Delft University of Technology

Users and Contemporary SERPs

A (Re-)Investigation: Examining User Interactions and Experiences

Roy, N.; Maxwell, D.M.; Hauff, C.

DOI
[10.1145/3477495.3531719](https://doi.org/10.1145/3477495.3531719)

Publication date
2022

Document Version
Final published version

Published in
SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Users and Contemporary SERPs: A (Re-)Investigation
Examining User Interactions and Experiences

Nirmal Roy
Delft University of Technology
Delft, The Netherlands
n.roy@tudelft.nl

David Maxwell
Delft University of Technology
Delft, The Netherlands
d.m.maxwell@tudelft.nl

Claudia Hauff
Delft University of Technology
Delft, The Netherlands
c.hauff@tudelft.nl

ABSTRACT

The *Search Engine Results Page (SERP)* has evolved significantly over the last two decades, moving away from the simple *ten blue links* paradigm to considerably more complex presentations that contain results from multiple verticals and granularities of textual information. Prior works have investigated how user interactions on the SERP are influenced by the presence or absence of *heterogeneous* content (e.g., images, videos, or news content), the layout of the SERP (*list* vs. *grid* layout), and *task complexity*. In this paper, we reproduce the user studies conducted in prior works—specifically those of Arguello et al. [4] and Siu and Chaparro [29]—to explore to what extent the findings from research conducted five to ten years ago still hold today as the average web user has become accustomed to SERPs with ever-increasing presentational complexity. To this end, we designed and ran a user study with four different SERP interfaces: *(i) a heterogeneous grid*; *(ii) a heterogeneous list*; *(iii) a simple grid*; and *(iv) a simple list*. We collected the interactions of 41 study participants over 12 search tasks for our analyses. We observed that SERP types and task complexity affect user interactions with search results. We also find evidence to support most (6 out of 8) observations from [4, 29] indicating that user interactions with different interfaces and to solve tasks of different complexity have remained mostly similar over time.

CCS CONCEPTS

• **Information systems** → **Search interfaces**; • **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

Human Computer Interaction; Interactive Information Retrieval; Search Interfaces; Search Tasks; Reproducibility

1 INTRODUCTION

The *Search Engine Results Page (SERP)* has evolved significantly over the last two decades, moving away from the *ten blue links* paradigm, to considerably more complex presentations that contain results from multiple verticals and multiple granularities of textual information (snippets, direct answers, entity cards, etc.)—all interleaved within one page. The incorporation of heterogeneous content in a SERP has been shown to change how users interact with web results [2, 3, 7, 10, 17, 21, 27, 36]. How (and where) content is displayed in a SERP affects user interactions as well [4, 28, 30]. While contemporary web SERPs maintain the original idea of a *list* of items that are ranked in decreasing order of relevance, alternative presentations such as a *grid* layout—as also recently (again) popularised by *You.com*—have also been explored [14, 20, 29, 37].

In addition, past research [4, 29, 30, 36] has shown that user behaviour on the SERP does not only depend on the *presentation* of information, but also on the *search task* at hand. For a navigational task such as *'find and access the homepage of SIGIR 2022'*, a user—in the ideal case—requires a single query and a single click. Contrast this to an informational task, such as *'good restaurants near the venue of SIGIR 2022'*. This requires the scanning of multiple results, and likely results in further query reformulations to learn more about specific suggestions.

As commercial web search engine SERPs have evolved over time (and thus end users have become accustomed to different types of SERPs), we explore in this paper to what extent user study findings from $5-10$ years ago still hold today. Specifically, we focus our attention on reproducing the experimental setup of two prior studies: Arguello et al. [4] (published in 2012) as well as Siu and Chaparro [29] (published in 2014)—these both investigated how user interactions on the SERP are influenced by the presence or absence of heterogeneous content, the layout of the SERP (list vs. grid), and task complexity. Inspired by the two papers we reproduce, our study is guided by the following research questions.

RQ1 How does a user's interactions with a SERP differ when results are presented in a list and grid layout?

RQ2 How does task complexity affect user interactions with a SERP?

RQ3 What is the interplay between task complexity and SERP layout on user interactions?

RQ4 How do users perceive the different SERP layouts?

To this end, we conducted a user study with $n = 41$ participants that each were given 12 search tasks of varying complexity (ranging from search tasks of type *Remember* to *Analyse*) to solve with one of four different SERP interfaces: *(i) heterogeneous grid*; *(ii) heterogeneous list*; *(iii) simple grid*; and *(iv) simple list*. We explore whether the following eight observations from [4, 29] about users and their

interactions with list vs. grid layouts—and heterogeneous vs. simple results—across different task complexities hold today.

O1 Users fixated significantly more on the grid layout SERP compared to the list layout SERP for completing more complex tasks [29].

O2 On the grid layout SERP, users fixated on search results significantly more for completing more complex tasks compared to simple tasks. A similar observation was found for the list layout SERP [29].

O3 On the list layout SERP, users fixated significantly longer for completing more complex tasks compared to simple tasks. For the grid SERP, there were no significant differences in fixation duration between varying task complexities [29].

O4 In the list layout SERP, more complex tasks required significantly greater levels of search interaction: longer search sessions, more clicks on SERP, and more web pages visited [4].

O5 In a SERP where web results are arranged in a list layout, users clicked on significantly more vertical results when they were present on the main page of the SERP (blended, heterogeneous display) compared to when they were only present as tabs (non-blended, simple display) [4].

O6 Task complexity did not have a significant effect on user interaction with vertical results in the list layout SERP [4].

O7 The interplay between task complexity and display of verticals (blended, heterogeneous display vs. non-blended, simple display) did not have a significant effect on user interaction with vertical results in the list layout SERP [4].

O8 Neither study [4, 29] found significant differences in user evaluation of the different SERP types, list vs grid layout for the former and blended vs non-blended display for the latter, in their experiments.

In our user study, we observed that SERP types and task complexity affect user interactions with search results. We also find evidence to support most—6 out of 8—observations from [4, 29].

2 RELATED WORK

2.1 Task Complexity and User Interactions

A number of works have focused on the effect of task types on user interactions on SERPs. Buscher et al. [8] performed a large-scale analysis using query logs to understand how individual and task differences might affect search behaviour. Their findings show that there are cohorts of users who examine search results in a similar way. They also showed that the type of task has a pronounced impact on how users engage with the SERP. Arguello et al. [4] observed that the more complex the task, the more users would interact with various components on the SERP whereas Thomas et al. [32] found that users tended to examine the result list deeper and more quickly when facing complex tasks. Jiang et al. [12] compared user interactions in relatively long search sessions (10 minutes; about 5 queries) for search tasks of four different types. Wu et al. [36] also observed differences in user interactions with the SERP based on whether they had to look for answers to a factoid question or a non-factoid question. In these studies, the SERPs were composed of web search results in the *de facto* list format.

2.2 SERP Presentation and User Interactions

Sushmita et al. [30] observed that positioning (top, middle, bottom) of different verticals on a SERP affects clickthrough rates of users when the verticals (news, image and video) were presented in a blended manner with the web search results. Arguello et al. [4] also looked into how task complexity affects user interactions and usage of aggregated vertical results when they are interleaved with web results, versus when they are presented as tabs. On a similar note, they observed that for more complex tasks, users clicked on more vertical results when they were interleaved with web search results. Bota et al. [7] conducted a crowdsourced online user study to investigate the effects of entity cards given ambiguous search topics. They found that the presence of entity cards has a strong effect on both the way users interact with search results and their perceived task workload. Furthermore, Levi et al. [16] performed a comprehensive analysis of the presentation of results from seven different verticals (including a community question answering vertical) based on the logs of a commercial web search engine. They observed that the community question answering vertical receives on average the highest number of clicks compared to other verticals. Wu et al. [36] studied how the presence of answer modules on SERPs affect user behaviour and whether that varies with question types (factoid vs. non-factoid). They found that the answer module helps users complete search tasks more quickly, and reduces user effort. In the presence of answer modules, users' clicks on web search results were significantly reduced while answering factoid questions. Shao et al. [28] conducted a user study to understand how user interaction is affected by the presence of results in the right rail of a heterogeneous SERP in addition to the traditional web results in the left-rail. They found that users have more interactions with the SERPs, appear to struggle more, and feel less satisfied if they examine the right-rail results. Overall, findings observed that the presence of verticals and other heterogeneous modalities of results and their position on the SERP affect user interactions. In these studies, results were also presented in the *de facto* list format.

