluated with members of the general population in our experiment, the Tree-visible tree test approximated user behavior in menus potentially so well that in standard menus, further validation via usability testing may be redundant. Mobile menus took more of the user's effort, but were still overall similar in user behavior in the Tree-visible variant. Future research is required to draw more generalized conclusions.

Prioritizing ecological validity, our study was conducted in an uncontrolled environment to yield results reflective of usability research conducted online in realistic conditions. To minimize the estimated impact of potential confounding factors such as different screen sizes, mouse device settings, or distracting factors in the surroundings, we used stratified random sampling with 30 participants per condition (180 in total) so that random distribution of uncontrolled factors is likely to be balanced between groups. Investigation of the reproducibility of our findings in more controlled conditions would be beneficial in future research.

By answering its research questions, this paper contributes to the field of user testing in information systems, potentially imparting researchers and designers with better understanding the tree testing method for validation of information architecture (IA). More methodologically sound tree testing can in turn contribute to more intuitive menus and other hierarchical information navigation systems where users can locate information effectively, efficiently and to their satisfaction. We hope this research opens further discussion about tree testing, resulting in future research towards robust tree testing methodology.

CRediT authorship contribution statement

Eduard Kuric: Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Peter Demcak:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Matus Krajcovic:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-23-0408, co-financed by the Cultural and Educational Grant Agency of Slovak Republic (KEGA) under grant No. KG 014STU-4/2024. We would like to thank UXtweak j.s.a. for their generous financial support of this research and for the technical and expert support provided by the UXtweak Research team.

Data availability

The data and analysis scripts are openly available in a public paper repository at https://github.com/treetest-research/information-architecture-validation.

References

[1] H. Gunatilake, J. Grundy, I. Mueller, R. Hoda, Empathy models and software engineering — A preliminary analysis and taxonomy, J. Syst. Softw. 203 (2023) 111747, http://dx.doi.org/10.1016/j.jss.2023.111747.

[2] I. Otaduy, O. Diaz, User acceptance testing for Agile-developed web-based applications: Empowering customers through wikis and mind maps, J. Syst. Softw. 133 (2017) 212–229, http://dx.doi.org/10.1016/j.jss.2017.01.002.

[3] E. Kuric, P. Demcak, M. Krajcovic, G. Nguyen, Cognitive abilities and visual complexity impact first impressions in five-second testing, Behav. Inf. Technol. 43 (13) (2024) 3209–3236, http://dx.doi.org/10.1080/0144929X.2023.2272747.

[4] P.G. Dunn, A. Hayes, The problems with usability testing, in: Human-Computer Interaction. Design and User Experience, Springer International Publishing, Cham, 2020, pp. 420–430, http://dx.doi.org/10.1007/978-3-030-49059-1_30.

[5] O.L. Aiyegbusi, Key methodological considerations for usability testing of electronic patient-reported outcome (ePRO) systems, Qual. Life Res. 29 (2) (2020) 325–333, http://dx.doi.org/10.1007/s11136-019-02329-z.

[6] L.A. Rojas, J.A. Macías, Bridging the gap between information architecture analysis and software engineering in interactive web application development, Sci. Comput. Program. 78 (11) (2013) 2282–2291, http://dx.doi.org/10.1016/j.scico.2012.07.020.

[7] A. Tapia, A. Moquillaza, J. Aguirre, F. Falconi, A. Lecaros, F. Paz, A process to support the remote tree testing technique for evaluating the information architecture of user interfaces in software projects, in: Design, User Experience, and Usability: UX Research, Design, and Assessment, Springer International Publishing, Cham, 2022, pp. 75–92, http://dx.doi.org/10.1007/978-3-031-05897-4_6.

[8] T. Le, S. Chaudhuri, J. Chung, H.J. Thompson, G. Demiris, Tree testing of hierarchical menu structures for health applications, J. Biomed. Inform. 49 (2014) 198–205, http://dx.doi.org/10.1016/j.jbi.2014.02.011.

[9] L. Rosenfeld, P. Morville, J. Arango, Information Architecture: For the Web and Beyond, O'Reilly Media, 2015, URL https://books.google.sk/books?id=vJWJCgAAQBAJ.

[10] W. Ding, X. Lin, M. Zarro, Information Architecture: The Design and Integration of Information Spaces, Morgan & Claypool Publishers, 2017.

[11] A. Schall, Information architecture and web navigation, in: Eye Tracking in User Experience Design, Morgan Kaufmann, Boston, 2014, pp. 139–162, http://dx.doi.org/10.1016/B978-0-12-408138-3.00006-6.

