Driven to Distraction: Examining the Influence of Distractors on Search Behaviours, Performance and Experience

LEIF AZZOPARDI, University of Strathclyde, Scotland
DAVID MAXWELL, Delft University of Technology, The Netherlands
MARTIN HALVEY, University of Strathclyde, Scotland
CLAUDIA HAUFF, Delft University of Technology, The Netherlands

Advertisements, sponsored links, clickbait, in-house recommendations and similar elements pervasively shroud featured content. Such elements vie for people's attention, potentially distracting people from their task at hand. The effects of such "*distractors*" is likely to increase people's cognitive workload and reduce their performance as they need to work harder to discern the relevant from non-relevant. In this paper, we investigate how people of varying cognitive abilities (measured using *Perceptual Speed* and *Cognitive Failure* instruments) are affected by these different types of distractions when completing search tasks. We performed a crowdsourced within-subjects user study, where 102 participants completed four search tasks using our news search engine over four different interface conditions: *(i)* one with no additional distractors; *(ii)* one with advertisements; *(iii)* one with sponsored links; and *(iv)* one with in-house recommendations. Our results highlight a number of important trends and findings. Participants perceived the interface condition without distractors as significantly better across numerous dimensions. Participants reported higher satisfaction, lower workload, higher topic recall, and found it easier to concentrate. Behaviourally, participants issued queries faster and clicked results earlier when compared to the interfaces with distractors. When using the interfaces with distractors, one in ten participants clicked on a distractor—and despite engaging with a distractor for less than twenty seconds, their task time increased by approximately two minutes. We found that the effects were magnified depending on cognitive abilities—with a greater impact of distractors on participants with lower perceptual speed, and for those with a higher propensity of cognitive failures. Distractors—regardless of their type—have negative consequences on a user's search experience and performance. As a consequence, interfaces containing visually distracting elements are creating poorer search experiences due to the *"distractor tax"* being placed on people's limited attention.

CCS Concepts: • **Information systems** → **Users and interactive retrieval**; • **Human-centered computing** → **HCI design and evaluation methods**.

ACM Reference Format:
Leif Azzopardi, David Maxwell, Martin Halvey, and Claudia Hauff. 2023. Driven to Distraction: Examining the Influence of Distractors on Search Behaviours, Performance and Experience. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3576840.3578298

1 INTRODUCTION

Advertisements, sponsored links, clickbait, recommendations and the like are commonplace on the Web [10, 11, 54, 57]. Today, nearly every page is filled with such elements as a way to monetise relevant content (via advertising revenue), or increase engagement through further recommendations—often to show yet more advertising and sponsored links [34, 41, 45]. Depending on the search portal/engine used, such elements may be presented before searching, during searching

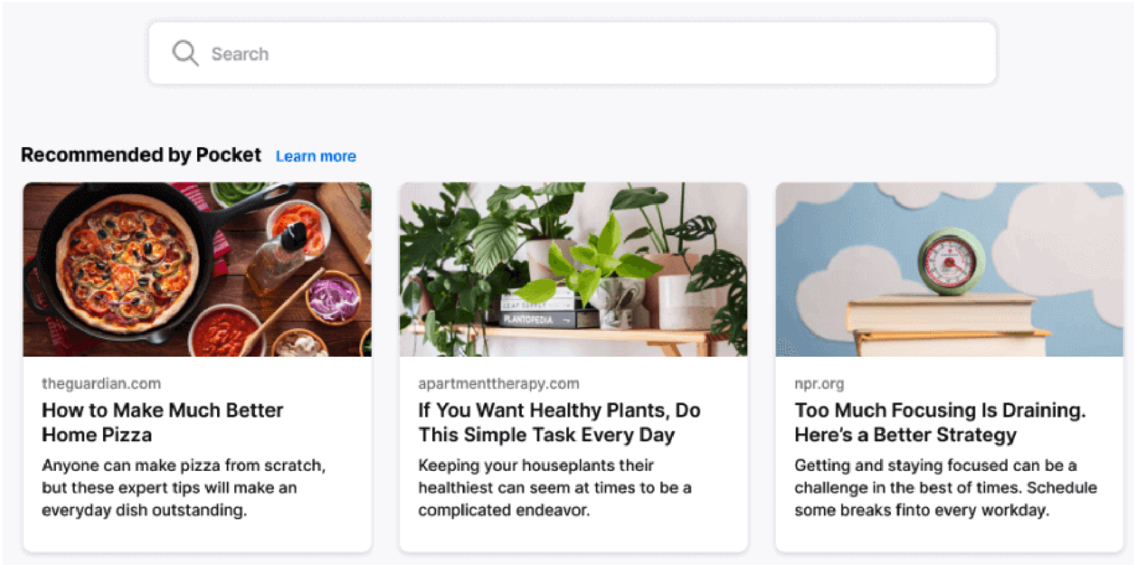


Fig. 1. A popular browser tempting users with clickbait in a so-called *chumbox* [62]. Will they ignore the distractors and search, or will they take the bait?

(on the *Search Engine Results Page (SERP)*) and/or while browsing (on landing pages). To satisfy their information needs, people need to traverse pages cluttered with these elements that are often not particularly relevant to their tasks. Such *"distractors"* need not be adversarial in nature, as they may be native recommendations that are presented alongside the primary set of results or content. For example, a user searching for movies on *IMDB* could—in addition to the organic search results—also be recommended popular movies, the latest movies, and so forth. However, all of these additional elements can potentially distract users from their primary information-seeking task.

Prior work has largely been focused on examining the effects of advertisements on users' search experience—where it has been largely shown that they are perceived negatively [20, 32, 35]. However, we argue that any additional visual element added to such pages—*whether it be ads, sponsored or native*—has the potential to be a distraction. With theories of attention suggesting that people have a limited capacity, these additional visual elements will consume part of this capacity regardless and thus leave *less capacity to focus on the relevant elements*. Therefore, it may be that whatever the type of distractor (e.g., advertisements, *clickbait*, etc.), the effects are similar—or, it may be that some distractors have a greater impact than others on the user's search experience. In this paper, we perform one of the first studies examining and comparing the influence of different types of distractors on information seeking. Given that people's cognitive abilities play a defining role in how well they can handle visual complexity, we also examine whether people who have higher *perceptual speed* and who are less prone to *cognitive failures* are able to be better cope with the different distractors presented to them.

2 BACKGROUND

Adversarial elements on the Web come in a variety of forms. These include sponsored links, banners, animations, native adverts, sponsored content, and clickbait [10, 11, 54, 57]. When presented on a web page, these elements often try to lure the user to another site to expose them to even more advertising, or try to sell them a product, service, or brand. Given advertising-based business models that are commonly employed, sites are a juxtaposition of content and advertisements, sponsored links, and related content/recommendations. As such, it is almost impossible to search and browse without encountering such distractions [33, 39]). The peril of this "Attention Economy" is that *every presented element on every site is vying for everyone's attention* [26].

2.1 Attention

Attention is considered as an individual's cognitive ability to focus on relevant items while ignoring the non-relevant ones [58]. Broadly, attention is considered to be controllable and selective [18]—but also limited [30, 43]. In the context of adversarial elements, people try to avoid paying attention to advertisements [20, 56]. However, even when such elements are within an individual's peripheral vision, they can still consume their attention [27]. As such, they are absorbed to a degree—even if avoided [59]. Despite people's efforts to avoid such distractions, some people have been shown to be more prone to distractions than others, which is a type of *cognitive failure* [19]. Since attention is limited, the amount of attention available to process relevant information is also reduced by the very presence of these distracting elements [43].

Since attracting the attention of users is the goal of advertisers, clickbaiters, and content creators [45], different techniques are used to draw the user's attention (e.g., sensationalised headlines, eye-catching images, etc.) and to attract views and clicks [36]. The more of a user's attention is allocated to processing these additional elements, the less attention will be available for the search task at hand—and the more damaging it may be to the task outcomes [47, 63]. For example, it may result in insufficient cognitive resources to attend to, store, and later recall relevant content [48]. Indeed, studies in *Interactive Information Retrieval (IIR)* have shown that the addition of elements (such as advertisements, entity cards, images, videos, etc.) on SERPs and content pages increases page complexity [8], temporal costs [9] as well as people's cognitive load [15, 32, 42, 44, 53]. This in turn, makes it more difficult to find relevant information. Taken together, this body of work suggests that regardless of the additional element types added to the interface, less attention will be available for the primary search task—which may result in lower search performance [47].

2.2 Perceptual Speed and Cognitive Failure

How well people can filter out distractions, identify relevant content and ignore non-relevant content depends on their cognitive abilities. In IIR, numerous studies have examined how different cognitive abilities—such as working memory, associative memory, visualisation ability, perceptual speed, inhibition, dyslexia, etc. [4, 7, 12, 17, 22, 24, 60]—affect and influence search performance and behaviours. From these studies, it has emerged that ***Perceptual Speed (PS)*** tends to play the biggest role in shaping an individual's search experience [31].