Kammerer and Gerjets [14] observed that when web search results are presented in a grid layout, the impact of search result positioning on selecting trustworthy sources is drastically reduced in comparison to the more traditional list approach. Users typically follow a top-down approach when scanning lists, and are more susceptible to select untrustworthy sources if they appear high up in the list. This effect is reduced for a grid-based presentation. However, the authors do not compare different types of tasks, nor do they explore user behaviours when results from various vertical features of the search engines are present on the SERP. Siu and Chaparro [29] compared the eye-tracking data of grid and list SERP layouts with two types of tasks (informational vs. navigational), and investigated potential differences in gaze patterns. The *'F-shaped'* pattern was less prominent on the grid in comparison to the list layout. These two studies explore how user interactions change when web results are presented in a grid vs. a list layout. They do not, however, include vertical results in their study.

3 METHODOLOGY

To address our four overarching research questions as outlined in §1, we conducted a user study with $n = 41$ participants. Each

participant was assigned to one of four experimental search interface conditions (**interfaces:** *between-subjects*), and completed 12 search tasks (**tasks:** *within-subjects*).

Our four experimental search interface conditions considered the **layout type** (*list-* vs. *grid-*based layout) and the **verticals present** on the SERP (*heterogeneous content* vs. *homogeneous content*). These combinations result in the interface conditions outlined below, with examples of the two layout types presented in Figure 1, with further details provided in §3.1.

SL Simple List Considered as our baseline interface condition (the standard and widely used *ten blue links* [11]), this interface presents results in a list, with each result presented one under the other. All results are *web results*, and as such are *homogeneous* in terms of presented content.

SG Simple Grid The same homogeneous approach to content is taken as for **SL**, but with results presented in a *grid-based* approach. Instead of scrolling along the vertical, participants subjected to this interface scroll along the *horizontal*.

HL Heterogeneous List Similar in approach to **SL**, **HL** presents results in a list. However, different verticals are mixed in with the standard web results. Beyond web results, heterogeneous content used in this study includes *image* and *video* results.

HG Heterogeneous Grid Similar to **HL** but now the content is displayed in grid form, with *web-based* results appearing in a grid, before additional image and video content.

3.1 Search Interface Design and System

Given the above, our goal is to find out to what extent the observations from prior studies by both Arguello et al. [4] and Siu and Chaparro [29] are valid after almost a decade of SERP design evolutions (and additions). To operationalise our four experimental interface conditions, we first needed to create a SERP template design that closely mirrors the design of a contemporary web search engine.

For this study, we selected the *Google* SERP as it presents information recognisably, and commands approximately 92% market share.[1] A replica template was created with particular attention paid to the colour schemes, fonts, width and height of components—as well as the spacing between them. The end result was a highly realistic template of a contemporary SERP, on which we based all of our study's results pages.[2]

SERP Template Overview Figure 1 presents the SERP template used, presenting results for the query 'how do dams generate electricity?'. Present are examples for both list- *(Figure 1(a))* and grid-based *(Figure 1(b))* interfaces.

A query box is provided ❶. However, this is disabled and provides no functionality for this study. It does however display the query terms that were used to derive the presented results *a priori* (see §3.2). This is presented next to the *information need* for a given task ❷, alongside which there is a button to take the participant to the next stage of the experiment. The SERP template also provides links to additional results pages, namely 🖾 *Images* and 🎥 *Videos*

❸, emulating the setup of the study by Arguello et al. [4]. As shown at ❹, a grid-based layout is shown for both image and video pages, as is the norm in commercial web search engines such as Google and *Bing*. For images, a total of 16 were displayed (in a 4×4 grid); for videos, a total of nine were shown (in a 3×3 grid).

On the SERP, standard web results are presented ❺, with the *de facto* 10 *results per page (RPP)* provided. For list-based interfaces ❺ (**SL** and **HL**), web results are displayed in the standard way, with one result following the other down the left rail of the SERP. Grid-based interfaces ❺ (**SG** and **HG**) present the results in a *carousel* user interface component (emulating the setup of Siu and Chaparro [29]), where the 10 results are arranged in the form of a 5×2 grid. A total of six (3×2) results were visible *above-the-(vertical)-fold*; access to the remaining four results (2×2) was made available through use of a button to scroll across.

As denoted by the red dashed boxes in Figure 1 ❻, heterogeneous content is also added to the SERP template. Present only in experimental search interface conditions **HL** and **HG**, these components provided inline image and video results to the participants. Like web results in the grid-based interfaces, these were also scrollable, mimicking the behaviour of contemporary web search engine SERPs. Sufficient content was placed within these components to ensure two complete scrolls could be completed; the number of images displayed varied as their widths were variable. On interface condition **SG**, image and video components were placed under the *third* web result; on interface condition **HG**, they were placed directly underneath the web results grid.

The SERP template was also fully interactive—participants could click on links of web results ❼, with a new browser tab then opening to present the page at the linked URL. In addition, images and videos within the SERP could also be interacted with. Clicking on an image ❽ took the participant to the webpage containing the image (again, in a new tab). Videos, all sourced from *YouTube*, could be played on the SERP itself ❾, with the necessary infrastructure in place to enable such functionality. If the participant wished to view the video on the YouTube website itself, they could click the link underneath ❿ to do so. Again, YouTube links opened a new tab in the participant's browser.

SERP Definition Note that for each query, there are three unique results pages to replicate the study of Arguello et al. [4]. These are the results *'landing'*, or 🔍 *All* page, containing the web results (and additional components, for interface conditions **HL** and **HG**)—as shown in Figure 1, as well as the 🖾 *Images* and 🎥 *Videos* pages. From hereon in, we refer to a SERP as the 'landing' page, containing web results. To replicate the study of Siu and Chaparro [29], the 'landing' page itself is sufficient.

Capturing Interactions and Experiences Integrated with the SERP template was **Log**UI [19], a framework-agnostic JavaScript library for capturing different interactions and other events within a web-based environment. **Log**UI was configured to capture a series of mouse events (including hovers, clicks, and scrolls) over the various components of the SERP. Interactions on components included (but were not limited to) web results, images and video contents (including the capturing of the playback, pausing, and completion of YouTube videos). We also recorded interactions to (and on) the supplementary 🖾 *Images* and 🎥 *Videos* pages. Browser-wide events