[12] J. Walhout, S. Brand-Gruwel, H. Jarodzka, M. van Dijk, R. de Groot, P.A. Kirschner, Learning and navigating in hypertext: Navigational support by hierarchical menu or tag cloud? Comput. Hum. Behav. 46 (2015) 218–227, http://dx.doi.org/10.1016/j.chb.2015.01.025.

[13] D. Spencer, Card-based classification evaluation, 2003, https://boxesandarrows.com/card-based-classification-evaluation/. (Accesed 18 January 2024).

[14] P. Laubheimer, Tree testing: Fast, iterative evaluation of menu labels and categories, 2023, https://www.nngroup.com/articles/tree-testing/. (Accesed 18 January 2024).

[15] M. Schmettow, J. Sommer, Linking card sorting to browsing performance – are congruent municipal websites more efficient to use? Behav. Inf. Technol. 35 (6) (2016) 452–470, http://dx.doi.org/10.1080/0144929X.2016.1157207.

[16] B. Albert, T. Tullis, Measuring the User Experience: Collecting, Analyzing, and Presenting UX Metrics, Morgan Kaufmann, 2022, http://dx.doi.org/10.1016/C2018-0-00693-3.

[17] K. Whitenton, Tree testing part 2: Interpreting the results, 2017, https://www.nngroup.com/articles/interpreting-tree-test-results/. (Accesed 18 January 2024).

[18] H. Arslan, A.G. Yüksek, M.L. Elyakan, Ö. Canay, Usability and quality tests in software products to oriented of user experience, Online J. Qual. High. Educ.-July 5 (3) (2018).

[19] W.E. Nurcahyanti, Suhardi, Information architecture assessment of BPS headquarter official website, in: 2014 International Conference on Information Technology Systems and Innovation, ICITSI, 2014, pp. 177–182, http://dx.doi.org/10.1109/ICITSI.2014.7048260.

[20] B.J. White, W.A. Kapakos, User experience (ux) in the cis classroom: better information architecture with interactive prototypes and ux testing, Issues Inf. Syst. 18 (2) (2017) http://dx.doi.org/10.48009/2_iis_2017_59-70.

[21] F. Paz, A. Lecaros, F. Falconi, A. Tapia, J. Aguirre, A. Moquillaza, A process to support heuristic evaluation and tree testing from a ux integrated perspective, in: ITNG 2023 20th International Conference on Information Technology-New Generations, Springer International Publishing, Cham, 2023, pp. 369–377, http://dx.doi.org/10.1007/978-3-031-28332-1_42.

[22] A. Moquillaza, F. Falconi, J. Aguirre, A. Lecaros, A. Tapia, F. Paz, Using remote workshops to promote collaborative work in the context of a UX process improvement, in: Design, User Experience, and Usability, Springer Nature Switzerland, Cham, 2023, pp. 254–266, http://dx.doi.org/10.1007/978-3-031-35699-5_19.

[23] N.R. Dayama, M. Shiripour, A. Oulasvirta, E. Ivanko, A. Karrenbauer, Foraging-based optimization of menu systems, Int. J. Hum.-Comput. Stud. 151 (2021) 102624, http://dx.doi.org/10.1016/j.ijhcs.2021.102624.

[24] L. Troiano, C. Birtolo, R. Armenise, Searching optimal menu layouts by linear genetic programming, J. Ambient. Intell. Humaniz. Comput. 7 (2) (2016) 239–256, http://dx.doi.org/10.1007/s12652-015-0322-7.

[25] C.K. Sione Paea, G. Bulivou, Information architecture: Using best merge method, category validity, and multidimensional scaling for open card sort data analysis, Int. J. Hum.–Comput. Interact. 40 (2) (2024) 203–223, http://dx.doi.org/10.1080/10447318.2022.2112077.

[26] P. Tang, Z. Yao, J. Luan, J. Xiao, How information presentation formats influence usage behaviour of course management systems: flow diagram navigation versus menu navigation, Behav. Inf. Technol. 41 (2) (2022) 383–400, http://dx.doi.org/10.1080/0144929X.2020.1813331.

[27] L. Jiang, Y.-H. Chen, Menu design on small display user interfaces: Measuring the influence of menu type, number of preview items, and menu breadth on navigation efficiency, Int. J. Hum.–Comput. Interact. 38 (7) (2022) 631–645, http://dx.doi.org/10.1080/10447318.2021.1954781.

[28] A.P. Chhetri, K. Zhang, E. Jain, A mobile interface for navigating hierarchical information space, J. Vis. Lang. Comput. 31 (2015) 48–69, http://dx.doi.org/10.1016/j.jvlc.2015.10.002.