Perceptual speed can be defined as *an individual's accuracy and speed to view, scan, and compare information during visual search tasks* [4]. Essentially, it is the ability of an individual to identify or compare relevant elements among non-relevant elements (distractors) [55], as cited by Arguello and Choi [7]. In the IIR literature, a number of key findings regarding perceptual speed have emerged. People with *High PS*:

- *(i)* issue more effective queries obtaining higher precision and recall scores [5];
- *(ii)* are faster at identifying the first relevant item [2];
- *(iii)* are more active searchers (i.e., they issue more queries, longer queries, and visit more documents) [17]; *yet*
- *(iv)* find a similar number of relevant items during their search [2];
- *(v)* report lower workloads [17]; and

(vi) spend less time on the task [17].

Given this prior work, it seems reasonable to assert that searchers with higher perceptual speed would be able to better deal with increased interface complexity. When examining perceptual speed and interface complexity, researchers have found that participants with *Low PS*:

(i) spent longer completing their tasks [60];
(ii) report lower usability [60];
(iii) experience less satisfaction [60];
(iv) struggle to locate relevant information [7]; and
(v) issue more queries of lower quality when using more complex blended and aggregated interfaces (i.e., interfaces with many different element types combined together on the same page) [7].

Taken together, these findings suggest that when distracting or non-relevant elements are included on webpages, it will make it harder for low PS individuals to search effectively. From this, we hypothesise that additional elements (regardless of the type presented) will have a negative impact on searchers with lower PS—more so than those that have a higher PS.

In addition to PS ability, we also consider *how prone people are to distractions*, which is a type of ***Cognitive Failure (CF)***. Broadbent et al. [19] defines a CF as a cognitive error that manifests during a task that an individual would otherwise successfully execute. Typically, a CF is characterised by concentration issues, memory loss, or a decreased perception ability—typically as a result of stress. We hypothesise that if a searcher is more prone to distraction, they will not be able to effectively filter out additional elements. Instead, they will become distracted, and this will negatively impact their search performance.

2.3 Advertisements and Search

Studies that describe how advertisements influence search behaviour have so far mainly focused on how people perceive such elements in terms of their quality and relevance to the user [3, 21, 25, 28, 35, 49], and how memorable and engaging such elements are [13, 37, 41, 45, 46, 51, 56]. Broadly, advertisements are seen to be annoying and result in a decrease of user satisfaction [3]. When people are presented with advertisements that are not related or congruent with the search task in hand (or the content), they are perceived more negatively [21, 40]. As such, there has been a drive to personalise and tailor advertisements to be as relevant as possible to minimise user annoyance and avoidance [21, 28, 62]. In studies examining how advertisements influence people's attitudes towards sites that combine advertisements within their own organic content, they were also perceived negatively [46]—and rated as less credible when compared to sites without advertisements [25].

Recent work by Foulds et al. [32]—which is the most related to the present study—showed the influence of advertisements on information seeking, when they were either present or not present on a news search site. They found that the presence of advertisements—in addition to being perceived negatively (more frustrated, less satisfied, more annoyed)—also resulted in lower search performance, and significantly longer task completion times. However, there were some reported positives. The presence of advertisements were seen by participants as making the site more engaging and less boring than compared to when none were shown (as no images or additional elements were shown in the non-advertisements condition). In addition to this recent work, Im et al. [41] studied the interplay of advertisements within the context of web pages to better understand how people process and evaluate the content under visual complexity, finding that the inclusion of some advertisements also led to improved perceptions about the site. However, too many advertisements led to more negative perceptions. Again, this suggests some positive effects of advertisements.

Kim [45] specifically examined the influence of advertisements on how people processed news articles. Their findings also confirm that people's attention to the news is reduced when advertisements are present. However, they hypothesised that when news articles included images, they may mitigate these negative effects [45]. Hsieh and Chen [39] subsequently found that participants were less likely to examine advertisements when the content pages contained an image (an *attractor*) than when it did not—and thus claimed that attention was diverted to the relevant article, rather than towards the advertisements. Kim [45] also found news images helped participants recall more about the articles when they had enough resources available to process the information. Otherwise, it increased cognitive load—and reduced the ability of participants to recall task-related concepts.

Taken together, these findings suggest that the presence of advertisements generally has a negative impact on people's information seeking experiences. However, it is also important to note that other visual elements included on a page may also have positive and negative impacts. While the focus of most research has been directed towards examining the influence of advertisements when information seeking, less effort has been paid towards investigating how other *"distractors"* (or *"attractors"*) may influence people's search experience, performance, and behaviour. For example, when sponsored links—which are designed to entice clicks (such as clickbait [11]) are included—then such elements may be more detrimental.

This paper aims to compare the influence of different types of distractors to determine the influence on people's search experience—and whether an individual's ability to filter distractions (i.e., their propensity for distraction) and their ability to identify relevant from non-relevant (i.e., their perceptual speed) influences whether they can cope (or not) with such distractions.

3 METHODOLOGY

To examine the impact of distractors on searching we performed a within-subjects study using simulated work tasks [14] to situate the information seeking. Participants were asked to collect relevant information in order to write a report on a specific topic. To achieve this, they needed to undertake an exploratory search session [52] to identify several different examples from news articles. For this work task, we used our bespoke news search engine.

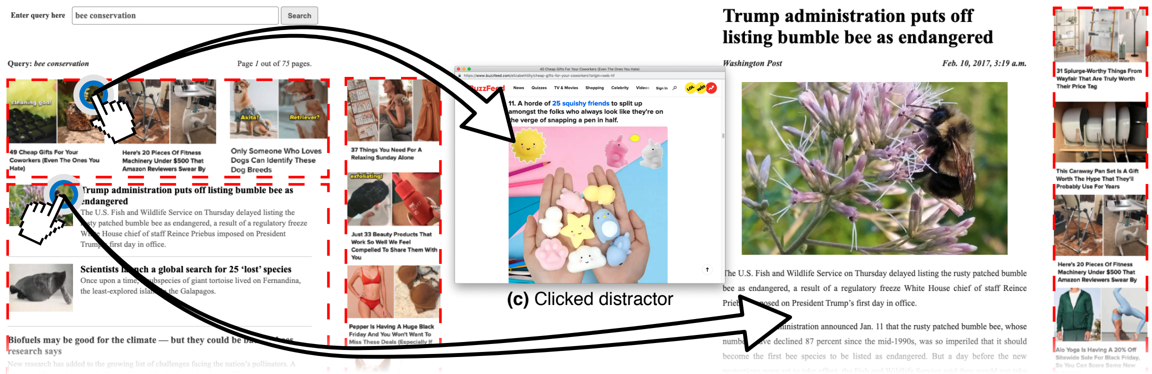
Given this context, we investigated three main research questions: *How does the presence of different types of distractors influence...*

- **RQ1** ...the observed search behaviours of participants;
- **RQ2** ...their assessment of workload, satisfaction, effort, etc.; and
- **RQ3** ...their overall search performance?

Our within-group variable for the study was the following four types of distractors.

- **None**: No distractors were shown—this acts as our baseline.
- **Recommendations**: Contemporary news articles and opinion pieces were shown for the same site.
- **Sponsored Links**: Sponsored links were shown, advertising quizzes, recipes, shopping, etc.
- **Ads**: Animated advertisements were shown advertising major brands.

To explore whether the cognitive abilities of participants also played a role in how they behaved, felt and performed, our study included two between-group independent variables based on the *(i)* perceptual speed and *(ii)* cognitive failures of participants.



(a) SERP view with distractors **(b)** Document view with distractors

Fig. 2. Screenshots of the experimental search interface, complete with distractors from **Sponsored Links**. *(a)* presents the SERP interface; *(b)* shows the document view; and *(c)* illustrates a popup shown for a sample distractor when clicked on.

3.1 Collection and System

For this study, we used the *TREC Washington Post Corpus (WAPO)* collection from the *TREC Common Core 2018*[1]. The collection consists of 608, 180 news articles and blog posts from January 2012 through to August 2017, and 50 topics. From these 50 topics, we selected five to use as part of this study (see §3.4). Our retrieval system used a pure Python search engine, *Whoosh*[2], using the *BM25* retrieval algorithm (where $\beta = 0.75$). We used a standard Web search interface, as illustrated in Figure 2. Our SERP (shown in Figure 2(a)) presented a total of ten results per page. Result summaries were composed of:

(i) the title of the news article;
(ii) the leading sentence; and
(iii) a thumbnail of the image associated with the article (if available from the WAPO website).

Participants could click either on the article's title or associated image to view the news article, which took them to the document view which is shown in Figure 2(b). Participants could bookmark/save the articles that they judged to be relevant and view a list of documents that they had bookmarked/saved.