---

[1]https://gs.statcounter.com/search-engine-market-share. All URLs in this paper were last checked on 2022-02-14.
[2]Templates are released for future user studies, available at https://github.com/roynirmal/sigir2022-serp-reproducibility.
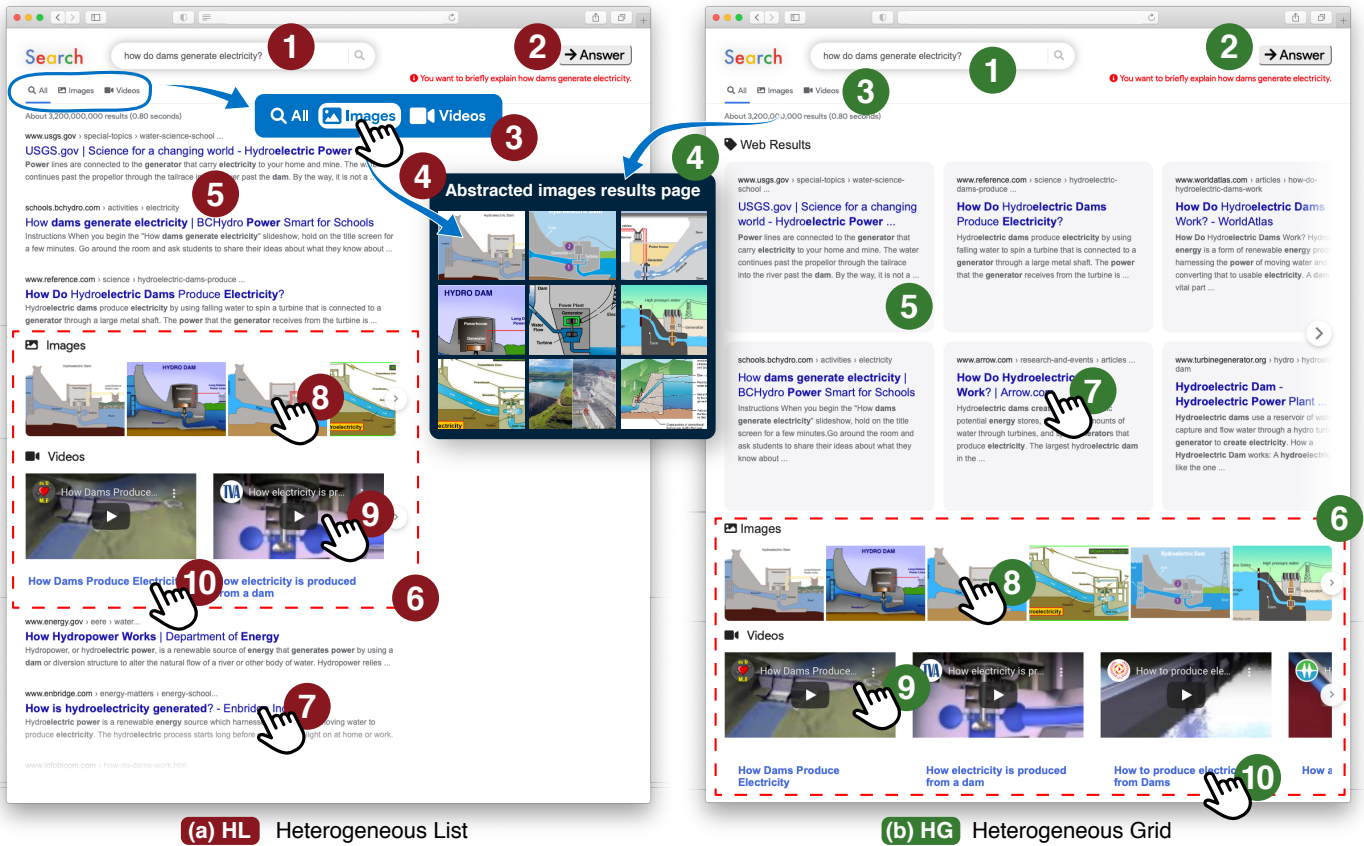
Figure 1: Examples of both the *(a)* list-based and *(b)* grid-based interfaces trialled. Note the inclusion of links for the separate 🔍 *All* (as shown), 🖼 *Images,* and 🎥 *Videos* result pages. Heterogeneous content is displayed in red boxes, and is *not* present in the two homogeneous content interface conditions (SG and SL). Circled numbers correspond to the narrative of Section 3.1.

were also captured, and included the ability for us to compute the time spent *away* from the SERP—when participants would click on a document/image/video link, which would open a new tab.

Experience data was captured via a number of *Qualtrics*[3] surveys; a pre- and post-experiment survey were completed by each participant, in addition to the small post-task summary the participants had to write. More details on these questions and the flow of the experiment can be found in §3.4. Our setup ensured that participants would jump between the Qualtrics surveys and SERPs as and when required.

Static SERPs As alluded to with the disabled query box ❶, our experimental setup featured no programmable backend or search functionality. This meant that there was *no* additional querying functionality. We served manually curated SERPs that we produced *a priori* for each of the 12 search tasks we asked participants to undertake. This setup ensured that all study participants viewed the same results (a setup also chosen in prior studies, such as those by Sushmita et al. [30] and Wu et al. [36]). While making the search experience somewhat less realistic, it did provide us with the benefit of not having to deal with participants submitting diverse queries.

The design also removed a confounding variable and allowed us to address our four RQs by calculating the user interaction measures on a fixed set of web results, images, and videos.

3.2 Search Tasks

For this study, we used four different types of information need: a **Navigational** type *(where individuals seek to find particular websites)*, and three different *informational* categories, belonging to the **Remember** *(involving the retrieval, recognition, and recalling of relevant knowledge)*, **Understand** *(constructing meaning from information sources)*, and **Analyse** *(involving the breakdown of information info constituent parts, and determining how they related to one another)* categories. For each category, we produced three unique information needs. This led to a total of 12 information needs which are listed in Table 1. Particular attention was paid to designing tasks that enticed participants to not only look at web results but also at image and video search results as well.

The choice of our information needs is based on the study designs used by both Arguello et al. [4] and Siu and Chaparro [29]. More specifically, Arguello et al. [4] designed a series of tasks that required different levels of diversity of *information* to complete—as well as different amounts of search effort. Tasks were grounded in

---

[3]https://www.qualtrics.com/

the revised *Bloom taxonomy*, as outlined by Anderson and Krathwohl [1].[4] Search tasks were informational and belong to the *Remember*, *Understand*, and *Analyse* categories. In addition, Siu and Chaparro [29] employed just two categories of tasks for their study—navigational and informational. Our design thus combines the setups of both prior works that we wish to examine.

3.3 Query Selection and SERP Curation

To ensure that participants received helpful search results, we required a common search query for each. To this end, we ran a small crowdsourced pilot study on the *Prolific* platform[5]. This pilot had $n = 25$ workers, with the design largely inspired by the study reported by Bailey et al. [6]. Workers took approximately 10 minutes to complete the task and were paid at the hourly rate of GBP8.00 for their time. All 12 information needs were presented to the workers. They were instructed to type the query terms that they would issue to their web search engine of choice if they were seeking information to address the information need. Collected queries were then normalised (case normalisation, stripped punctuation, whitespace cleanup) and passed through the *Bing Spell Check API* to generate a final canonical form of each submitted query. Subsequently, we determined the most frequently occurring query variation for the 12 tasks, taking this query forward as the one to use for the next stage of our study. These are listed in the parentheses in Table 1.[6]

Curating SERPs We then used a combination of the *Bing Web Search API*, *Bing Image Search API*, and *Bing Video Search API* to curate a collection of: *web* (title, snippet text, and target URL); *image* (source image and document URL); and *video* (video source URL) results for each of the 12 queries. Snippet text was truncated to the equivalent of two sentences/lines, as this has been previously shown to be a good trade-off in terms of providing a sufficient *information scent* and encouraging interaction (i.e., clicks) [18]. Video links were filtered to YouTube only, as utilising only one video content provider reduced complexity for playback on our SERPs. Any URLs that proved non-functional or redirected to a 404 page were also removed. The content was then placed on our SERP templates, allowing us to construct SERPs, an image results page, and a video results page for each query. SERP variations for all four search interface conditions were produced.

3.4 Experimental Procedure

The 12 search tasks undertaken by each participant were preceded and followed by pre- and post-experiment surveys. We first performed screen and browser viewport resolution checks, requiring that all participants use a maximised browser window with a resolution of 1920×1080 or greater. This ensured that we could guarantee the SERPs displayed to the participants could be viewed without scrollbars along the horizontal. If the checks were successful, participants began the experiment by providing basic demographic information and were also asked minor questions on their search engine usage, specifically on what components on a contemporary

SERP they often make use of. In addition, we asked what their preferred search engine is. They were then randomly assigned to one of the four search interface conditions (**SL**, **SG**, **HL**, or **HG**).

Participants were primed to summarise their findings after each search task (in no more than 50 words). Upon acceptance of this instruction, the first search task began, with a SERP similar to the one presented in Figure 1*(a)* or Figure 1*(b)*. With the selected query ❶ and information need ❷ present, participants then began to examine the content. Participants were *not* given a minimum or maximum amount of time to search. We reiterate that they were also *not* given the opportunity to issue their queries. Once they were satisfied with what they had found, they clicked the ➜Answer button at the top of the SERP ❷, and entered their summary. Once complete, the next task began. This process was repeated for the remaining 11 tasks which were displayed to them in random order. Other researchers have also employed randomisation for condition allocation to minimise topic ordering effects [15, 36].

