Factors Influencing Viewing Behaviour on Search Engine Results Pages: A Review of Eye-Tracking Research

Dirk Lewandowski [a] & Yvonne Kammerer [b, c]

[a] Hochschule für Angewandte Wissenschaften Hamburg, Hamburg, Germany

[b] Leibniz-Institut für Wissensmedien, Tübingen, Germany

[c] Open University of the Netherlands, The Netherlands

This is a preprint of an article accepted for publication in

Behaviour & Information Technology, https://doi.org/10.1080/0144929X.2020.1761450

Corresponding author

Dirk Lewandowski, Hochschule für Angewandte Wissenschaften Hamburg, Finkenau 35, 22082 Hamburg, Germany. Phone: + 49 40 42875 3621, e-mail: dirk.lewandowski@haw-hamburg.de

Factors Influencing Viewing Behaviour on Search Engine Results Pages: A Review of Eye-Tracking Research

ABSTRACT

Eye-tracking research is beneficial for better understanding user behaviour in search engines. The present paper presents a narrative literature review of eye-tracking studies examining factors influencing users' viewing behaviour on results pages of search engines. Discipline-specific databases from Psychology, Computer Science, and Library and Information Science, as well as one multidisciplinary database have been searched for relevant articles. Criteria for inclusion were that a paper reported empirical results from an eye-tracking study in which effects of a specific factor on users' viewing behaviour on search engine results pages (SERPs) were examined, with inferential statistical results being reported. This led to a set of 41 papers that were further examined. The papers were grouped into three categories according to three types of factors that may affect individuals' web search activities: contextual factors, resource factors, and individual factors. Papers were assigned to these categories and subsequently to sub-categories. Overall, while for some sub-categories, robust findings can be reported, we found results in many sub-categories to be inconclusive. For future research, we recommend a shift from small-scale studies examining single factors to more comprehensive and theory-driven research using larger sample sizes.

KEYWORDS

Search engines; search engine results pages (SERPs); eye-tracking research; literature review; Human-Computer Interaction

1. INTRODUCTION

As eye-tracking technology has become increasingly reliable and less expensive over the years, eye-tracking now has turned into one of the standard methods for conducting user research. As Poole and Ball (2006) put it, "eye movements provide a window into so many aspects of cognition". That is, eye-tracking methodology allows determining whether, for how long, and in which order individuals pay attention to particular elements presented, for instance, on a computer screen (cf. Scheiter & van Gog, 2009). Two main types of eye-tracking data can be obtained with eye-tracking systems: saccades and fixations. Whereas saccades are very fast, ballistic eye movements during which no perceptual intake is possible, fixations are short time periods between saccades when the eye remains still and “fixated” to a certain point. According to the eye-mind-assumption introduced by Just & Carpenter, (1980), what is being fixated by the eyes is assumed to indicate what is being processed in the mind. Thus, eye-tracking data is considered a strong indicator for individuals’ moment-by-moment cognitive processing (cf. Rayner, 1998), which also provides insights into cognitive processes that do not lead to overt actions (e.g., when deciding *not* to click on a particular hyperlink). Moreover, it seems suited to unravel quick and automated or unconscious cognitive processes, which are difficult to express verbally (cf. Scheiter & van Gog, 2009). Common measures that can be derived from eye-tracking recordings are time-based fixation measures, such as the total fixation time, that is, the sum of all fixation durations, directed to an area of interest (AOI), or the average fixation duration, respectively, or count-based fixation measures, such as the number of fixations falling into a defined AOI, or the number of lookbacks to a previously fixated AOI, or the number of particular AOIs that have been fixated (for a minimum fixation duration of, e.g., 100 ms). Other common measures are sequence-based measures that, for instance, indicate the order in which different AOIs are fixated by a participant. Besides, recordings of the pupil size allow to analyse users’ pupil

dilation, which can be an indicator of mental load (e.g., Beatty, 1982) or emotional arousal (e.g., Bradley, Miccoli, Escrig, & Lang, 2008).

Eye-tracking has also been used to gain a better understanding of users' attention allocation and cognitive processing on search engine results pages (SERPs). A SERP is "the complete HTML page output that a search engine serves in response to a search query entered by a user." (Lewandowski, Kerkmann, Rümmele, & Sünkler, 2018). Search engines are a central service of the Internet without which it would be inconceivable for users to find their way on the Web (Tavani, 2012; Varian, 2006). Thus, unsurprisingly, search is one of the two most popular activities on the Internet, rivalled only by e-mail. For instance, 91% of the U.S. population uses search engines (Purcell, Brenner, & Raine, 2012) and Google alone answers more than 2 trillion search queries per year (Sullivan, 2016). Besides, search engines and Google, in particular, enjoy a high level of trust among users. This has been shown, for instance, in studies on the credibility of search results (Purcell et al., 2012), on the assumption of the factual correctness of search results (Purcell et al., 2012), on trust in the "right" result ranking (Pan, Hembrooke, Joachims, Lorigo, Gay, & Granka, 2007), and on the use of search engine rankings as a criterion for the quality of content (Westerwick, 2013).