For the conditions where distractors were shown (**Ads**, **Sponsored Links**, and **Recommendations**), participants were initially presented with a search page, with a query box and *chumbox* (see Figure 1). The chumbox consisted of three distractors. Once a query had been issued, the SERP was presented. As shown in Figure 2(a), three distractors were placed beneath the query box, with four placed on the right rail. When viewing a document, four distractors were also placed to the right of the article's content (Figure 2(b)); a row of three distractors were placed beneath the second paragraph to mimic how distractors are presented in many popular websites. Distractors for the given condition were randomly selected from the available pool when displayed on each page.

[1] https://trec-core.github.io/2018/—last accessed October 18th, 2022.
[2] `$pip install whoosh==2.7.4`

Table 1. SAM scores [16] for each of the three distractor types. $^{\mathcal{R}}$(**Recommendations**) and $^{\mathcal{A}}$(**Ads**) denote significant differences between respective distractor types.

Distractor	Valence	Arousal	Dominance
Ads	**5.10**$^{\mathcal{R}}$	**4.58**$^{\mathcal{R}}$	**5.14**
Sponsored Links	4.76$^{\mathcal{R}}$	4.30$^{\mathcal{R}}$	3.53$^{\mathcal{A}}$
Recommendations	3.74	3.86	3.80$^{\mathcal{R}}$

3.2 Distractors

Distractors were sourced from three different websites. **Sponsored Links** were extracted from *Buzzfeed*, **Recommendations** *(news articles)* were taken from the WAPO website[3], and **Ads** were extracted from various animated GIF hosting websites, such as *Giphy* and *Tenor*[4]. The prominent focus of each selected animated GIF was a specific brand or product. A random selection of approximately 50 items from each website were taken across categories such as *Shopping*, *TV*, *Animals*, *Quizzes*, and *Current Affairs*. Each distractor was downloaded, cropped and resized to the same dimensions (200 × 210). Target URLs for **Sponsored Links** and **Recommendations** distractors were also recorded. For **Ads**, the website of the associated brand and/or product was recorded in its place.[5] URLs were stored such that if a participant clicked on a distractor, they would be taken to the respective page in a popup window, keeping the search interface for the experiment present behind the new popup. Distractors were vetted internally to ensure that there were no inappropriate, distasteful, or offensive content. We collected a total of 121 distractors, consisting of 43 **Recommendations** articles, 44 **Sponsored Links**, and 34 **Ads**.

To measure the differences between each type of distractor, we performed an initial study to further vet the distractors and their quality across several dimensions. Using the *Self Assessment Manikin (SAM)* scales rating [16], each item was shown to an annotator, who rated it in terms of its:

(i) **valence** *(how pleasing is this item?)*;
(ii) **arousal** *(how exciting is this item?)*; and
(iii) **dominance** *(how distracting is this item?)*.

For this study, we recruited $n = 50$ participants to rate 20 items each, and compensated the equivalent of US$3 for their time (approximately 10 minutes). Items were randomly selected to avoid ordering effects. A total of 978 annotations were collected (approximately eight per distractor); some participants did not fully complete the survey.

Table 1 presents the mean ratings associated with each of the three distractor types from our pilot study. Using an ANOVA test, we detected significant differences across the ratings of *valence* (pleasing), *arousal* (exciting), and *dominance* (distracting). Follow-up tests (using the Bonferroni correction) indicated that **Recommendations** elements were significantly less pleasing than **Ads** ($p < 0.001, np^2 = 0.568$) and **Sponsored Links** ($p < 0.001, np^2 = 0.454$). In terms of how arousing, **Recommendations** articles were also significantly less exciting than **Ads** ($p < 0.001, np^2 = 0.314$) and **Sponsored Links** ($p = 0.032, np^2 = 0.192$). Pertaining to dominance, **Ads** were seen as significantly more distracting than **Sponsored Links** ($p < 0.001, np^2 = 0.697$) and **Recommendations** elements ($p < 0.01, np^2 = 0.549$). These results gave us confidence that sufficient differences were present between the three distractor types—and thus between our three experimental interface conditions.

[3] See https://www.buzzfeed.com/ and https://www.washingtonpost.com/—last accessed October 18$^{\text{th}}$, 2022.
[4] See https://giphy.com/ and https://tenor.com/—last accessed October 18$^{\text{th}}$, 2022.
[5] E.g., a GIF of a *McDonalds* cheeseburger redirected to https://www.mcdonalds.com.

3.3 Measuring Cognitive Abilities

To measure the cognitive capabilities of participants in relation to searching, we measured their perceptual speed—as this has been shown to play a major role in IIR studies [4, 6, 17, 32, 60]. Furthermore, we used the *Cognitive Failures Questionnaire (CFQ)* [19] to measure the extent to which participants were prone to distractions.

Perceptual Speed Test: Finding ϵ and ¥'s Participants were given two minutes to navigate through five pages, containing a total of 100 artificially generated 'words' composed of non-alphanumeric symbols. Participants were asked to find and mark the 'words' that contained *both* ϵ and ¥ symbols. This is similar to the *Finding A's test* defined by Ekstrom et al. [29]. However, later work by Ackerman and Cianciolo [1] showed that the *Finding ϵ and ¥'s test* has a higher correlation with PS than other PS tests. A total of 30% of the 'words' contained both symbols. The score for each participant was computed by considering how many words they *correctly* identified, minus how many they *incorrectly* identified (as per the instructions used by Ekstrom et al. [29]). Participants were able to practice on a small sample, where they were given feedback on their selections. To continue to the full test, they were required to correctly identify all the words containing both ϵ and ¥. This served as a check to ensure that they had understood the instructions correctly.

Cognitive Failures Questionnaire The CFQ developed by Broadbent et al. [19] was used to assess the frequency with which people suffer from reduced attention, and subsequently yield lower performance when performing tasks. Consisting of 12 questions [61], the survey uses a five-point scale from 1 *(never)* to 5 *very often*. Questions probed on aspects of behaviour, with examples including:

(i) whether a participant was reading something, found they weren't thinking about it, and had to read it again;
(ii) whether they daydream when they should be listening to something; and
(iii) whether the participant finds they forget whether they've turned off a light, fire, or locked the door.

The final score for each participant was the sum over all question responses. A higher score indicates that they are more likely to exhibit cognitive failure, and thus be more prone to distraction.

3.4 Search Topics and Tasks

From the 50 topics associated with the WAPO collection, five (four main plus one practice) were selected for this study. The four main topics chosen were:

- **№ 341** *Airport Security*;
- **№ 347** *Wildlife Extinction*;
- **№ 363** *Transportation Tunnel Disasters*; and
- **№ 408** *Tropical Storms*.

These topics were selected as they offered similar levels of performance on our experimental system, and the underlying task for each was similar. We also selected topic **№ 367** *(Piracy)* for use as a practice topic for participants to familiarise themselves with our experimental system.

With these topics, participants were asked to undertake a simulated work task [14], as previously mentioned. They were explicitly instructed to find as many different and relevant examples as *they felt necessary to provide evidence for the article*. Participants were instructed to first save documents that contained relevant examples, and then later (post-task) asked to recall as many of the examples that they had found. This was done so that we could quantify how much evidence they found (and learnt) about the topics. Using **№ 408** as an example, participants were asked to find a number of different tropical storms that caused widespread destruction and loss of life. Post-task, they were asked to recall the names of these storms, and where they occurred. Examples requested for the other topics were:

- **№ 341** the airport and security measures employed;
- **№ 347** the country and species; and
- **№ 363** the name of the tunnel and the cause of the disaster.

3.5 Outcome Measures

The dependent variables for this study were split into three groups: *search behaviours*, *search experiences*, and *search performance*.

Measuring Search Behaviours To provide us with insights into the search behaviours of participants, we logged a number of interactions with the search system. These included:

- the number of queries issued;
- the number of documents clicked/viewed (and by proxy, the number of documents clicked *per query*);
- the total number of SERPs viewed;
- the total number of documents saved (identified as relevant); and
- by incorporating TREC judgments, the total number of saved, TREC relevant documents.

From the recorded interaction logs, we could also compute a series of time-based measures, including:

- the total session time per task;
- the total time spent querying per topic (from query box `focus` to `submit` events);
- the total time spent on SERPs;
- the time spent examining each result summary (snippet);
- the total time spent examining documents; and
- *per-unit* times, such as the time spent examining *per document.*

Measuring Search Experience After each task, participants completed a subset of the *Post System Search Usability Questionnaire (PSSUQ)* [50]. The statements were slightly adapted to be more specific to our study's conditions, and broken down into two parts: ***(i)*** focusing on the *information and task found using the system*; and ***(ii)*** their *perceptions about the system*. Statements for ***(i)*** included whether participants:

(a) found it *difficult* to complete the task;
(b) were able to complete the task *quickly*;
(c) could *concentrate* on the task;
(d) found the results *effective* in completing the task; and
(e) were *satisfied* with what they had found.