After the search tasks had been completed, participants then moved on to the post-experiment survey. We used the sub-scales from O'Brien's *Engagement Scale* [22, 23] as was done by Arguello et al. [4]. These are aimed at eliciting their evaluation of the interface they used on the following aspects of engagement: *focused attention*; *perceived usability*; *experience*; *aesthetics*; and *felt involvement*. The engagement scale was originally designed to evaluate shopping websites, and hence we modified/removed the statements pertaining to shopping to suit our needs. For example, we changed the original statement (belonging to *aesthetics* sub-scale) *"This shopping website was aesthetically appealing"* to *"The layout of the results page is aesthetically appealing"*. For all statements in the sub-scales, participants indicated their level of agreement (1=*strongly agree*; 5=*strongly agree*). We also used the *search effectiveness* sub-scale used by Arguello et al. [4] to evaluate how effective the interfaces were in helping participants find information. In total, we used 26 statements from the six sub-scales to elicit user evaluation of the search interfaces. The reliability scores (Cronbach's Alpha) for the sub-scales are reported in Table 3. They were also asked to rate the perceived usefulness of web, image and video results.

3.5 Study Participants

Like our pilot, we recruited participants from the Prolific platform. Our $n = 41$ participants were native English speakers from the United Kingdom, with a 95% approval rate on the platform, and had a minimum of 250 prior successful task submissions. From our participants, 32.5% identified as female, and 67.5% as male. The mean age of our participants was 36.5 ± 9.7, with a minimum age of 22 and a maximum of 68. 92% of participants listed *Google* as their preferred search engine, with the remaining 8% identified as *DuckDuckGo* users. With respect to the highest completed education level, 51.2% possess a Bachelors (or equivalent), 24.4% have a Masters (or equivalent), 19.5% have a high school degree, and 4.9% have an Associate (or equivalent). 95% of participants cited using web results on a contemporary SERP, 78% made use of image results, with 37% citing that they used video results.

In our random assignment, 11 participants were assigned to **HG**, with ten participants each assigned to **HL**, **SL**, and **SG**. The experiment lasted on average 40 minutes for the 41 participants. Like our pilot participants, they were compensated at the rate of

---

Table 1: Overview of information needs and their type. The rightmost column shows the most popular query obtained from our query selection pilot study, outlined in Section 3.3. Numbers in parentheses indicate how many crowdworkers ($n = 25$) submitted the most popular query variation. Here, *Nav.=Navigation, Remem.=Remember,* and *Underst.=Understand.*

Type	Information Need	Most Popular Query Variation
Nav.	You want to find the homepage of Andrew Zimmern, the chef.	`andrew zimmern chef` (14)
	You want to find the page of Air Jordan on the Nike website.	`nike air jordan` (14)
	You want to find the page displaying the Flixbus route map in Europe.	`flixbus europe route map` (6)
Remem.	You want to know where is the pituitary gland located in the body.	`where is the pituitary gland` (9)
	You want to find out what clothes the famous cartoon character Mickey Mouse typically wears.	`what clothes does mickey mouse wear` (5)
	You want to find out how to calculate the volume of an ellipsoid.	`ellipsoid volume formula` (4)
Underst.	You want to find out the steps required to make a paper airplane.	`how to make a paper airplane` (10)
	You want to briefly explain how dams generate electricity.	`how do dams generate electricity` (17)
	You want to find out how to prevent shower mirrors from fogging.	`stop shower mirror fogging` (3)
Analyse	You want to get into martial arts, but you have no fighting experience. Which form of martial arts is more suitable for beginners?	`best martial arts for beginners` (3)
	You want to find out the main things to look for while installing solar panels on the roof of a house.	`things to consider before installing rooftop solar panels` (1)
	You want to buy a new camera lens for taking professional pictures of your friend. Which camera lenses are best for portrait photography?	`best camera lenses for portrait photography` (3)

Table 2: Results of a factorial mixed ANOVA, where interface is between-subjects, and task is within-subjects variable. A ✓ indicates significant effect ($p < 0.05$) on the particular user interaction and ✗ indicates no significant effect.

User Interactions	SERP Main Effect	Task Main Effect	B/W SERP & Task
Web results clicks	✓ (F = 4.27, $p = 0.01$)	✓ (F = 4.18, $p = 0.01$)	✗
Mean web result reading time (s)	✗	✓ (F = 3.97, $p = 0.01$)	✗
Mean session duration (s)	✗	✓ (F = 12.72, $p < 0.0001$)	✗
Mean web result hover duration (s)	✗	✗	✗
Image clicks (SERP)	✗	✓ (F = 7.24, $p = 0.004$)	✗
Video clicks (SERP)	✗	✗	✗
Image hovers (SERP)	✗	✓ (F = 6.98, $p = 0.009$)	✗
Video hovers (SERP)	✗	✗	✗
Image clicks (image results page)	✗	✓ (F = 4.66, $p = 0.01$)	✗
Video clicks (video results page)	✗	✗	✗
Image hovers (image results page)	✗	✓ (F = 5.39, $p = 0.01$)	✗
Video hovers (video results page)	✓ (F = 3.36, $p = 0.02$)	✗	✗

GBP 8.00 per hour. All participants who registered completed the study; post-hoc checks confirmed that they had provided sensible answers for each task, and as such we approved all who took part for payment. As such, our base analyses are reported over $41 * 12 = 492$ search sessions and their corresponding interaction logs.

4 RESULTS AND DISCUSSION

For our analyses[7], we conduct a series of mixed factorial ANOVA tests to observe if task complexity, SERP types or the interplay between them have a significant effect on interactions. We follow up the ANOVA with post-hoc *t*-tests with Bonferroni correction ($p < 0.05$) to observe where significant differences occur. We evaluate if observations **O4**-**O7** also hold in grid SERPs, **HG** and **SG**.

4.1 RQ1: SERP Type and User Interactions

Table 2 presents results that are relevant to our first three research questions. Here, a ✓ indicates a significant effect ($p < 0.05$) on the particular user interaction, and a ✗ indicates no significant effect.

As seen in Tables 2 and 3, different SERP types do *not* have a significant effect on user interactions except for: *(i)* the number of web results clicked (row **I**, Table 3); and *(ii)* the number of hovers on videos present in the video results page (**XV**, Table 3). Post-hoc

tests reveal that participants in the **HL** condition have significantly more web result clicks than their **HG** and **SG** counterparts (**I**, Table 3). **HL** participants also have longer web result reading times compared to participants in any of the other SERP conditions (**II**, Table 3)—albeit not significant. Furthermore, participants with the list interfaces (**HL** and **SL**) have a greater number of hovers over web results compared to **HG** and **SG**. As a result, we cannot confirm **O1** where Siu and Chaparro [29] found significantly more fixation counts on the grid interface than on the list interface. We note that, since we did not record eye gaze data, we are approximating fixation counts by user interactions such as web result clicks and snippet text hovers, as mouse position has been shown to correlate with gaze positions in prior studies [21, 24, 25]. One of the possible reasons for the difference in observation with **O1** can be that our participants are more familiar with the standard list layout of web results, as a majority use Google as their main search engine.

Arguello et al. [4] do not compare user interactions with web results on heterogeneous SERPs vs. simple SERPs. However, we observe that participants using the simple SERP interfaces (**SG** and **SL**) scan web search results to lower depths than those of their heterogeneous interface counterparts (**IV**, Table 3). The lack of information (i.e., fewer verticals) on the SERP requires participants to scan web results to a greater depth in the ranked list.

---

[7]All data and code pertaining to our analyses are available here.