Perhaps the most popular way to measure users' behaviour on SERPs is by measuring click-throughs, i.e., the ratio of users clicking on different search results links provided by the search engine. However, while from this data one can derive indirect assumptions about users' attention to search results, directly measuring users' eye movements allows for gaining a deeper insight into users' attention to and processing of elements on the SERPs.

In recent years, many studies have shown by the use of eye-tracking methodology that most attention is directed to the first few search results presented on the first SERP, that is, the search results in the top of the list (e.g., Cutrell & Guan, 2007; Granka, Joachims, & Gay, 2004; Guan & Cutrell, 2007; Pan et al., 2007) and that users often show a rather linear viewing behaviour from top to bottom (e.g., Cutrell & Guan, 2007; Papoutsaki, Laskey, &

Huang, 2017; Turpin, Scholer, Billerbeck, & Abel, 2006). This typical behaviour that is referred to as position bias will be further discussed in Section 3.2.1. Position bias does not mean, however, that users' scan paths, that is, the order in which search results are fixated, are strictly linear. Of course, regressions to previously inspected search results occur (cf. Lorigo et al., 2006; Thomas, Scholer, & Moffat, 2013), but still, the search results presented further down in a list are typically inspected later and to a lesser extent than the top results.

There have been some previous review articles on eye-tracking research. For instance, Lund (2016) focuses on eye-tracking research in the field of library and information science, and White's (2016) handbook on interactive information retrieval deals with eye-tracking research in that area in some detail. Furthermore, Alemdag & Cagiltay (2018) recently have reviewed eye-tracking research in the field of multimedia learning. An earlier review article on eye-tracking on SERP viewing behaviours (Granka, Feusner, & Lorigo, 2008) summarizes major methodological issues (such as different types of measurements) together with empirical finding. Besides, Lorigo et al. (2008) summarize the authors' own studies on SERP viewing behaviours in Google and Yahoo and put them into context with related studies.

However, to the best of our knowledge, there is no comprehensive review article on eye-tracking research focusing on SERPs and, specifically, on what factors might influence users' viewing behaviour on SERPs, which is the central question of the present paper. Lazonder and Rouet (2008) distinguished three different types of factors that may affect individuals' web search activities (and, thus, potentially also their viewing behaviour on SERPs). These are (a) contextual factors, that is, characteristics of the context that pre-exists to the search activity, such as different task types or task instructions, (b) resource (or design) factors concerning the tools or the SERP interface available for the web search activity (e.g., how the search results are displayed), and (c) individual factors, that is, characteristics and (cognitive) prerequisites of the users (cf. Lazonder & Rouet, 2008).

The major goal of the present paper, thus, is to provide a narrative literature review of previous research on factors potentially fnfluencing viewing behaviour on SERPs, by (1) classifying previous studies according to the three types of factors proposed by Lazonder and Rouet (2008), (2) identifying factors that may affect viewing behaviour on SERPs, and (3) synthesizing the results achieved so far and to show potential areas for further research.

The remainder of this paper is structured as follows: In the next section, we will describe the body of literature on the topic, as found through our literature search. Then, we define the major topic areas and synthesize the work done in these areas. After that, we discuss theoretical and methodological approaches. Then, we discuss the state of research in the areas and point to fruitful areas for future research. We conclude by highlighting the major issues we found in the body of work on viewing behaviour on SERPs.

2. METHODS

2.1. Literature search

To find literature from a wide variety of disciplines, in May 2018 we conducted searches in the multidisciplinary database Scopus, as well as in discipline-specific databases from Psychology (PsycARTICLES, Psychology and Behavioral Sciences Collection, PsycINFO), Computer Science (ACM Digital Library), and Library and Information Science (LISTA). We did not restrict our searches to particular publication years. We queried the databases for articles containing words related to eye-tracking (e.g., eye-tracking, gaze-tracking) as well as words related to search engines (e.g., search engine, Google, SERP) in the title, abstract or keyword fields.

2.2. Inclusion and exclusion criteria

Criteria for inclusion were that the paper reports empirical results from an eye-tracking study in which the effects of a certain factor on users' viewing behaviour on SERPs were examined. Criteria for exclusion were: (1) The paper reports only preliminary results, on which a follow-up publication is available (this was the case with some conference papers which were later extended to journal papers), (2) the paper does not provide tests for statistical significance of eye-tracking measures regarding the factors under examination, (3) dissertations and grey literature, and (4) publications in a language other than English. Applying our criteria for inclusion and exclusion led to a total of 41 papers that are considered in the analysis presented below. For the papers included, see Table 1, which also includes basic information on these studies. The exclusion table is provided as supplementary data in the appendix.