For part ***(ii)*** of the survey, questions included whether participants:

- *(f) learnt* about the topic using the *system*;
- *(g)* found the system *annoying*;
- *(h) liked* using the system;
- *(i)* felt *productive* using the system; and
- *(j)* whether they felt *satisfied* using it.

Table 2. Breakdown of participant groupings with respect to *Perceptual Speed (PS)* and *Cognitive Failures (CF)*.

	High CF	**Low CF**	**Total**
High PS	21	28	49
Low PS	26	27	53
Total	**47**	**55**	**102**

Statements were rated on a scale of 1 *(strongly disagree)* to 6 *(strongly agree)*. The PSSUQ survey was then followed by the *NASA Task Load Index* [38] to obtain the perceptions on the participants' workloads in terms of *cognitive*, *physical*, *temporal*, *performance*, *frustration*, and *effort*. Statements were rated from 1 *(low)* to 9 *(high)*. The total workload computed from the instrument was the sum of the six dimensions.

Measuring Search Performance By using the *TREC Common Core 2018* relevance judgements, we were also able to provide an estimation of a participant's search performance. For each query that was submitted by a participant, we evaluated the query's *P@*10. Given all of the documents that participants clicked on and saved, we could use the aforementioned relevance judgements as a ground truth, allowing us to compute the *accuracy* of a participant's searching ability. This was summarised as the proportion of relevant items saved (i.e., documents that are identified as relevant in the relevance judgements) vs. the total number saved. Finally, overall success was determined by the number of *examples recalled*, which was how many relevant examples they could recall post-task. Each recalled example was checked against saved documents, and counted towards their score if correct.

3.6 Experimental Procedure

To ensure that topic and ordering effects were minimised, we used a *Graeco-Latin Square* experimental design to rotate topics and interfaces. Participants were recruited from the online crowdsourcing platform *Prolific*, and upon successful completion of the study, received the equivalent of US$20 for their time. Participants were recruited from the United Kingdom, and were asked to complete the study on a desktop or laptop computer. Programmatic checks were used to ensure that these requirements were met, and that they could see distractors in their browser before they began (ensuring *ad blockers*, if installed, would not interfere). Participants gave their consent after reading the on-screen information sheet, and then completed the demographics survey. To decide whether they wanted to continue with the study, they were first given the practice task, as mentioned in §3.4. Once complete, their practice accuracy was shown. After looking at this, participants decided to continue, or not. If they chose to do so, they then moved on to performing the PS test (§3.3), before moving to the four search tasks. After each of the search tasks, participants completed the post-task surveys, as outlined in §3.5. Lastly, they filled in the CFQ (§3.3). Ethics approval was obtained from the University of Strathclyde Department of Computer and Information Sciences Ethics Committee (№ 1636).

3.7 Participant Demographics

Of the 102 participants who fully completed the study, 69 were female, 32 were male, with 1 not disclosing their sex. Their ages ranged from 18 to 65, with most participants in the 18 – 25 age bracket (35%). Other age brackets and percentages included: 26 – 35 (27%); 36 – 45 (16%); the remainder were over 45. 86% of participants were native English speakers, 11% were bilingual, and 3% had professional working English. 83% were university or college graduates, with the remaining 17% stating that they had completed their high school education.

Table 3. Search behaviours, with the mean number of actions performed per user, per task. Here, *Q.* denotes *Queries*, and *Docs.* denotes *Documents*. Asterisks (*) denote a significant difference between groups ($p < 0.05$). Highest accuracy values are **bolded**.

Condition	#Q.	Docs./Q.	Depth/Q.	#Clicked	Documents Saved		Accuracy
					Docs.	Rel. Docs.	
None	4.49 ± 4.28	3.64 ± 3.08	15.98 ± 20.17	8.32 ± 3.95	5.70 ± 2.60	3.93 ± 1.79	0.73 ± 0.25
Recommendations	3.83 ± 3.04	3.45 ± 3.37	13.47 ± 11.69	7.60 ± 3.62	5.38 ± 2.37	3.73 ± 1.88	0.72 ± 0.24
Sponsored Links	4.06 ± 3.70	3.78 ± 3.23	16.36 ± 15.79	8.39 ± 3.41	5.68 ± 2.39	4.05 ± 1.88	0.73 ± 0.23
Ads	4.14 ± 4.27	3.37 ± 2.49	16.18 ± 15.33	7.67 ± 2.83	5.42 ± 2.10	**4.11 ± 1.79**	**0.77 ± 0.21**
High PS	4.22 ± 3.82	3.41 ± 2.73	13.68 ± 12.63*	8.11 ± 3.15	5.48 ± 2.06	**4.12 ± 1.76**	**0.77 ± 0.21***
Low PS	4.04 ± 3.89	3.70 ± 3.34	17.24 ± 18.56*	7.89 ± 3.78	5.61 ± 2.62	3.80 ± 1.90	0.71 ± 0.24*
High CF	3.98 ± 4.02	4.08 ± 3.59*	18.08 ± 17.20*	8.27 ± 3.62	5.82 ± 2.57*	**4.17 ± 1.94***	**0.74 ± 0.24**
Low CF	4.25 ± 3.71	3.13 ± 2.46*	13.38 ± 14.68*	7.77 ± 3.35	5.32 ± 2.16*	3.78 ± 1.74*	0.73 ± 0.23

Groupings Given our participant's scores on the PS test and CF Questionnaire, we assigned them to different groups for each cognitive ability. Groups were *High* and *Low* based on their median score — in line with prior studies [6, 17, 60]. For PS, the median score was 19, while for CF, the median score was 25 (out of a maximum 60). Table 2 shows the breakdown of participants over each grouping. Interestingly, a Chi-squared test found that there was no relationship between PS CF scores of participants ($p = 0.668$). Given these groupings, we hypothesised that participants who had high PS but were less likely to be distracted (low CF) would be the most efficient and effective in completing their search tasks, and vice versa.

4 RESULTS

To determine whether there were any significant differences between the conditions and the measures we examined, we performed n-way ANOVAs. The main effects were examined with $\alpha = 0.05$, with post-hoc analyses performed using pairwise t-tests using the *Benjamini-Hochberg FDR* correction. For the reported tests, the F-score, p-value, and effect size η_p^2 are reported to 3 decimal places (where value/ranges of η_p^2 indicate: *small* (< 0.06); *medium* (0.06–0.14); or *large* effect sizes (> 0.14) [23]). ± values reported in subsequent tables denote the *standard deviation* of the mean.

4.1 Search Behaviours

Table 3 reports the average number of behavioural-related events as computed from the interactions of participants over each condition/group. Also included is the performance-related accuracy measure that reports how well participants were able to identify the relevant from the non-relevant (i.e., the proportion of relevant items saved vs. the total number saved). Firstly, in terms of the different interface conditions, we observe very little difference in the behaviours of participants. Regardless of the condition, for a given search task, participants issued approximately four queries, clicked on 3 – 4 documents per query, and examined content to a depth of around 13 – 16, on average. During their sessions, participants examined/clicked a total of approximately eight documents, saved around 5 – 6 of the documents, identified 4 relevant items, and thus achieved an accuracy of around 0.75. ANOVA testing revealed that there were no significant differences in terms of these behaviours. However, when examining how our groupings performed, we see a number of differences for the CF groups. Firstly, **High CF** participants examined approximately one more document per query than **Low CF** participants (4.08 vs. 3.13, $F(1) = 9.518, p = 0.002, np^2 = 0.024$). **High CF** inspected approximately five more results on the SERP (18.08 vs. 13.38, $F(1) = 8.180, p = 0.0045, np^2 = 0.020$), saved on average 0.5 more documents

Table 4. Average timings (in seconds) for session and various actions. High PS participants issued queries and examined snippets faster, and were quicker to click—while they spent more time per document, and more time to save the first document. $^{\mathcal{N}}$(**None**), $^{\mathcal{A}}$(**Ads**), and $^{\mathcal{R}}$(**Recommendations**) denote significant differences between respective distractor types.