Based on Table 3 (**VII-X**), we find that on average **HL** participants interact more with image and video results that are present on the SERP compared to their **HG** counterparts. **SG** and **SL** participants interact more with vertical results present in the image and video results page than those of their heterogeneous counterparts (*XI-XIV*, Table 3). Post-hoc tests also reveal that **SL** participants have significantly more hovers on video results on the video results page than participants in the other SERP conditions (**XIV**, Table 3). The lack of vertical results on the SERP makes the participants interact with them in the respective vertical results pages which shows that our informational needs indeed require participants to seek out image and video search results as well. **HG** and **HL** participants seem to be satisfied with vertical results present on the SERP and the former barely interacted with vertical results present in the respective results pages (*XI-XIV*, Table 3). Looking at overall interactions with vertical results (adding interactions with vertical results present on the SERP and the vertical results pages for **HG** and **HL**), we see that **HG** and **HL** have slightly more interactions than **SG** and **SL** respectively. This difference is not significant, but we do see a trend in the line of **O5** where Arguello et al. [4] observed a higher number of vertical result clicks when they were blended with the web results in the SERP. This is compared to when they were only present on the respective vertical results page. On a side note, the higher interactions with vertical results present on the SERP by **HL** participants compared to **HG** participants (**VII-X**, Table 3, also depicted in Figure 2*(c)*) can be attributed to the fact that images and videos in **HL** SERPs appear in the middle of the web results (between rank 3 and 4) whereas they appear below the web results in **HG** SERPs. Participants in the latter interface expend comparatively more effort to access the vertical results, thereby reducing their interaction. We leave further analysis on the effect of positioning of vertical results on user interaction for future work.

Addressing **RQ1**, we found that the interface has a significant main effect on the clicks on web results and hovers on videos on the video results—page but not on other user interactions.

4.2 RQ2: Task Complexity and User Interactions

Table 4 shows that the informatioon needs of the *Analyse* type, which are the most complex among our information needs, warrant most web result clicks (**I**), web result dwell time (**II**) and session duration from participants (**III**). Table 2 shows that the main effect of task complexity on these interactions is significant. Participants reach greater web result click depth (**IV**, Table 4) for *Analyse* tasks, albeit not significant. Post-hoc tests reveal that *(i) Analyse* tasks receive significantly more web result clicks than *Remember* tasks; *(ii) Analyse* and *Understand* tasks lead to significantly higher web result dwell times than *Navigation* tasks; and *(iii)* the session duration for *Analyse* and *Understand* tasks are significantly higher than for *Navigational* tasks, while the session duration for *Analyse* tasks is also significantly greater than that for *Remember* tasks. Overall, we find that user interactions on web search results increase as the complexity of information needs increase which is inline with the observations of Arguello et al. [4] and we can partially confirm **O4**.

Arguello et al. [4] did not include *Navigational* tasks in their experiments. We argue that they can be considered as tasks requiring the lowest level of cognition, and as such follow the trend of **O4**—they receive the least interaction among all task categories.

The only exception to this was web result clicks—the nature of the task requires participants to click web result links to ascertain that they found the correct page.

We approximate fixation duration in **O3** by observing hover duration over the web results, akin to fixation count in §4.1. Although participants hover longer over web results (**V**, Table 4) and snippet text (**VI**) for *Remember* tasks compared to other tasks, the difference across tasks is not significant. Moreover, the mean hover duration on web results (snippet and title) for participants belonging to the grid SERP types (**HG** and **SG**) is longer than for those belonging to the list SERP types (**VI**, Table 3). As seen from Table 2, the interplay between SERP type and task complexity do not have a significant effect on hover duration over web results. As a result, we can only partially confirm **O3** where Siu and Chaparro [29] also do not find significant differences in fixation duration for grid layout for the tasks but they *did* find significantly longer fixation duration on the list layout for more complex tasks.

Among interactions with vertical results present on the SERP (**VII-X**, Table 4), we observe that *Remember* tasks receive the most interactions on average. Post-hoc tests reveal that: *(i)* participants click significantly more on images present on the SERP (**VIII**, Table 4) for *Remember* and *Navigational* tasks compared to *Analyse*; and *(ii)* they hover significantly more on images present on the SERP (**X**, Table 4) for *Remember* tasks compared to all other task categories. For images present on image results page, we again observe significantly more image clicks (**XII**, Table 4) and hovers (**XIV**, Table 4) for *Remember* tasks compared to the more complex *Understand* or *Analyse* tasks. Findings regarding user interactions with vertical results (present on the SERP and the vertical results pages) and their relationship with task complexity is contrary to the observations of Arguello et al. [4], and hence we cannot confirm **O6**. The high interaction with vertical results for *Remember* tasks together with the fact that participants hover over web results and snippet text longer (on average) for the same task (**V** & **VI**, Table 4) shows that participants prefer to address information needs of the *Remember* type by either hovering over web results and interacting with verticals rather than clicking the link. Arguello et al. [4] do not observe hover duration in their analysis.

To answer **RQ2**, we find that task complexity does have a significant effect on several user interactions. With participants interacting more with web results as tasks get more complex, we observe significantly more interactions with image results for *Remember* tasks compared to more complex *Analyse/Understand* tasks.

4.3 RQ3: Task Complexity, SERP Type and User Interactions

As seen in Table 2, we do not observe a significant effect of the interplay between SERP types and task complexity on user interaction with web results or verticals which is similar to what Arguello et al. [4] found. Hence, we can confirm **O7**.

From Figure 2*(a)*, we observe that participants across all SERP types click the most web results for *Analyse* tasks (in line with **O4**). For each task type, **HG** participants click the least number of web results and for most tasks participants with grid SERPs click lower ranked web results than those with list SERPs (in contrary to **O1**). Approximating fixation count by web result clicks, as done in §4.2, we see for each SERP type, the complex *Analyse* tasks receive

Table 3: User interactions for different interfaces across all tasks. † indicates that there is a significant main effect of SERP layout on that particular user interaction. $\mathcal{HG}$, $\mathcal{HL}$, $\mathcal{SG}$, $\mathcal{SL}$ indicate significant difference with HG, HL, SG and SL respectively. Maximum values for each interaction is highlighted in bold.

| Row	Interaction	Interface Condition			
		HG	HL	SG	SL
I	Web result clicks†	$11.27(\pm8.43)^{\mathcal{HL}}$	$\mathbf{21.30(\pm8.99)}^{\mathcal{HG,SG}}$	$18.20(\pm9.91)^{\mathcal{LH}}$	$18.70(\pm11.89)$
II	Mean web result reading time (s)	$17.96(\pm12.74)$	$\mathbf{27.00(\pm28.82)}$	$16.82(\pm9.05)$	$25.09(\pm10.64)$
III	Mean session duration (s)	$94.89(\pm56.43)$	$106.96(\pm46.85)$	$98.35(\pm47.52)$	$\mathbf{109.24(\pm64.55)}$
IV	Maximum web result click depth	$3.36(\pm2.60)$	$3.62(\pm1.45)$	$\mathbf{4.22(\pm1.89)}$	$4.15(\pm1.61)$
V	Total web result hovers	$66.55(\pm29.77)$	$83.40(\pm61.38)$	$122.70(\pm87.07)$	$\mathbf{124.40(\pm100.79)}$
VI	Mean web result hover duration (s)	$\mathbf{2.91(\pm6.95)}$	$2.49(\pm4.74)$	$2.20(\pm4.96)$	$0.80(\pm0.69)$
VII	Image clicks (SERP)	$0.82(\pm1.17)$	$\mathbf{2.10(\pm2.38)}$	-	-
VIII	Video clicks (SERP)	$1.55(\pm2.81)$	$\mathbf{11.20(\pm34.04)}$	-	-
IX	Image hovers (SERP)	$13.27(\pm14.88)$	$\mathbf{16.30(\pm16.73)}$	-	-
X	Video hovers (SERP)	$6.18(\pm9.66)$	$\mathbf{13.40(\pm22.62)}$	-	-
XI	Image clicks (image results page)	$0.00(\pm0.00)$	$0.40(\pm0.70)$	$0.70(\pm1.06)$	$\mathbf{1.10(\pm1.45)}$
XII	Video clicks (video results page)	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$\mathbf{0.70(\pm1.49)}$
XIII	Image hovers (image results page)	$0.00(\pm0.00)$	$4.10(\pm10.67)$	$8.80(\pm18.58)$	$\mathbf{13.00(\pm19.96)}$
XIV	Video hovers (video results page)†	$0.00(\pm0.00)^{\mathcal{SL}}$	$0.00(\pm0.00)^{\mathcal{SL}}$	$0.10(\pm0.32)^{\mathcal{SL}}$	$\mathbf{2.80(\pm4.85)}^{\mathcal{HG,HL,SG}}$
XV	Usefulness of image results	$2.45(\pm0.93)$	$2.60(\pm0.84)$	$2.40(\pm0.84)$	$\mathbf{2.90(\pm1.37)}$
XVI	Usefulness of video results	$2.27(\pm1.10)$	$2.30(\pm1.06)$	$\mathbf{2.40(\pm0.84)}$	$2.10(\pm0.74)$
XVII	Usefulness of web results	$4.64(\pm0.50)$	$\mathbf{4.70(\pm0.67)}$	$4.50(\pm0.71)$	$\mathbf{4.70(\pm0.48)}$
XVIII	Focused attention ($\alpha = 0.846$)	$3.45(\pm0.80)$	$\mathbf{4.33(\pm0.49)}$	$3.70(\pm1.20)$	$3.40(\pm0.86)$
XIX	Experience ($\alpha = 0.791$)	$\mathbf{4.39(\pm0.39)}$	$4.12(\pm0.78)$	$3.98(\pm0.49)$	$4.15(\pm0.68)$
XX	Aesthetics ($\alpha = 0.942$)	$3.45(\pm0.77)$	$\mathbf{3.62(\pm1.13)}$	$3.33(\pm0.91)$	$3.50(\pm0.66)$
XXI	Felt involved ($\alpha = 0.647$)	$\mathbf{4.00(\pm0.56)}$	$\mathbf{4.00(\pm0.67)}$	$3.80(\pm0.74)$	$3.77(\pm0.55)$
XXII	Effectiveness ($\alpha = 0.735$)	$\mathbf{4.35(\pm0.40)}$	$4.30(\pm0.36)$	$3.92(\pm0.43)$	$4.12(\pm0.61)$
XXIII	Usability ($\alpha = 0.891$)	$3.18(\pm0.38)$	$2.68(\pm1.28)$	$2.87(\pm0.66)$	$\mathbf{3.22(\pm0.69)}$