Table 1: Overview of studies included in the literature review

No.	Reference	Country	Sample	Sample size	Device type	Dependent variable(s)	Between-subjects variables	Number of tasks	Task types	Within-subjects variables	Information need (self-chosen vs. given)	SERP type (fabricated vs. real SERPs)	Search engine
1	Athukorala, Głowacka, Jacucci, Oulasvirta, & Vreeken (2015)	Finland	Computer science researchers	15	Desktop computer	Total fixation time Percentage of gaze points per AOI	N/a	6	3 exploratory tasks (knowledge acquisition, planning, and comparison) and 3 look-up tasks (informational and navigational)	Tasks	Given	Real SERPs	Google Scholar
2	Aula, Majaranta, & Räihä (2005)	Finland	Students from different majors	28	Desktop computer	Number and order of search results fixated before first click	2 participant clusters: Exhaustive evaluators vs. Economic evaluators	10	Informational fact-finding tasks (3 with poor results, 3 with good results, 4 with mixed results)	N/a	Given	Fabricated SERPs (but could be altered)	Google
3	Balatsoukas & Ruthven (2012)	Scotland	Students from different majors with a good level of experience in Web searching	17	Desktop computer	Number of fixations Total fixation time	N/a	1	Freely chosen by the test persons	Search result components (title, description, URL)	Self-chosen	Real SERPs	Google
4	Bilal & Gwizdka (2016)	United States	Students from grades 6 and 8	16	Desktop computer	Number of fixations (reading states vs. scanning states) Total fixation time (reading states vs. scanning states) Rank of first fixated search result	Grade level (6th grade vs. 8th grade)	2 (+ 1 that wasn't analysed)	1 fact-finding task (concrete task differed between grade levels), 1 complex research task (climate change)	Search result position Task type	Given	Real SERPs	Google

5	Brand-Gruwel, Kammerer, van Meeuwen, & van Gog (2017)	Netherlands	college and university students; psychology staff	35	Desktop computer	Number of search results fixated before first click	Domain expertise (experts vs. novices)	2	2 complex research tasks (from psychology)	Concurrent think-aloud vs. cued retrospective think-aloud	Given	Fabricated SERPs	Google
6	Buscher, Dumais, & Cutrell (2010)	United States	Diverse range of backgrounds and professions, between 26 and 60 years old	38	Desktop computer	Fixation impact (modified version of total fixation time)	N/a	32	16 navigational tasks, 16 informational tasks	Task type Quality of ads (relevant vs. irrelevant) Sequence in which ads of different quality were presented	Given	Real SERPs (but first query predefined so that all users see the same first SERP)	Not mentioned
7	Chen & Pu (2010)	N/a	Students or employees in the university (between 20 and 40 years old) from various countries	21	Desktop computer	Total fixation time Number of fixated search results (products)	3 different SERP interface designs	1	Product search task	N/a	Given	N/a	Product search engine
8	Cutrell & Guan (2007)	United States	Diverse range of jobs, backgrounds and education levels, between 18 and 50 years old	18	Desktop computer	Total fixation time	N/a	12	6 informational and 6 navigational tasks	Task type Snippet length (short, medium, long)	Given	Real SERPs	MSN
9	Dickerhoof & Smith (2014)	United States	Undergraduate students	18	Desktop computer	Number of fixations and fixation ratios	N/a	11	The task was to find a website that provided information about the association between two given nouns (e.g., dog and sheep). Participants were instructed to not use the two given nouns as keywords in their queries.	Query terms vs.no-query terms	Given	Real SERPs	Google

10	Dinet, Bastien, & Kitajima (2010)	France	School students from grades 5, 7, 9, 11	89	Desktop computer	Total fixation time	Grade level (5th, 7th, 9th, and 11th grade)	2	2 fact finding tasks (on history and geography topics; concrete topics differed between participants)	Keywords in bold face vs. Keywords not in bold face Familiar search topic vs. unfamiliar search topic	Given	Fabricated SERPs	Google
11	Dumais, Buscher, & Cutrell (2010)	United States	Diverse range of backgrounds and professions, between 26 and 60 years old	38	Desktop computer	Fixation impact (modified version of total fixation time) Number of fixations before the first click	3 participant clusters: Exhaustive searchers vs. Economic-results searchers vs. Economic-ads searchers	32	16 navigational tasks, 16 informational tasks	N/a	Given	Real SERPs (but first query predefined so that all users see the same first SERP)	Not mentioned
12	Eickhoff, Dungs, & Tran (2015)	N/a	University students	17 (Study 1)	Desktop computer	Average fixation duration Number of fixations	N/a	3 out 10	Complex informational tasks (from a set of 10 topics)	Query terms vs.no-query terms	Given (but free choice between two options per task)	Real SERPs	Google
13	Fernquist & Chi (2013)	United States	Active Google+ users	9	Desktop computer	Number of fixations	N/a	16-20	Informational and navigational.	N/a	Given, but individualized for each participant	Real SERPs	Google
14	Gerjets, Kammerer, & Werner (2011)	Germany	University students from different majors	30	Desktop computer	Number of fixated search results Total fixation time	Think-aloud instructions with evaluation prompts vs. Think-aloud instructions without evaluation prompts	1	Complex research task (alternative weight loss methods)	N/a	Given	Fabricated SERPs	Google
15	González-Caro & Marcos (2011)	Spain	Diverse range of professions	58	Desktop computer	Total fixation time Number of fixations	N/a	6 out of 18	11 informational, 3 navigational, 4 transactional	Task type Organic search results vs. sponsored search results	Given	Fabricated SERPs	Google and Yahoo! (half of the queries with each search engine)