Condition	Session Time	Time per... Query	Time per... Doc.	Time per... Snippet	Time to First... Click	Time to First... Saved Doc.
None	455.45 ± 277.64	**8.83 ± 5.50**	26.68 ± 20.89	**3.08 ± 2.05**	14.41 ± 11.10$^{\mathcal{A}}$	68.90 ± 74.79
Recommendations	**437.84 ± 239.58**	10.00 ± 9.60	**24.20 ± 17.02**	4.26 ± 5.51	15.01 ± 10.66$^{\mathcal{A}}$	66.36 ± 52.76
Sponsored Links	466.71 ± 253.88	9.66 ± 8.13	26.74 ± 17.58	3.92 ± 4.05	**13.42 ± 9.72**	**66.37 ± 65.71**
Ads	454.80 ± 277.64	10.05 ± 6.92	26.61 ± 20.76	3.84 ± 4.81	17.99±13.28$^{\mathcal{R},\mathcal{N}}$	76.28 ± 73.51
High PS	**441.13 ± 218.72**	**8.20 ± 4.95***	27.67 ± 19.18*	**3.27 ± 4.21***	**14.58 ± 10.98**	76.30 ± 65.14*
Low PS	465.78 ± 281.67	11.01 ± 9.40*	**24.50 ± 18.95***	4.26 ± 4.58*	15.81 ± 11.70	**62.92 ± 68.52***
High CF	476.16 ± 265.29	9.55 ± 6.28	27.31 ± 21.13	4.37 ± 5.78*	15.85 ± 11.75	73.54 ± 76.23
Low CF	**435.32 ± 241.11**	**9.70 ± 8.66**	**25.03 ± 17.26**	**3.29 ± 2.79***	**14.68 ± 11.02**	**66.15 ± 58.63**

(5.82 vs. 5.32, $F(1) = 4.432, p = 0.036, np^2 = 0.011$)—and found on average 0.5 more relevant documents as a result (4.17 vs. 3.78, $F(1) = 4.718, p = 0.030, np^2 = 0.012$). When looking at PS groups, we found no significant differences in terms of how participants behaved—bar the depths to which they inspected documents. **High PS** participants inspected around 3 – 4 fewer items per SERP than **Low PS** participants ($F(1) = 4.525, p = 0.034, np^2 = 0.011$). No interaction effects were noted between PS and CF.

Table 4 presents an overview of the average time (in seconds) taken for different actions. We can see that for the interface without distractions **None**, participants spent less time per query—and were faster to process result snippets—than the other conditions. However, no significant differences were observed, except for the time to first click ($F(3) = 3.479, p = 0.016, np^2 = 0.026$), where follow-up tests revealed that participants took longer to click on the **Ads** interface than on the **Recommendations** ($p = 0.008$) and **None** ($p = 0.008$) interfaces. In terms of cognitive abilities, we observed differences across the board for our PS groups, where **High PS** participants spent approximately 20 seconds less completing their tasks, took approximately three seconds less per query, one second less per result snippet, and clicked on the first item approximately one second earlier. However, **High PS** participants also spent about three seconds longer assessing documents, and took around 15 seconds longer deciding on what item to save first. Of these, we found significant differences for PS over: time per query ($F(1) = 13.953, p < 0.001, np^2 = 0.034$); time per document ($F(1) = 5.746, p = 0.0172, np^2 = 0.021$); time per snippet ($F(1) = 4.687, p = 0.031, np^2 = 0.012$); and for time to the first saved document ($F(1) = 6.337, p = 0.0122, np^2 = 0.016$). For our CF grouping, we observed that **High CF** participants spent approximately 40 seconds longer completing their task, spent about one second more per result snippet, and about seven seconds more to save the first relevant item. However, the only significant difference for our CF grouping was in terms of the time spent per result snippet (4.37 vs. 3.29, $F(1) = 5.546, p = 0.019, np^2 = 0.014$). Again, we found no interaction effects across groups and conditions.

4.2 Search Experience and Perceptions

NASA TLX In terms of workload, we observed the following trends. Firstly, the total workload for each of the different interfaces increases from: 29.31 on **None**; 30.93 on **Recommendations**; 31.36 on **Sponsored Links**; and 31.74 on **Ads**. In line with total workload, participants reported that when using **None**, they had the lowest cognitive, physical, effort, levels of frustration, and overall workload. On the **Ads** interface, participants reported the highest loads, with the highest amount of frustration and effort. Across conditions however, it was only in terms of frustration that any

Table 5. Results of the PSSUQ for task and system questions. Refer to §3.5 for details on the questions asked. $^{\mathcal{N}}$(**None**), $^{\mathcal{A}}$(**Ads**), $^{\mathcal{R}}$(**Recommendations**), and $^{\mathcal{S}}$(**Sponsored Links**) denote significant differences between respective distractor types.

	PSSUQ Question	None	Recommendations	Sponsored Links	Ads
Info./Task	*(a)* **Difficulty**	$2.97^{\mathcal{A},\mathcal{R}}$	$3.36^{\mathcal{A}}$	3.43	$3.60^{\mathcal{N}}$
	(b) **Quick**	3.60	3.48	3.23	3.14
	(c) **Concentration**	$4.38^{\mathcal{A},\mathcal{R},\mathcal{S}}$	$3.69^{\mathcal{N}}$	$3.82^{\mathcal{N}}$	$3.44^{\mathcal{N}}$
	(d) **Effective**	3.63	3.41	3.16	3.16
	(e) **Info. Satisfaction**	3.62	3.48	3.23	3.33
System	*(f)* **Learnt**	3.70	3.62	3.57	3.57
	(g) **Annoying**	$2.54^{\mathcal{A},\mathcal{R},\mathcal{S}}$	$3.71^{\mathcal{N}}$	$3.34^{\mathcal{N}}$	$3.78^{\mathcal{N}}$
	(h) **Liked**	$3.77^{\mathcal{A},\mathcal{R},\mathcal{S}}$	$3.18^{\mathcal{N}}$	$3.26^{\mathcal{N}}$	$3.08^{\mathcal{N}}$
	(i) **Productive**	$3.86^{\mathcal{A},\mathcal{R},\mathcal{S}}$	$3.45^{\mathcal{N}}$	$3.28^{\mathcal{N}}$	$3.23^{\mathcal{N}}$
	(j) **System Satisfaction**	$3.83^{\mathcal{A},\mathcal{R},\mathcal{S}}$	$3.40^{\mathcal{N}}$	$3.42^{\mathcal{N}}$	$3.22^{\mathcal{N}}$

significant differences were observed ($F(3) = 2.9465, p = 0.033, np^2 = 0.022$). Follow-up tests concluded that participants found the condition without distractions (3.89) to be significantly less frustrating than **Recommendations** ($p = 0.021$), **Sponsored Links** (4.71, $p = 0.038$) and **Ads** (4.83, $p = 0.009$) conditions.

When examining with workload by PS, participants in the **High PS** group reported a significantly lower workload than those in the **Low PS** group (29.69 vs. 31.94, $F(1) = 6.579, p = 0.011, np^2 = 0.023$). Participants also reported significantly lower cognitive load ($F(1) = 9.418, p = 0.002, np^2 = 0.023$), lower temporal load ($F(1) = 14.274, p = 0.0002, np^2 = 0.035$), higher performance ($F(1) = 7.541, p = 0.006, np^2 = 0.027$), and experienced less effort ($F(1) = 6.018, p = 0.015, np^2 = 0.021$) than those in the **Low PS** group. These findings are consistent with prior works [7, 17]. For our CF groupings, participants in the **High CF** group reported higher overall workload. However, it was not significantly different to that reported by the **Low CF** group (31.14 vs. 30.59).

Task Perceptions The upper part of Table 5 reports the subset of PSSUQ questions relating to the task and information found during their search session. First, participants consistently reported that the task was: *(a) less difficult*, *(b) quicker*; *(c) easier to concentrate*; *(d) more effective*—and that they were more satisfied with their task outcome; and *(e)* when no distractions were present on the interface. Significance testing revealed that for difficulty there was a significant difference ($F(3) = 2.967, p = 0.033, np^2 = 0.033$). Follow-up tests indicated participants found that when no distractors were present, it was significantly less difficult than when **Ads** ($p = 0.004$) or **Recommendations** ($p = 0.032$) *were* present. There were also significant differences ($F(3) = 8.981, p = 0.0, np^2 = 0.064$) when looking at concentration. Follow-up tests indicated that there was a significant difference between **None** and all the other conditions (**Ads** $p < 0.001$, **Sponsored Links** $p < 0.001$, **Recommendations** $p = 0.003$). Furthermore, participants also reported that the **Ads** interface was the most difficult and hardest to concentrate on. It required the most time, lowered their effectiveness, and resulted in the least satisfaction with what they had found. There were however no significant differences between the **Ads** interface and the **Recommendations** and **Sponsored Links** interfaces—suggesting that participants perceived them in a similar manner. Interestingly, in terms of our PS and CF groupings, there were no significant differences reported across these dimensions (not shown in Table 5)—and there were also no interaction effects.

System Perceptions. The lower part of Table 5 reports the system-focused questions from the PSSUQ. Again, we see that participants consistently rated the interface without distractors as the least annoying, most liked, most productive, and most satisfying where they felt they learnt the most. The **Ads** interface again was generally rated the worst, being

Table 6. The average number of topic concepts recalled, along with other concepts recalled (i.e., distractors)—and the mean total.