Table 4: User interactions for different task complexity across all search interfaces. † indicates that there is a significant main effect of task complexity on that particular user interaction. $\mathcal{N}$, $\mathcal{R}$, $\mathcal{U}$, $\mathcal{A}$ indicate significant difference with navigational, remember, understand and analyse tasks. Maximum values for each interaction is highlighted in bold.

Row	Interactions	Navigational	Remember	Understand	Analyze
I	Web result clicks †	$4.00(\pm2.65)$	$3.76(\pm3.25)^{\mathcal{A}}$	$4.12(\pm2.55)$	$\mathbf{5.34(\pm4.07)}^{\mathcal{R}}$
II	Mean web result reading time (s)†	$14.99(\pm11.62)^{\mathcal{A,U}}$	$20.57(\pm21.18)$	$24.05(\pm22.64)^{\mathcal{N}}$	$\mathbf{26.91(\pm29.18)}^{\mathcal{N}}$
III	Mean session duration (s)†	$69.95(\pm52.66)^{\mathcal{A,U}}$	$97.11(\pm76.00)^{\mathcal{A}}$	$106.90(\pm62.22)^{\mathcal{N}}$	$\mathbf{134.75(\pm73.98)}^{\mathcal{N,R}}$
IV	Maximum web result click depth	$3.32(\pm2.59)$	$3.88(\pm2.27)$	$3.85(\pm2.37)$	$\mathbf{4.27(\pm2.77)}$
V	Mean web result hover duration (s)	$0.86(\pm1.98)$	$\mathbf{3.26(\pm14.43)}$	$2.05(\pm8.28)$	$2.31(\pm9.89)$
VI	Mean snip. text hover duration (s)	$0.08(\pm0.22)$	$\mathbf{0.12(\pm0.24)}$	$0.07(\pm0.11)$	$0.08(\pm0.15)$
VII	Image clicks (SERP)†	$0.17(\pm0.38)^{\mathcal{A}}$	$\mathbf{0.44(\pm1.00)}^{\mathcal{A}}$	$0.12(\pm0.40)$	$0.00(\pm0.00)^{\mathcal{N,R}}$
VIII	Video clicks (SERP)	$0.00(\pm0.00)$	$1.15(\pm5.59)$	$0.12(\pm0.40)$	$\mathbf{1.88(\pm11.40)}$
IX	Image hovers (SERP)†	$1.10(\pm2.45)^{\mathcal{R}}$	$\mathbf{4.83(\pm10.20)}^{\mathcal{N,U,A}}$	$1.07(\pm2.53)^{\mathcal{R}}$	$0.54(\pm1.98)^{\mathcal{R}}$
X	Video hovers (SERP)	$0.88(\pm2.61)$	$\mathbf{4.49(\pm15.71)}$	$1.90(\pm5.51)$	$1.29(\pm4.47)$
XI	Image clicks (image results page) †	$0.07(\pm0.35)$	$\mathbf{0.32(\pm0.65)}^{\mathcal{U,A}}$	$0.12(\pm0.40)^{\mathcal{R}}$	$0.02(\pm0.16)^{\mathcal{R}}$
XII	Video clicks (video results page)	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$\mathbf{0.15(\pm0.69)}$	$0.02(\pm0.16)$
XIII	Image hovers (image results page)†	$1.15(\pm4.11)$	$\mathbf{4.37(\pm10.92)}^{\mathcal{A}}$	$0.73(\pm2.55)$	$0.07(\pm0.35)^{\mathcal{R}}$
XIV	Video hovers (video results page)	$0.02(\pm0.16)$	$0.00(\pm0.00)$	$\mathbf{0.56(\pm2.21)}$	$0.12(\pm0.64)$

more interaction than the less complex *Remember* or *Understand* tasks. Although pairwise comparisons do not show a significant difference in web result clicks between different tasks for each SERP, we observe a trend similar to **O2**—more complex tasks requiring higher document clicks. From Figure 2*(b)*, we see that participants across all SERP types take longest to finish *Analyse* tasks and least amount of time to finish *Navigational* tasks (also in line with **O4**). Finally, Figure 2*(c)* corroborates our findings from §4.2, as we see that participants across both SERP types interact most with image results for *Remember* tasks compared to other tasks. As mentioned earlier, this observation is contrary to what Arguello et al. [4] observed in their study (**O6**).

In Figure 3, we plot the distribution of where (which rank) participants made their first click of web results for *Navigation* and *Analyse* tasks. The **HL** SERP is the only one where the web results are "broken" by vertical results at rank three, and as a result, we observe that most of the first clicks for both tasks appear before rank four (subplot **(b)** of Figure 3). For the other SERPs, the first click distribution for the tasks is more uniform. This is especially prominent for *Analyse* tasks, where we see due to the absence of verticals on **SL** SERP (comparing *Analyse HL* and *Analyse SL* in subplot **(b)** of Figure 3) participants are willing to go further down the list before their first click. We also expect a peak around the first result for *Navigation* tasks, which is true for all SERP types
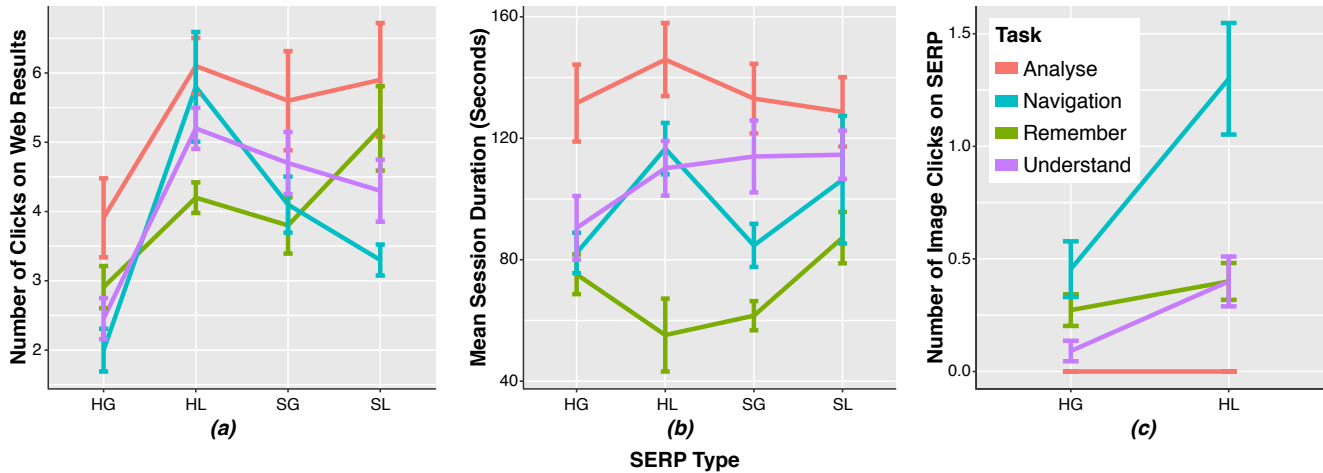
Figure 2: Interaction plots, showing effects of SERP types and task complexity over: *(a)* **clicking on web results;** *(b)* **the mean session duration (in seconds); and** *(c)* **clicks on images presented on the SERP.**