16	Guan & Cutrell (2007)	United States	Diverse range of jobs, backgrounds and education levels, between 18 and 50 years old	18	Desktop computer	Number of search results fixated before first click Total fixation time	N/a	12	6 navigational, 6 informational fact-finding tasks	Task type Target position (position 1, 2, 4, 5, 7, or 8)	Given	Fabricated SERPs	MSN
17	Hautala, Kiili, Kammerer, Loberg, Hokkanen, & Leppänen (2018)	Finland	School students from grade 6	36	Desktop computer	First-pass fixation likelihood (in %) Second-pass fixation likelihood (in %)	3 participant clusters: Title readers vs. Title and description readers vs. All components readers	9	9 informational fact-findings tasks	Search result relevance (relevant description vs. irrelevant-description vs. irrelevant-URL vs. all-components irrelevant)	Given	Fabricated SERPs	Experimental (Google-style)
18	Jiang, He, & Allan (2014)	United States	College and university students from different majors	20	Desktop computer	Number of fixated search results Total fixation time Fixation likelihood (in %) Linearity of viewing sequence	N/a	4 out of 20 tasks	Tasks from 2012 TREC session track	Task type (known item, known subject, interpretative, exploratory)	Given	Real SERPs	Experimental search system providing modified Google search results
19	Kammerer & Gerjets (2012)	Germany	University freshmen from different majors	58	Desktop computer	Total fixation time	List interface vs. tabular interface	1	Complex research task (therapies Bechterew's disease)	Type of search results (objective, subjective, commercial)	Given	Fabricated SERPs	Google
20	Kammerer & Gerjets (2013)	Germany	University students from different majors	44	Desktop computer	Total fixation time Linearity of viewing sequence	Think-aloud vs. prompted think-aloud vs. no think-aloud; In addition self-reported prior domain knowledge as continuous predictor (correlational)	1	Complex research task (alternative weight loss methods)	N/a	Given	Fabricated SERPs	Google

21	Kammerer & Gerjets (2014)	Germany	Students from social and natural sciences and humanities	40 (Study 1), 40 (Study 2)	Desktop computer	Number of search results fixated before first click Viewing sequence	Search result order (optimal trustworthiness order vs. reversed trustworthiness order)	1	Complex research task (therapies Bechterew's disease)	Search result trustworthiness (high, medium, low)	Given	Fabricated SERPs	Google
22	Kammerer, Bråten, Gerjets, & Strømsø (2013)	Germany	University students from different majors	79	Desktop computer	Total fixation time	N/a epistemic beliefs as continuous predictors (correlational)	1	Complex research task (therapies Bechterew's disease)	N/a	Given	Fabricated SERPs	Google
23	Kim, Thomas, Sankaranarayana, Gedeon, & Yoon (2015)	Australia	Students from various disciplines	32	Desktop computer (large screen vs. small screen)	Total fixation time Linearity of viewing sequence	N/a	20	10 navigational, 10 informational (fact-finding) tasks	Large screen vs. small screen (i.e., simulated mobile screen) Task type	Given	Fabricated SERPs	Google
24	Lagun, Hsieh, Webster, & Navalpakkam (2014)	United States	Diverse range of occupations, between 18 and 65 years old	24	Smartphone	Total fixation time	N/a	20	Informational fact-finding tasks	Relevance of knowledge graph result (relevant vs. irrelevant) Relevance of instant answer result (relevant vs. irrelevant) Presence of knowledge graph result (present vs. absent)	Given	Fabricated SERPs	Google
25	Liu, Liu, Zhou, Zhang, & Ma (2015)	China	University students from different majors	32	Desktop computer	Percentage of total fixation time (attention distribution in %)	N/a	30	Informational	Type of vertical results (4 different types or none) Position of vertical results Relevance of vertical result	Given	Fabricated SERPs	N/a
26	Lo, Hsieh, & Chiu (2014)	China	Participants with website searching experience (approx.. half male and female, mean	451	Desktop computer	Total gaze time Total number of gazes	N/a	1 (out of 3)	E-commerce product search task	Organic results vs. sponsored results at the top vs. sponsored results at the right-hand side	Given	Fabricated SERPs	Yahoo!