Condition	Concepts	Other	Total Recalled
None	**3.80**$^{\mathcal{A},\mathcal{R},\mathcal{S}}$	-	**3.80**
Recommendations	3.13$^{\mathcal{N}}$	0.46	3.59
Sponsored Links	3.06$^{\mathcal{N}}$	0.47	3.53
Ads	2.99$^{\mathcal{N}}$	0.55	3.54
High PS	**3.41**	**0.49***	**3.91***
Low PS	3.08	0.25*	3.33*
High CF	3.09	0.21*	3.30*
Low CF	**3.37**	**0.50***	**3.87***
High PS/Low CF	**3.57**	**0.71***	**4.28***
High PS High CF	3.20	0.21*	3.42*
Low PS/Low CF	3.16	0.30	3.45
Low PS/High CF	3.00	0.20	3.20

the most annoying, least liked, least productive, and least satisfying. When investigating whether there were significant differences, we found that over *(g) annoying* there was a clear difference ($F(3) = 13.195, p < 0.001, np^2 = 0.091$). Follow-up tests showed a difference between **None** and all the other conditions (**Ads** $p < 0.001$, **Sponsored Links** $p < 0.001$, **Recommendations** $p < 0.001$). Significant differences were found for *(h) liked* ($F(3) = 5.059, p = 0.002, np^2 = 0.037$), with follow-up tests showing that participants liked no distractions compared to all the other conditions (**Ads** $p = 0.001$, **Sponsored Links** $p = 0.004$, **Recommendations** $p = 0.013$). For *(i) productivity* ($F(3) = 4.307, p = 0.005, np^2 = 0.032$) and *(j) system satisfaction* ($F(3) = 3.733, p = 0.011, np^2 = 0.027$), **None** was again perceived as more *(i)* productive and *(j)* satisfying than the other interfaces.

There were few differences across PS and CG groupings. Differences that did exist however included that the **High CF** participants liked the interfaces significantly more than **Low CF** participants (3.52 vs 3.17, $F(1) = 6.097, p = 0.014, np^2 = 0.015$), and there were interaction effects for *(f) learnt* over PS and CG groupings ($F(1) = 3.965, p = 0.0471, np^2 = 0.01$)—where for **Low PS** participants, those that had **High CF** reported learning more than those with **Low CF** (3.84 vs. 3.45, $p = 0.041$).

4.3 Performance and Concepts Recalled

In terms of retrieval effectiveness, we found no significant differences across the interface conditions, nor by examining the cognitive abilities of the participants. Observed mean $P@5$ and $P@10$ was approximately 0.57 and 0.54 respectively. While **High PS** and **Low CF** participants obtained slightly higher performance than average—and **Low PS** and **High CF** participants were slightly lower than average—no significant differences were observed.

However, in terms of the post-task performance of participants in recalling example concepts, we see a number of striking differences. Table 6 reports the average number of concepts recalled (i.e., examples from relevant documents), other concepts (i.e., examples of advertisements, sponsored links, and recommended news articles that were remembered), and total concepts (the sum of recalled and other concepts). Here, we can see that in terms of the concepts recalled for the task, interface **None** showed that a significantly higher number of concepts were recalled by participants ($F(3) = 5.329, p = 0.001$), $np^2 = 0.039$) when compared to all the other three interfaces (and confirmed by follow-up tests). In terms of the total concepts recalled (i.e., task examples recalled plus and the distractors recalled, if present), we found no significant differences.

Table 7. Mean time spent per distraction, along with the total time spent by participants, over the three conditions utilising distractions. Also included is the total number of times participants were distracted (where they clicked on a distractor).

Condition	Time/Distractions	Total Time	Total Occurrences
Ads ($n = 2$)	7.67±1.15	23.43	3
Recommendations ($n = 2$)	11.50 ±14.85	23.28	2
Sponsored Links ($n = 8$)	20.53 ±13.73	349.03	17
Any ($n = 10$)	17.52±13.36	395.74	22

In terms of perceptual speed, **High PS** participants recalled more task concepts, significantly more other concepts ($F(1) = 5.548, p = 0.019, np^2 = 0.014$), and a significantly higher number of total concepts recalled overall ($F(1) = 8.461, p = 0.004, np^2 = 0.021$). Turning our attention to cognitive failures, **Low CF** participants recalled more task concepts, significantly more other concepts ($F(1) = 7.649, p = 0.006, np^2 = 0.019$), and a significantly higher total number of concepts ($F(1) = 7.507, p = 0.006, np^2 = 0.019$). Interaction effects were observed for **High PS** participants, such that those also in the **Low CF** group recalled significantly more other concepts—and significantly more total concepts than those in the **High CF** group. In line with expectations, participants in **Low PS** and **High CF** group recorded the lowest number of concepts recalled.

4.4 Post-Hoc Analysis of the Distracted

Given the interactions that we have reported, we decided to further investigate how many participants became *"distracted"* during the course of the experiment.[6] We then examined how long they spent being distracted by the visual elements, and how this influenced their total session times.

Ten participants (approx. 10%) were distracted at least once (i.e., they clicked on a distractor) during their search tasks—and two participants were distracted during two of their search tasks. The average length of each distraction was approximately 18 seconds. Table 7 provides an overview of the distractions—where we can see that approximately 2% of participants were distracted on the **Ads** interface on three occasions, where they spent approx. 8 seconds per advertisement before returning to their task. 2% of participants clicked on distractors on the **Recommendations** interface twice, and spent approximately 12 seconds away per article. Finally, 8% of participants were enticed by distractors presented in the **Sponsored Links** interface 17 times, spending approximately 20 seconds off task per article. The total session time, per task, on average for the distracted was approximately 520 seconds—while for the non-distracted, it was 447 seconds on average. This difference of approximately 30 seconds, per task (approximately 2 minutes overall tasks), however, was not significantly different ($p = 0.088$).

Table 8 shows the retrieval effectiveness (P@10), accuracy, number of concepts recalled, and the overall workload for the **Not Distracted** and **Distracted** groups. We can see that the **Distracted** group experienced lower retrieval performance, had lower accuracy in finding relevant items, and recalled fewer concepts, all while experiencing a higher workload. While we observed no significant differences (as our group size for the **Distracted** was very small), it appears that these participants performed worse on the whole. It is worth noting that these averages are across all four search tasks, as we suspected there may be carryover effects across conditions.

[6] For this analysis, our definition of *"distracted"* was based on whether a participant clicked on at least one distractor, or not.

Table 8. Distracted Group vs. Non Distracted Group over session direction, performance and workload.

Measure	**Not Distracted** $n = 92$	**Distracted** $n = 10$
Session Time	446.53 ± 234.36	518.26 ± 378.66
***P*@10**	0.53 ± 0.22	0.49 ± 0.23
Accuracy	0.74 ± 0.23	0.70 ± 0.23
Concepts	3.29 ± 1.61	2.85 ± 1.84
Workload	30.75 ± 7.17	31.67 ± 8.66

5 DISCUSSION AND CONCLUSION

In this paper, we have examined how different types of visual elements that try to attract people's attention away from their primary work task (and thus serve as a distractor) influence search behaviours and performance. Our findings show that any kind of distractor leads to negative perceptions and poorer overall performance. We observed little difference in terms of search behaviours when compared to each other, and against the baseline without distractors. The lack of differences may be a result of people being able to selectively filter out (and thus avoid) the blocks containing distractors. However, the perceptions of participants revealed that this avoidance comes at a cost—being that it increases the task difficulty, making it harder to concentrate, reducing satisfaction, and inducing greater annoyance. When distractors were not present by contrast, our participants felt more productive and liked the system more. Moreover, participants reported experiencing higher levels of workload when distractors were present. This cognitive drain on people's attention also resulted in lower task performance (in terms of post-task topic recall). Without distractors, participants recalled approximately four examples per task. However, participants could only recall approximately three when distractors were present. When distractors were present, participants could recall approximately 0.5-0.6 distractors, per topic, on average.

These findings suggest that when people were in any one of the three distracted conditions, their limited attention capacity was partly filled by the distractors, resulting in a decrease in their ability to recall information relevant to their actual task. Such visually distracting elements can therefore be seen to be imposing a *"distractor tax"* on people's limited attention—reducing their search efficacy and task outcomes. Here, we found that the difference in post-task topic recall was significantly lower due to the presence of distractors. These findings largely confirm past work which shows that advertisements of various types lead to negative experiences [32]. However, we also show that other types of visual elements (e.g., sponsored links and recommendations) have a similar negative effect on people's search performance and perceptions. As such, search interfaces that try to drive up engagement—even through the promotion of in-house, non-adversarial content—are also leading to unintended consequences for information seekers.