except *Navigation HG* in subplot **(a)**. Either the participants using that SERP type prefer to not click a lot as is evident from Table 3 (fewest web result clicks by **HG** participants), or they chose to explore more before their first click. It has been observed in earlier works [13, 14] that participants have a trust bias for list SERPs (they click on web results appearing higher up the ranked order). The trust bias had been found previously to be less prevalent in grid SERPs [14]. We also find evidence of similar user interaction in subplot **(b)** compared to subplot **(a)** where participants are more open to exploration before their first click. To conclude, we find for **RQ3**, that the interplay between SERP types and task complexity does not have a significant effect on user interactions.

4.4 RQ4: Perceived Experience of SERPs

Turning our attention to the post-experiment surveys, we observe little difference in participant ratings of the systems (**XVIII-XXIII**, Table 3). This is in line with both Arguello et al. [4] and Siu and Chaparro [29], who also did not find significant differences in user ratings for different interfaces. Therefore, we can confirm **O8**.

We also observe that web search results on average are perceived to be more useful (**XVII**, Table 3) than image or video results (**XV-XVI**, Table 3). This is in line with the click behaviour of participants. Across all SERP types, they clicked on more web results than they did on images or videos. Arguello et al. [4] also found the overall number of vertical clicks to be lower than that on web results. Image results were perceived to be more useful by **SL** participants followed by their **HL** counterparts (**XV**, Table 3), which is reflected in their behaviour as well. While the former has the most interactions with images present on the image results page (**XI-XIII**, Table 3) compared to participants in other cohorts, the latter interacted most with images present on the SERP (**VII-IX**, Table 3).

5 CONCLUSION

Summary In this work, we set out to answer the question of how four different types of SERP and four different types of tasks of varying levels of complexity affect user interaction with web, image

and video results. We also explore whether observations about users and their interactions from the studies of Arguello et al. [4] and Siu and Chaparro [29] hold with contemporary SERPs. We observed the following findings with respect to our research questions.

RQ1 The SERP has a significant main effect on the number of clicks on web results and the number of hovers on videos on the video results page, but not on other user interactions.

RQ2 Task complexity has a significant effect on user interactions. While participants interact more with web results as the task becomes more complex, we observe significantly more interactions with image results for *Remember* tasks compared to the more complex *Analyse* or *Understand* tasks.

RQ3 The interplay between SERP types and task complexity does not have a significant effect on user interactions.

RQ4 There is little difference in the evaluation of the four SERP types by participants.

Out of eight observations, we found evidence to confirm two (**O7**, **O8**), with partial evidence for a further four (**O2**, **O3**, **O4**, **O5**). These findings indicate that the user interactions over different interfaces for solving tasks of varying complexity have remained mostly similar over time. However, we employed different information needs—and recruited different participants—from the prior studies. Nevertheless, the evidence contrary to **O1** and **O6** has interesting implications—introducing SERPs that users are not familiar with might result in a decrease in interaction. Although the grid layout can present search results in a condensed format (displaying more items in a given screen space compared to the list layout), users might still end up exploring more in the familiar list layout. Additionally, interactions with vertical results are not only dependent on the complexity of the tasks, but also the type of information need. As we observed, certain simpler tasks might warrant more interaction with vertical results than more complex tasks [30].

Reproducing IIR studies Several variables exist that might affect the observations of an IIR study. An unexhaustive list includes the selection of users, interfaces, and task types. Although both Arguello et al. [4] and Siu and Chaparro [29] described how their
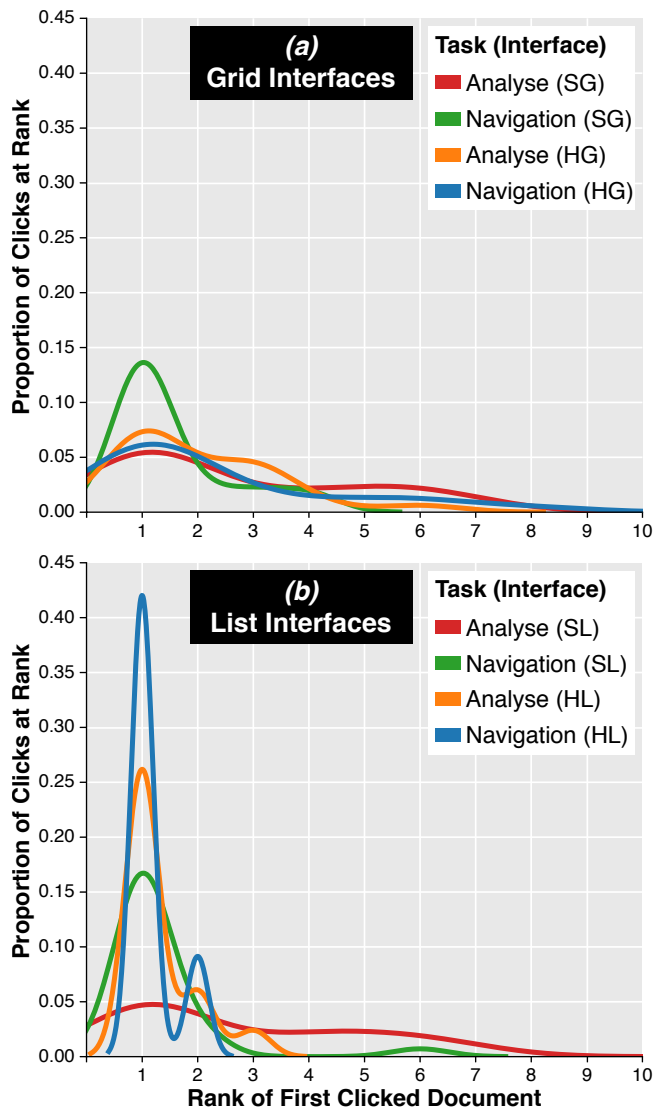
Figure 3: Distribution of ranks of the first clicked web results for participants over both grid-based interfaces SL and HL (a), and list-based interfaces SG and HG (b).

respective interfaces looked, they did not point to any resources which would help us replicate them. Moreover, we believe that the more users become familiar with a particular interface, the more important it is to present a similar interface to them during a study examining their behaviours. As mentioned in Section 3.1, we have created templates of SERPs that resemble google.com and you.com, and released them for further use. We believe our templates will be useful for the community to eliminate confounding variables in IIR studies that might arise due to SERP presentation. Secondly, Arguello et al. [4] and Siu and Chaparro [29] did not mention the entire set of tasks used in their studies. As a result, we came up with our tasks of different complexity, as presented in Table 1. Two studies by Urgo et al. [33, 34] both list examples of tasks pertaining to different complexities which also offer useful resources for

future IIR studies. Our tasks differ with respect to the fact that we designed tasks that specifically enticed participants to not only look at web results, but also to image and video search results as well. It is important to have a fixed set of tasks and similar interfaces to reproduce and enable reliable comparison of observations (e.g., the number of queries, documents opened, etc.) with prior IIR studies. Lastly, in most cases, it will not be possible to have the same participants while reproducing IIR studies. Crowdsourcing provides a solution for capturing user interactions as it has been shown that there is little difference in the quality between crowdsourced and lab-based studies [39]. Power analysis can be used to determine the number of participants required given the experimental conditions of a particular study. It also might be useful to release experimental logs from these studies, after careful ethical checks and considerations. This will permit future researchers to examine them closely, and use them to develop, for example, models of user interaction and search behaviour.