			age: 23- 24 years)										
27	Lorigo, Pan, Hembrooke, Joachims, Granka, & Gay (2006)	United States	Undergraduate students from different majors	23	Desktop computer	Number of fixations Average fixation duration Pupil dilation Linearity of viewing sequence	Gender (males vs. females)	10	5 navigational and 5 informational tasks	Task type	Given	Real SERPs	Google
28	Lu & Jia (2014)	China	Undergraduate and graduate students	58	Desktop computer	Number of fixated search results Number of regressions Total fixation time Number of fixations Average fixation duration Pupil dilation	Ranking order of images (normal vs. reversed)	10	5 specific and 5 general image search tasks	Task type	Given	Fabricated SERPs	Imagine search engine
29	Marcos, Gavin, & Arapakis (2015)	Spain	Participants all were frequent users of web search engines (between 18 and 58 years old)	60	Desktop computer	Time to first fixation on an AOI Number of fixations before fixating an AOI Total fixation time Number of fixations Visit duration	N/a	10	Simple tasks (informational and navigational) from the tourism domain	Top-ranked results vs. bottom-ranked results Rich snippets vs. plain snippets	Given	Fabricated SERP	Google

30	Muntinga & Taylor (2017)	Netherlands	Hospital patients (from 18 to over 60 years old)	50	Desktop computer	Number of fixations	Participants who selected genuine pharmacy search results vs. Participants who selected rogue pharmacy search results	5	1 transactional, 1 navigational, and 1 informational task plus 2 tasks with only two search results each	Task type	Given	Fabricated SERP	Google
31	Muralidharan, Gyongyi, & Chi (2012)	United States	Non-computer-programmers from the authors' organization	12 (Study 2)	Desktop computer	Number of fixations	N/a	36	Simple informational tasks	Snippet length (1-line, 2-lines, 4-lines) Picture size (large vs. small) Annotation above snippet vs. Annotation below snippet	Given	Fabricated SERP	Google
32	Oliveira, Aula, & Russell (2009)	United States	Google employees	16 (Study 1); 13 (Study 2)	Desktop computer	Pupil dilation	N/a	24	N/a	Relevant vs. irrelevant text results; Relevant vs. irrelevant image results	Given	Three pre-selected search results per task	Google
33	Pan, Hembrooke, Joachims, Lorigo, Gay, & Granka (2007)	United States	Undergraduate students from various majors	16	Desktop computer	Number of fixated search results Total fixation time Number of fixations Pupil dilation	Normal ranking vs. swapped ranking vs. reversed ranking	10	5 navigational and 5 informational fact-finding tasks	Top-five search results vs. bottom-five search results	Given	Real SERP (but the order of the search results was experimentally manipulated	Google
34	Rele & Duchowski (2005)	United States	Participants all had a minimum of 5 years internet experience (between 20 and 29 years old)	16	Desktop computer	Average fixation duration Number of fixations	N/a	4 (out of 8)	4 navigational and 4 informational fact-finding tasks	Interface type (list vs. tabular) Task type	Given	Fabricated SERP	Google

35	Rovira, Capdevila, & Marcos (2014)	Spain	Regular Internet users (between and 18 and over 50 years old)	50	Desktop computer	Total fixation time	N/a	2	Tasks related to news content	Search result component (title vs. description vs. source)	Given	Fabricated SERP	Google Noticias (GoogleNews, Spain)
36	Saito, Terai, Egusa, Takaku, Miwa, & Kando (2009)	Japan	Graduate students and undergraduates (LIS and other majors)	16	Desktop computer	Number of eye-gaze points (no automatic fixation-based analysis of eye-tracking data)	N/a	2	1 report-writing task and 1 trip-planning task	Task type	Self-chosen	Real SERP	Participant's favorite search engine
37	Schultheiß, Sünkler, & Lewandowski (2018)	Germany	Undergraduate students from different majors	25	Desktop computer	Number of fixations	Normal ranking vs. swapped ranking vs. reversed ranking	10	5 navigational and 5 informational fact-finding tasks	Search result position	Given	Fabricated SERP	Google
38	Siu & Chaparro (2014)	United States	Participants were recruited from a university (between 18 and 46 years old)	45	Desktop computer	Total fixation time Number of fixations	N/a	12	6 navigational and 6 informational tasks	Interface type (grid vs. list) Task type Top row search results vs. left-column search results	Given	N/a	Experimental (similar to Google)
39	Turpin, Scholer, Billerbeck, & Abel (2006)	Australia	Predominately postgraduate students	9	Desktop computer	Number of fixations Number of fixated search results Total fixation time	N/a	10	5 navigational tasks and 5 informational fact-finding tasks	Task type	Given	N/a	Sensis.com.au search engine