Second, we found greater differences in how people were influenced by different types of distractors (regardless of whether it was from the **Ads**, **Sponsored Links**, or **Recommendations** conditions) depending on their individual cognitive abilities. We observed that perceptual speed was predominately associated with the speed at which people performed their different search interactions (as observed in prior works [2, 7, 17]). We found that **High PS** participants were quicker at querying and examining snippets, but spent longer assessing the documents. They also reported lower overall workload and completed their search task approximately 20 seconds faster than **Low PS** participants, and recalled more concepts overall. However, unlike past work, we found little to no evidence to suggest that distractors influenced the number of search actions participants performed [17, 31]. On the other hand, we found that people's cognitive failures were more indicative of the actions that people performed. For our cognitive failure groupings, we observed that participants with **High CF** examined more documents per query, went to greater depths, saved more documents, and found more relevant items. This resulted in longer overall session times per task, where they spent approximately 40 seconds longer than those in the **Low CF** grouping (or over two minutes over all distractor conditions). Taken together, these findings suggest that participants who were more prone to distraction (**High CF**) explored more of the results. What was particularly surprising was the lack of interaction effects between perceptual speed and cognitive failure, possibly due to the small sample sizes. In terms of task outcomes, however, participants with **High PS** and **Low CF** recalled the most concepts, while those with **Low PS** and **High CF** recalled the least. Restated, those with lower perceptual speeds and a greater propensity to be distracted were affected the most negatively with respect to their post-search task outcomes. Even if interventions by the user were employed (such as using an ad blocker in their Web browser), sites with native recommendations and other elements (such as entity cards, videos, etc.) could also bring about a similar negative effect on outcomes and experiences, despite not being adversarial in nature.

Finally, our post-hoc analysis revealed another dimension to this work. For those who were distracted and clicked on at least one of our distractors, their session lengths for the distractor conditions were considerably longer over all their search tasks than those that did not become distracted (approximately 630 seconds vs. 440 seconds). This suggests that there may be carryover or cumulative effects due to distraction. As the amount of time spent investigating a distractor (i.e., clicking on an **Ads**, **Sponsored Links**, or **Recommendations** element) was approximately 20 seconds. For the 10% of distracted users, it seems that the presence of distractors slowed them down considerably, and beyond the amount of time that they were actually distracted for (i.e., their session time per task was approximately 120 seconds longer, on average—yet they only spent approximately 20 seconds on the distractor).

While we have very few instances of participants being distracted, these results are concerning as 1 in 10 participants were affected. Results show that the search tasks would take approximately two minutes longer when performed under distraction. In addition, this also resulted in a reduction in the number of successfully recalled task concepts. Considering these findings in the larger scope of a hypothetical search platform that has one million users per day—where 1 in 10 becames distracted when performing one of their search tasks—this would amount to the population of users losing approximately $200,000$ minutes (approx. 140 hours per day in total). It would also lead to users missing/forgetting approximately $50,000 - 100,000$ examples that were previously identified.

However, a larger scale study is required to more deeply investigate whether these findings hold—and confirm whether they are practically and statistically significant. Interestingly, participants that were more prone to distraction investigated and considered more results and documents during their search, which does suggest that this lowered focus to stay on task gave way to more exploratory searching. Conversely, those who were more focused and less prone to distraction appeared to be more exploitative when searching. Since we found that perceptual speed and cognitive failure were largely independent, our results also suggest that we need to consider people's *propensity to be distracted* (and other cognitive failures, such as forgetfulness and how easily interrupted they are) in future IIR studies regarding individual cognitive differences.

To summarise, this work shows that additional visual elements—whether intended to be adversarial or not—can have serious ramifications in terms of increased demand on people's attention, higher workloads, and longer task times. This all comes at the cost of reducing the efficiency and efficacy of people's post-search task outcomes (in terms of concept recall, or *knowledge gain*), experience (in terms of satisfaction and workload), and time. Our findings suggest that cleaner, more minimalist interfaces/sites that are devoid of distracting clutter and instead help to preserve people's attention should be prioritised, *if the goal is to optimise for the user's search experience and post-search task outcomes*. Our work also motivates further research into examining the influence of other types of elements (entity cards, suggestions, explanations, etc.) which on the surface may appear appealing, but may also act as a distraction. Investigating adversarial elements which intentionally try to subvert people's filtering abilities by appearing to be relevant should also be examined. Further studies are also required to investigate to a greater degree the longer term and larger scale impacts of distractors when information seeking.

REFERENCES

[1] Phillip L Ackerman and Anna T Cianciolo. 2000. Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied* 6, 4 (2000), 259.

[2] Azzah Al-Maskari and Mark Sanderson. 2011. The Effect of User Characteristics on Search Effectiveness in Information Retrieval. *Information Processing & Management* 47, 5 (Sept. 2011), 719–729. https://doi.org/10.1016/j.ipm.2011.03.002

[3] Afrah Olayan Alanazi, Mark Sanderson, Zhifeng Bao, and Jaewon Kim. 2020. The Impact of Ad Quality and Position on Mobile SERPs. In *Proc. 5th ACM CHIIR*. 318–322. https://doi.org/10.1145/3343413.3377990

[4] Bryce Allen. 1994. Cognitive Abilities and Information System Usability. *IP&M* 30, 2 (1994), 177–191.

[5] Bryce Allen. 1994. Perceptual Speed, Learning and Information Retrieval Performance. In *SIGIR '94*, Bruce W. Croft and C. J. van Rijsbergen (Eds.). Springer London, London, 71–80. https://doi.org/10.1007/978-1-4471-2099-5_8

[6] Jaime Arguello and Rob Capra. 2016. The Effects of Aggregated Search Coherence on Search Behavior. *ACM Trans. Inf. Syst.* 35, 1, Article 2 (2016), 30 pages.

[7] Jaime Arguello and Bogeum Choi. 2019. The Effects of Working Memory, Perceptual Speed, and Inhibition in Aggregated Search. *ACM Transactions on Information Systems* 37, 3 (July 2019), 1–34. https://doi.org/10.1145/3322128

[8] Jaime Arguello, Wan-Ching Wu, Diane Kelly, and Ashlee Edwards. 2012. Task Complexity, Vertical Display and User Interaction in Aggregated Search. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 435–444. https://doi.org/10.1145/2348283.2348343

[9] Leif Azzopardi, Paul Thomas, and Nick Craswell. 2018. Measuring the Utility of Search Engine Result Pages: An Information Foraging Based Measure. In *Proc. 41st ACM SIGIR*. 605–614. https://doi.org/10.1145/3209978.3210027

[10] Becker-Olsen, Karen L. 2003. And Now, a Word from Our Sponsor: A Look at the Effects of Sponsored Content and Banner Advertising. *Journal of Advertising* 32, 2 (2003), 17–32.

[11] Iva Beleslin, Biljana R. Njegovan, and Maja S Vukadinović. 2017. Clickbait Titles: Risky Formula for Attracting Readers and Advertisers. In *Proc. XVII Int'l. Sci. Conf. Industrial Systems*. 364–369.

[12] Gerd Berget and Andrew MacFarlane. 2020. What Is Known About the Impact of Impairments on Information Seeking and Searching? *Journal of the Association for Information Science and Technology* 71, 5 (May 2020), 596–611. https://doi.org/10.1002/asi.24256

[13] Benjamin Boeuf, François. A. Carrillat, and Alian D'Astous. 2018. Interference Effects in Competitive Sponsorship Clutter. *Psychology & Marketing* 35, 12 (2018), 968–979.

[14] Pia Borlund and Peter Ingwersen. 1997. The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *J. of Documentation* 53 (1997). Issue 3.

[15] Horatiu Bota, Ke Zhou, and Joemon M. Jose. 2016. Playing Your Cards Right: The Effect of Entity Cards on Search Behaviour and Workload. In *Proc. 1st ACM CHIIR*. 131–140.

[16] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[17] Kathy Brennan, Diane Kelly, and Jaime Arguello. 2014. The Effect of Cognitive Abilities on Information Search for Tasks of Varying Levels of Complexity. In *Proceedings of the 5th Information Interaction in Context Symposium*. ACM, Regensburg Germany, 165–174. https://doi.org/10.1145/2637002.2637022

[18] Donald E. Broadbent. 1957. A mechanical model for human attention and immediate memory. *Psychological review* 64, 3 (1957), 205.

[19] Donald E. Broadbent, P. F. Cooper, P. FitzGerald, and Katharine R. Parkes. 1982. The Cognitive Failures Questionnaire (CFQ) and Its Correlates. *British Journal of Clinical Psychology* 21, 1 (1982), 1–16. https://doi.org/10.1111/j.2044-8260.1982.tb01421.x arXiv:https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8260.1982.tb01421.x

[20] Moira Burke, Anthony Hornof, Erik Nilsen, and Nicholas Gorman. 2005. High-cost banner blindness: Ads increase perceived workload, hinder visual search, and are forgotten. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 4 (2005), 423–445.

[21] Georg Buscher, Susan Dumais, and Edward Cutrell. 2010. The Good, the Bad, and the Random: An Eye-Tracking Study of Ad Quality in Web Search. In *Proc. 33rd ACM SIGIR*. 42–49.