Limitations and Future Work There are several areas with scope for future refinement. First, although we tried to select information needs that cover a broad range of topics, we cannot be certain that the results generalise to information needs with other characteristics. Second, we did not provide querying functionality to users—and hence it will be worthwhile to explore if that has an overall effect on user interactions. Thirdly, the positions of vertical results on the main page of the SERP were fixed, and we know from previous work [28, 30] that user interactions with verticals is affected by where they are displayed on the SERP. In the future, we aim to investigate varying the position of verticals on list and grid interfaces, and their effect on user interaction. The findings from this study can be further applied to designing and evaluating SERP presentations and the placement of heterogeneous content. Understanding and modelling user interactions will also help us work on methodologies for interface optimisation [35] and SERP evaluation, along the same veins of prior studies [5, 9, 26, 31, 38].

REFERENCES

[1] L. W. Anderson and D. R. Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* Longman,.

[2] J. Arguello and R. Capra. 2014. The effects of vertical rank and border on aggregated search coherence and search behavior. In *Proceedings of the 23rd acm international conference on conference on information and knowledge management.* 539–548.

[3] J. Arguello, F. Diaz, J. Callan, and B. Carterette. 2011. A methodology for evaluating aggregated search results. In *European Conference on Information Retrieval.* Springer, 141–152.

[4] J. Arguello, W.C. Wu, D. Kelly, and A. 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.* 435–444.

[5] L. Azzopardi, P. Thomas, and N. Craswell. 2018. Measuring the utility of search engine result pages: an information foraging based measure. In *The 41st international acm sigir conference on research & development in information retrieval.* 605–614.

[6] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2016. UQV100: A test collection with query variability. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.* 725–728.

[7] H. Bota, K. Zhou, and J.M. Jose. 2016. Playing your cards right: The effect of entity cards on search behaviour and workload. In *Proceedings of the 2016 acm on conference on human information interaction and retrieval.* 131–140.

[8] G. Buscher, R.W. White, S. Dumais, and J. Huang. 2012. Large-scale analysis of individual and task differences in search result page examination strategies. In *Proceedings of the fifth ACM international conference on Web search and data mining.* 373–382.

[9] A. Chuklin and M. de Rijke. 2016. Incorporating clicks, attention and satisfaction into a search engine result page evaluation model. In *Proceedings of the 25th acm international on conference on information and knowledge management.* 175–184.

[10] S.T. Dumais, G. Buscher, and E. Cutrell. 2010. Individual differences in gaze patterns for web search. In *Proceedings of the third symposium on Information interaction in context.* 185–194.

[11] M. Hearst. 2009. *Search user interfaces.* Cambridge University Press.

[12] J. Jiang, D. He, and J. Allan. 2014. Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval.* 607–616.

[13] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *Acm Sigir Forum,* Vol. 51. Acm New York, NY, USA, 4–11.

[14] Y. Kammerer and P. Gerjets. 2014. The role of search result position and source trustworthiness in the selection of web search results when using a list or a grid interface. *International Journal of Human-Computer Interaction* 30, 3 (2014), 177–191.

[15] D. Lagun, C.H. Hsieh, D. Webster, and V. Navalpakkam. 2014. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval.* 113–122.

[16] O. Levi, I. Guy, F. Raiber, and O. Kurland. 2018. Selective cluster presentation on the search results page. *ACM Transactions on Information Systems (TOIS)* 36, 3 (2018), 1–42.

[17] Z. Liu, Y. Liu, K. Zhou, M. Zhang, and S. Ma. 2015. Influence of vertical result in web search examination. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval.* 193–202.

[18] D. Maxwell, L. Azzopardi, and Y. Moshfeghi. 2017. A Study of Snippet Length and Informativeness: Behaviour, Performance and User Experience. In *Proceedings of the 40th ACM SIGIR.* 135–144.

[19] D. Maxwell and C. Hauff. 2021. LogUI: Contemporary Logging Infrastructure for Web-Based Experiments. In *Proceedings of the 43th ECIR.* (In press).

[20] A. Namoun. 2018. Three column website layout vs. grid website layout: An eye tracking study. In *International Conference of Design, User Experience, and Usability.* Springer, 271–284.

[21] V. Navalpakkam, L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. 2013. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *Proceedings of the 22nd international conference on World Wide Web.* 953–964.

[22] H.L. O'Brien and E.G. Toms. 2010. The development and evaluation of a survey to measure user engagement. *Journal of the American Society for Information Science and Technology* 61, 1 (2010), 50–69.

[23] H.L. O'Brien. 2010. The influence of hedonic and utilitarian motivations on user engagement: The case of online shopping experiences. *Interacting with computers* 22, 5 (2010), 344–352.

[24] K. Rodden and X. Fu. 2007. Exploring how mouse movements relate to eye movements on web search results pages. (2007).

[25] K. Rodden, X. Fu, A. Aula, and I. Spiro. 2008. Eye-mouse coordination patterns on web search results pages. In *CHI'08 extended abstracts on Human factors in computing systems.* 2997–3002.

[26] T. Sakai and Z. Zeng. 2020. Retrieval evaluation measures that agree with users' SERP preferences: Traditional, preference-based, and diversity measures. *ACM Transactions on Information Systems (TOIS)* 39, 2 (2020), 1–35.

[27] S. Salimzadeh, D. Maxwell, and C. Hauff. 2021. The Impact of Entity Cards on Learning-Oriented Search Tasks. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval.* 63–72.

[28] Y. Shao, J. Mao, Y. Liu, M. Zhang, and S. Ma. 2022. From linear to non-linear: investigating the effects of right-rail results on complex SERPs. *Advances in Computational Intelligence* 2, 1 (2022), 1–16.

[29] C. Siu and B.S. Chaparro. 2014. First look: Examining the horizontal grid layout using eye-tracking. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* Vol. 58. SAGE Publications Sage CA: Los Angeles, CA, 1119–1123.

[30] S. Sushmita, H. Joho, M. Lalmas, and R. Villa. 2010. Factors affecting click-through behavior in aggregated search interfaces. In *Proceedings of the 19th ACM international conference on Information and knowledge management.* 519–528.

[31] P. Thomas, A. Moffat, P. Bailey, F. Scholer, and N. Craswell. 2018. Better effectiveness metrics for serps, cards, and rankings. In *Proceedings of the 23rd australasian document computing symposium.* 1–8.

[32] P. Thomas, F. Scholer, and A. Moffat. 2013. What users do: The eyes have it. In *Asia information retrieval symposium.* Springer, 416–427.

[33] K. Urgo, J. Arguello, and R. Capra. 2019. Anderson and krathwohl's two-dimensional taxonomy applied to task creation and learning assessment. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval.* 117–124.

[34] K. Urgo, J. Arguello, and R. Capra. 2020. The Effects of Learning Objectives on Searchers' Perceptions and Behaviors. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval.* 77–84.

[35] C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. 2013. Incorporating vertical results into search click models. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval.* 503–512.

[36] Z. Wu, M. Sanderson, B.B. Cambazoglu, W.B. Croft, and F. Scholer. 2020. Providing Direct Answers in Search Results: A Study of User Behavior. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.* 1635–1644.

[37] X. Xie, Y. Liu, X. Wang, M. Wang, Z. Wu, Y. Wu, M. Zhang, and S. Ma. 2017. Investigating examination behavior of image search users. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval.* 275–284.

[38] F. Zhang, J. Mao, Y. Liu, X. Xie, W. Ma, M. Zhang, and S. Ma. 2020. Models versus satisfaction: Towards a better understanding of evaluation metrics. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval.* 379–388.

[39] G. Zuccon, T. Leelanupab, S. Whiting, E. Yilmaz, J.M. Jose, and L. Azzopardi. 2013. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval* 16, 2 (2013), 267–305.