40	Walhout, Oomen, Jarodzka, & Brand-Gruwel (2017)	Netherlands	Secondary-school students (avg. age 14.5 years)	15	Desktop computer	Number of fixated search results Total fixation time	N/a	3	1 fact-finding task, 1 cause-and-effect task, and 1 controversial research task	Task type	Given	Real SERPs	Google
41	Xie, Liu, Wang, Wang, Wu, Wu, … Ma (2017)	China	Undergraduate students (science, engineering and arts)	40	Desktop computer	Total fixation time Time to first fixation Viewing sequence	N/a	21	8 specific queries (e.g., “image of Doraemon”), 7 generic queries (e.g., “image about New York City’s beautiful scenery”), 6 abstract queries (e.g., “image that expresses pleasant surprise”)."	Search result position	Given	Fabricated SERPs	Results from a commercial search engine

Note. * Country in which the study was conducted not stated in text; information given in the table derived from first author’s affiliation or information given in the text (e.g., screenshots showing SERPs in Chinese).

2.3. Descriptive statistics

In the following we describe the selected articles concerning countries of origin, number and type of participants, materials and study design, and dependent measures.

Seventeen (41.4 %) of the studies were conducted in Europe, 15 (36.6%) in North America, 5 (12.2%) in Asia, 2 (4.9%) in Australia, and for two studies no respective information was available. Of the 17 studies conducted in Europa, most studies were conducted in Germany (6), the Netherlands (3), Finland (3) and Spain (3). The number of participants per study ranged from 9 to 451 (median: 28, $M = 41.51$, $SD = 66.62$), with low-scale studies being the norm. 58.5% of the studies were conducted with university or college students (and/or with academic staff) as study participants, 34.1% were conducted with other adults (e.g., diverse samples or patients), and 7.3% with school students.

58.5% of the studies examined viewing behaviour on fabricated SERPs, 34.1% on actual SERPs (i.e., when searching on the open Internet). For the rest of the studies, no information on the kinds of SERPs used was available. All but one study used an experimental or quasi-experimental design, with 23 studies using a within-subjects design, five studies a between-subjects design, and 11 studies a mixed design comprising both between-subjects and within-subjects variables; one study was only correlational (examining continuous predictors). Most of the studies analysed time-based and/or count-based fixation measures for particular Areas of Interest (AOIs), such as individual search results. Total fixation time (26 studies) and number of fixations (30 studies) were the most common dependent measures. A few studies additionally analysed other measures, for instance, related to viewing sequences. In four studies, pupil dilation was recorded.

Furthermore, for our analysis of the literature, we clustered the articles according to the three types of factors proposed by Lazonder and Rouet (2008) that may affect individuals' web search activities (and, thus, also their viewing behaviour on SERPs). As outlined above,

these are contextual factors, resource factors, and individual factors. In our sample, 13 studies examined contextual factors, 29 resource factors, and eight individual factors (note that categories are not mutually exclusive, i.e., some studies examined factors from different categories in combination). Furthermore, we grouped the articles into several sub-categories of which some were already proposed by Lazonder and Rouet (e.g., prior domain knowledge as an individual factor, task type as a contextual factor, or the interface design as a resource factor). Others (e.g., search result elements as a resource factor) emerged bottom-up, based on the number of papers published on these topics. Figure 1 illustrates the factor groups and gives examples of the sub-categories.

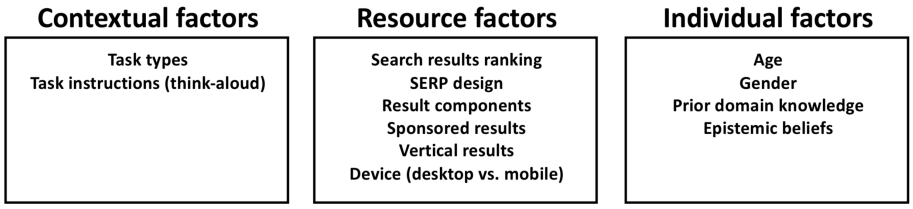

Figure 1: Three types of factors that may affect individuals' viewing behaviour on SERPs, with sub-categories that have been investigated in studies included in this review

3. RESULTS

3.1. Contextual factors

As reported in the following, most studies that have addressed contextual factors that might influence individuals' viewing behaviour on SERPs have examined effects of different task types. Two more studies investigated effects of task instructions (or more specifically, the effect of concurrent thinking-aloud methodology).

3.1.1. Effects of task types

Individuals typically conduct web searches in order to meet a specific need that is related to some task at hand (Broder, 2002). Based on the task, which can be an external task assignment received by a teacher or by the experimenter or an internally defined task, a specific need is defined (which might further evolve during the search). The need is then translated into a concrete query (consisting of one or several query terms) that is posed to a search engine, which in turn provides a list of search results. From the provided search results, one or multiple websites are accessed. This process of entering queries, evaluating search results, and accessing websites can be iterated several times. According to Broder's (2002) "taxonomy of web search", three different types of query intents can be distinguished: informational queries, where the user aims at acquiring information on a topic he or she is interested in by retrieving one or several web pages; navigational queries, where the user aims at navigating to a particular web page already known or where the user at least assumes that a specific web page exists; and transactional queries, where the user aims at finding a web page where a further transaction (e.g., shopping, downloading software, or playing a game) can take place (cf. Lewandowski, 2015).