[22] Bogeum Choi, Robert Capra, and Jaime Arguello. 2019. The Effects of Working Memory during Search Tasks of Varying Complexity. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM, Glasgow Scotland UK, 261–265. https://doi.org/10.1145/3295750.3298948

[23] Jacob Cohen. 1973. Eta-Squared and Partial Eta-Squared in Fixed Factor ANOVA Designs. *Educational and psychological measurement* 33, 1 (1973), 107–112.

[24] Lynne Cole, Andrew MacFarlane, and Stephann Makri. 2020. More than Words: The Impact of Memory on How Undergraduates with Dyslexia Interact with Information. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. ACM, Vancouver BC Canada, 353–357. https://doi.org/10.1145/3343413.3378005

[25] Henriette Cramer. 2015. Effects of Ad Quality & Content-Relevance on Perceived Content Quality. In *Proc. 33rd ACM CHI*. 2231–2234.

[26] Thomas H. Davenport and John C. Beck. 2001. The Attention Economy. *Ubiquity* 2001, May, Article 6 (May 2001). https://doi.org/10.1145/375348.376626

[27] Jon Driver and Gordon C. Baylis. 1989. Movement and visual attention: the spotlight metaphor breaks down. *Journal of Experimental Psychology: Human perception and performance* 15, 3 (1989), 448.

[28] Susan Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual Differences in Gaze Patterns for Web Search. In *Proc. 3rd IIiX*. 185–194.

[29] Ruth B. Ekstrom, John W. French, Harry H. Harman, and Diran Dermen. 1976. *Manual for Kit of Factor-Referenced Cognitive Tests* (education ed.). Princeton N.J:.

[30] Susan T Fiske and Shelley E Taylor. 2013. *Social Cognition: From Brains to Culture*. Sage.

[31] Olivia Foulds, Leif Azzopardi, and Martin Halvey. 2020. Reflecting upon Perceptual Speed Tests in Information Retrieval: Limitations, Challenges, and Recommendations. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. ACM, Vancouver BC Canada, 234–242. https://doi.org/10.1145/3343413.3377982

[32] Olivia Foulds, Leif Azzopardi, and Martin Halvey. 2021. Investigating the Influence of Ads on User Search Performance, Behaviour, and Experience during Information Seeking. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. ACM, Canberra ACT Australia, 107–117. https://doi.org/10.1145/3406522.3446024

[33] Gary L. Geissler, George M. Zinkhan, and Richard T. Watson. 2006. The Influence of Home Page Complexity on Consumer Attention, Attitudes, and Purchase Intent. *Journal of Advertising* 35, 2 (May 2006), 69–80. https://doi.org/10.1080/00913367.2006.10639232

[34] Richard Glassey and Leif Azzopardi. 2011. Finding interest in the stream. *Proceedings of the American Society for Information Science and Technology* 48 (01 2011). https://doi.org/10.1002/meet.2011.14504801297

[35] Daniel G. Goldstein, Siddharth Suri, R. Preston McAfee, Matthew Ekstrand-Abueg, and Fernando Diaz. 2014. The Economic and Cognitive Costs of Annoying Display Advertisements. *Journal of Marketing Research* 51, 6 (Dec. 2014), 742–752. https://doi.org/10.1509/jmr.13.0439

[36] Louisa Ha. 2008. Online Advertising Research in Advertising Journals: A Review. *Journal of Current Issues & Research in Advertising* 30, 1 (2008), 31–48.

[37] Bianca Harms, Tammo H.A. Bijmolt, and Janny C. Hoekstra. 2019. You Don't Fool Me! Consumer Perceptions of Digital Native Advertising and Banner Advertising. *Journal of Media Business Studies* 16, 4 (2019), 275–294.

[38] Sandra G Hart. 2006. NASA-task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[39] Yu-Chen Hsieh and Kuo-Hsiang Chen. 2011. How Different Information Types Affect Viewer's Attention on Internet Advertising. *Comput. Hum. Behav.* 27, 2 (mar 2011), 935–945. https://doi.org/10.1016/j.chb.2010.11.019

[40] Yi-Ting Huang. 2018. The Female Gaze: Content Composition and Slot Position in Personalized Banner Ads, and How They Influence Visual Attention in Online Shoppers. In *Computers in Human Behavior*, Vol. 82. 1–15.

[41] Hyunjoo Im, Hae Won Ju, and Kim K.P. Johnson. 2021. Beyond Visual Clutter: The Interplay among Products, Advertisements, and the Overall Webpage. *Journal of Research in Interactive Marketing* 15, 4 (Oct. 2021), 804–821. https://doi.org/10.1108/JRIM-10-2020-0213

[42] Jaroslaw Jankowski, Juho Hamari, and Jaroslaw Wątróbski. 2019. A Gradual Approach for Maximising User Conversion without Compromising Experience with High Visual Intensity Website Elements. *Internet Research* 29, 1 (2019), 194–217.

[43] Daniel Kahneman. 1973. *Attention and Effort*. Prentice-Hall, Englewood Cliffs, N.J.

[44] Diane Kelly and Leif Azzopardi. 2015. How Many Results per Page? A Study of SERP Size, Search Behavior and User Experience. In *Proc. 5th ACM ICTIR*. 183–192.

[45] Minchul Kim. 2013. *The Effects of Online Advertisements and News Images on News Reception*. Ph.D. Dissertation. The University of Wisconsin - Milwaukee.

[46] Nam Young Kim and S. Shyam Sundar. 2010. Relevance to the Rescue: Can "Smart Ads" Reduce Negative Response to Online Ad Clutter? *Journalism and Mass Communication Quarterly* 87, 2 (2010), 346–362.

[47] Youngseek Kim. 2009. How Does Web Advertising Affect Users' Information Seeking, Website Evaluation, and Source Evaluation?. In *Proc. iConference*. 2.

[48] Annie Lang. 2000. The Limited Capacity Model of Mediated Message Processing. *Journal of communication* 50, 1 (2000), 46–70.

[49] Dirk Lewandowski. 2017. Users' Understanding of Search Engine Advertisements. *Journal of Information Science Theory and Practice* 5, 4 (2017), 6–25.

[50] James R. Lewis. 1995. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *Int. J. Hum.-Comput. Interact.* 7, 1 (Jan. 1995), 57–78. https://doi.org/10.1080/10447319509526110

[51] Yujie Li. 2019. User Perception Affects Search Engine Advertising Avoidance: Moderating Role of User Characteristics. *Social Behavior and Personality: an international journal* 47, 4 (2019), 1–12.

[52] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[53] Antti Oulasvirta, Janne P. Hukkinen, and Barry Schwartz. 2009. When More Is Less: The Paradox of Choice in Search Engine Use. In *Proc. 32nd ACM SIGIR*. 516–523.

[54] Dolors Palau-Sampio. 2016. Reference Press Metamorphosis in the Digital Context: Clickbait and Tabloid Strategies in Elpais.Com. *Communication and Society* 29, 2 (2016), 63–71.

[55] K. Pawlik. 1966. Concepts and Calculations in Human Cognitive Abilities. In *Handbook of Multivariate Experimental Psychology*. Rand McNally.

[56] Marc Resnick and William Albert. 2014. The Impact of Advertising Location and User Task on the Emergence of Banner Ad Blindness: An Eye Tracking Study. *International Journal of Human-Computer Interaction* 30, 3 (2014), 206–219.

[57] Kate Scott. 2021. You Won't Believe What's in This Paper! Clickbait, Relevance and the Curiosity Gap. *Journal of Pragmatics* 175 (April 2021), 53–66. https://doi.org/10.1016/j.pragma.2020.12.023

[58] Elizabeth A. Styles. 2005. *The Psychology of Attention*. Taylor & Francis e-Library, London.

[59] Anne Treisman. 1991. Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology: Human Perception and Performance* 17, 3 (1991), 652.

[60] Lauren Turpin, Diane Kelly, and Jaime Arguello. 2016. To Blend or Not to Blend?: Perceptual Speed, Visual Memory and Aggregated Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Pisa Italy, 1021–1024. https://doi.org/10.1145/2911451.2914739

[61] Oliver Wilhelm, Michael Witthöft, and Stefan Schipolowski. 2010. Self-Reported Cognitive Failures: Competing Measurement Models and Self-Report Correlates. *Journal of Individual Differences* 31, 1 (Jan. 2010), 1–14. https://doi.org/10.1027/1614-0001/a000001

[62] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. 2021. What Makes a "Bad" Ad? User Perceptions of Problematic Online Advertising. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–24. https://doi.org/10.1145/3411764.3445459

[63] Ping Zhang. 2000. The Effects of Animation on Information Seeking Performance on the World Wide Web: Securing Attention or Interfering with Primary Tasks? *Journal of the Association for Information Systems* 1, 1 (2000), 30.