[29] S. Leuthold, P. Schmutz, J.A. Bargas-Avila, A.N. Tuch, K. Opwis, Vertical versus dynamic menus on the world wide web: Eye tracking study measuring the influence of menu design and task complexity on user performance and subjective preference, Comput. Hum. Behav. 27 (1) (2011) 459–472, http://dx.doi.org/10.1016/j.chb.2010.09.009.

[30] U.V. Mari Carmen Puerta Melguizo, H. van Oostendorp, Seeking information online: the influence of menu type, navigation path complexity and spatial ability on information gathering tasks, Behav. Inf. Technol. 31 (1) (2012) 59–70, http://dx.doi.org/10.1080/0144929X.2011.602425.

[31] A. Burrell, A. Sodan, Web interface navigation design: Which style of navigation-link menus do users prefer? in: 22nd International Conference on Data Engineering Workshops, ICDEW'06, 2006, http://dx.doi.org/10.1109/ICDEW.2006.163, 42–42.

[32] F. Zhang, S. Lin, X. Li, Y. Shuai, H. Jiang, C. Yao, F. Ying, F. Gao, Navigation configuration and placement influences the visual search efficiency and preference, in: Proceedings of the 26th International Conference on World Wide Web Companion, in: WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, pp. 871–872, http://dx.doi.org/10.1145/3041021.3054236.

[33] J.R. Kingsburg, A.D. Andre, A comparison of three-level web menu navigation structures, Proc. Hum. Factors Ergon. Soc. Annu. Meet. 48 (13) (2004) 1513–1517, http://dx.doi.org/10.1177/154193120404801309.

[34] X. Faulkner, C. Hayton, When left might not be right, J. Usability Stud. 6 (4) (2011) 245–256.

[35] E.P. dos Santos, S.M. de Lara, W.M. Watanabe, M.C. Filho, R.P. Fortes, Usability evaluation of horizontal navigation bar with drop-down menus by middle aged adults, in: Proceedings of the 29th ACM International Conference on Design of Communication, SIGDOC '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 145–150, http://dx.doi.org/10.1145/2038476.2038504.

[36] S.S. Bodrunova, A. Yakunin, Impact of menu complexity upon user behavior and satisfaction in information search, in: Human Interface and the Management of Information. Information in Applications and Services, Springer International Publishing, Cham, 2018, pp. 55–66, http://dx.doi.org/10.1007/978-3-319-92046-7_5.

[37] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén, et al., Experimentation in Software Engineering, vol. 236, Springer, 2012.

[38] G. Richard, T. Pietrzak, F. Argelaguet, A. Lécuyer, G. Casiez, Within or between? Comparing experimental designs for virtual embodiment studies, in: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces, VR, 2022, pp. 186–195, http://dx.doi.org/10.1109/VR51125.2022.00037.

[39] G. Charness, U. Gneezy, M.A. Kuhn, Experimental methods: Between-subject and within-subject design, J. Econ. Behav. Organ. 81 (1) (2012) 1–8, http://dx.doi.org/10.1016/j.jebo.2011.08.009, URL https://www.sciencedirect.com/science/article/pii/S0167268111002289.

[40] E. Kuric, P. Demcak, M. Krajcovic, P. Nemcek, Is mouse dynamics information credible for user behavior research? An empirical investigation, Comput. Stand. Interfaces 90 (2024) 103849, http://dx.doi.org/10.1016/j.csi.2024.103849.

[41] E. Kuric, A. Puskas, P. Demcak, D. Mensatorisova, Effect of low-level interaction data in repeat purchase prediction task, Int. J. Hum.–Comput. Interact. 40 (10) (2024) 2515–2533, http://dx.doi.org/10.1080/10447318.2023.2175973.

[42] A.J. Berinsky, M.F. Margolis, M.W. Sances, Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys, Am. J. Political Sci. 58 (3) (2014) 739–753, http://dx.doi.org/10.1111/ajps.12081.

[43] D.P. Anil Jadhav, K. Ramanathan, Comparison of performance of data imputation methods for numeric dataset, Appl. Artif. Intell. 33 (10) (2019) 913–933, http://dx.doi.org/10.1080/08839514.2019.1637138.

[44] J. Sweller, J.J.G. van Merriënboer, F. Paas, Cognitive architecture and instructional design: 20 years later, Educ. Psychol. Rev. 31 (2) (2019) 261–292, http://dx.doi.org/10.1007/s10648-019-09465-5.

[45] M.E. Cornejo, J. Medina, E. Ramírez-Poussa, C. Rubio-Manzano, Preferences in discrete multi-adjoint formal concept analysis, Inform. Sci. 650 (2023) 119507, http://dx.doi.org/10.1016/j.ins.2023.119507.