Most of the studies reported in the following used Broder's (2002) classification to distinguish between three different tasks types, that is, informational tasks (i.e., participants are instructed to retrieve information in order to answer a question at hand), navigational tasks (i.e., participants are instructed to find a specific web page), and transactional tasks (i.e., participants are instructed to make a transaction). In this regard, it should be noted that completing a task can require more than one query.

Lorigo et al. (2006) examined participants while performing five informational (fact-finding) tasks and five navigational tasks, analysing the number of fixations, average fixation duration, pupil dilation, and linearity of viewing sequences on SERPs. However, no effects of task type on any eye-tracking measure on SERPs were found.

In contrast, González-Caro and Marcos (2011), who examined participants performing a set of informational, navigational, and transactional tasks, found significant differences between the three task types for total fixation times and number of fixations on SERPs. Participants, on average, had longer total fixation times and more fixations for informational and transactional searches than for navigational searches (however, no pairwise comparisons were calculated). Moreover, in the transactional searches, individuals fixated sponsored search results for a higher proportion of time (29%) than in the two other types of searches (2.6% and 7.6%, respectively).

Rele and Duchowski (2005) also examined differences in navigational and informational tasks. Task type did not significantly affect average fixation duration. However, for navigational tasks, the number of fixations on the description element of the search results was significantly higher than for the informational tasks. Yet, one navigational task was perceived as particularly difficult by participants, which might alternatively explain the finding (as this task might have required more careful reading).

Guan and Cutrell (2007) compared participants' viewing behaviour while performing six navigational and six informational search tasks. For each task, only one page (i.e., the target result) in the SERP contained the required information, which participants had to identify. The position of the target search result was varied between tasks across six different positions (position 1, 2, 4, 5, 7, or 8). Results showed that the lower the target position, the more search results were fixated, without any effects of task type. Moreover, total fixation time on search results decreased with lowering target positions, though also to a similar extent in both task types. Additional results from this study reported in Cutrell and Guan (2007) will be outlined below (see Section "Effects of search result components").

In a study investigating the effects of advertisements (sponsored results) on the SERPs (also see section 3.2.5), Buscher, Dumais, and Cutrell (2010) also investigated effects of task types on viewing behaviour on organic and sponsored results. They found that participants

spent more time on the SERPs for informational tasks than for navigational tasks, and that this additional time was mostly spent on organic results (and particularly on the first two search result positions). Similarly, Kim, Thomas, Sankaranarayana, Gedeon, and Yoon (2015) also found that for informational tasks fixation times on SERPs were longer and users looked more on search results of lower ranks than for navigational tasks (both when SERPs were presented on a small and a large screen).

Other studies have focused explicitly on informational tasks, which can be further distinguished into simple fact-finding (or look-up) tasks (closed-ended tasks) and more complex or exploratory information search tasks with the purpose of learning or making informed decisions about controversial issues (e.g., Kellar, Watters, & Shepherd, 2006; Marchionini, 2006). Whereas the former usually possess one correct answer or a set of correct answers (closed-ended tasks), for the latter multiple competing viewpoints exist (open-ended tasks). Accordingly, Athukorala, Głowacka, Jacucci, Oulasvirta, and Vreeken, (2015) compared users' (computer scientists) search behaviour when performing three look-up tasks and three exploratory search tasks in Google Scholar. Concerning eye-tracking data obtained, they analysed the proportion of time spent on each of the first ten search results before the first click. No significant differences were found between the tasks.

Walhout, Oomen, Jarodzka, and Brand-Gruwel (2017) examined secondary-school students while performing web searches with Google (i.e., no fabricated SERPs and no predefined search queries) for three informational search tasks that varied in complexity: one simple fact-finding task and two more complex tasks of which one required to explain the effect of a mirage (cause-and-effect task) and the other to learn about potential consequences of mobile phone radiation (controversial task). Regarding eye-tracking data, the authors analysed the number of search results that, on average, were fixated by participants and the inspection likelihood of a search result as a function of its position in the SERP (i.e., the percentage of participants that fixated a search result at a particular position in the SERP).

Results showed that in the cause-and-effect task, on average, more search results were fixated than in the controversial task (which might be due to the higher number of queries conducted), and inspection likelihood of search results further down the SERP declined considerably faster in the controversial task than in the other two tasks.

Saito, Terai, Egusa, Takaku, Miwa, and Kando (2009) compared participants' eye movements when performing two open Web search tasks on self-chosen topics related to (a) a research report and (b) a trip planning. Results revealed more eye-gaze points (note that in this study, eye-gaze points were identified manually) on organic search results for the report-writing task than for the trip-planning task. By contrast, in the trip-planning task, more eye-gaze points were directed towards sponsored search results. Moreover, in the report-writing task, participants directed more eye-gaze-points on search result positions 4 and 7 than in the trip-planning task. However, these effects might be simply due to specific characteristics of the search results presented in positions 4 and 7 (e.g., more text, higher relevance, more difficult, etc.).

Lu and Jia (2014) compared general and specific tasks during image search. Specific tasks required searching for specific objects (e.g., the Eiffel tower), whereas general tasks required the search for broader image categories (e.g., a large crowd of people). Images were presented in a 5x3 grid format in the SERP. Results showed that in general tasks participants viewed significantly more results, made more regressions, and had larger pupil diameters than in specific tasks, which could be indications for a higher effort invested in the former tasks.

Jiang, He, and Allan (2014) investigated four task types. First, tasks could be either fact-finding or "pursuing intellectual understanding of a topic" (p. 607), second, they could either have a specific goal or an ill-developed and undefined goal. This leads to known-item tasks (fact-finding + specific), known-subject tasks (fact-finding + ill-defined), interpretive tasks (intellectual + specific), and exploratory tasks (intellectual + ill-defined). In total 20 tasks were used (5 per task type), out of which each participant performed 4 tasks (1 per task

type). Results revealed several differences between task types in terms of users' viewing behaviour on the SERPs, which, however, were hard to interpret. Users fixated more search results in the exploratory tasks than in the interpretative and known-subject tasks, but not more than in the known-item tasks. The number of fixations in the known-item tasks was higher than in the interpretive tasks. The total time fixating search results was longer in the known-item and exploratory tasks than in the known-subject and interpretive tasks. The total fixation times per search result were longer in the known-item and exploratory tasks than in the known-subject tasks. In the known-item and the known-subject tasks, linear viewing sequences (from top to bottom) were more likely than in the other two task types. Scan paths were wider (i.e., users scanned a larger area of the SERPs and skipped more search results in between) in the exploratory tasks than in known-item tasks. In the known-item and the known-subject tasks, users were less likely to move up in the results lists again than in the interpretive tasks. As a search session progressed, the number of fixations per query decreased in all tasks except for the interpretive tasks.

To summarize, results of the reviewed studies regarding effects of task types on viewing behaviour are somewhat unclear, with many studies having found no systematic differences, whereas a few studies found various differences (e.g. Jiang et al., 2014). In our opinion, a general problem with these studies is that effects of task types might be heavily influenced by the concrete topics, by participants' familiarity with these topics, and last but not least by the concrete individual search results provided by the search engine, which can differ in their length, relevance, number of bold terms included, etc.

3.1.2. Effects of thinking-aloud instructions

Eye-tracking methodology is often combined with thinking-aloud methodology, that is, by asking participants to verbalize everything that comes to their mind while performing a search task (cf. Ericsson & Simon, 1993). In order to examine whether concurrent thinking-aloud

alters users' viewing behaviour on SERPs, Brand-Gruwel, Kammerer, van Meeuwen, and van Gog (2017) compared a situation, in which participants were asked to think aloud during Web search to a situation without concurrent thinking-aloud. In both cases, participants' task was to conduct a Web search task on a complex scientific issue, for which they were provided with fabricated SERP. Results revealed that thinking-aloud had no effect on the number of search results that were visually inspected by the participants before the first click. However, no other eye-tracking measures on the SERPs were reported.

In a similar study, Gerjets, Kammerer, and Werner (2011) examined the effects of neutral thinking-aloud instructions (i.e., to verbalize everything that comes to mind, cf. Ericsson & Simon, 1993), as compared to directed thinking-aloud instructions that prompted to mention evaluation criteria while exploring SERPs and websites. In both conditions, participants' task was to perform a Web search on a controversial health-related issue, for which they were provided with three fabricated SERPs. According to the results, the type of thinking-aloud instructions had no effects on participants' viewing behaviour on SERPs, as measured by the number of search results that were visually inspected and the total fixation time on search results. Besides, Kammerer and Gerjets (2013) compared the data from the neutral thinking-aloud instructions condition, and the directed thinking-aloud instructions condition from the study reported in Gerjets et al. (2011) to a condition without thinking-aloud. They also found no significant differences between the three conditions in terms of total fixation time on search results. However, they found differences regarding the linearity of the viewing sequences (i.e., in which order the search results on a SERP were visually inspected). Participants in the two thinking-aloud conditions had less linear viewing sequences than participants in the condition without thinking-aloud.

To conclude, instructions to evaluate the search results or to verbalize one's thoughts during task processing do not seem to affect the number of or the extent to which search

results are inspected. However, there are first indications that such instructions might result in a more sequential viewing behaviour from top to bottom.

3.2. Resource factors

The studies reported in the following have examined resource (or design) factors concerning aspects of the SERP, that is, effects of how the search results are presented (i.e., the order of the search results or the layout of the SERP) or aspects of the search results themselves (e.g., regarding the search result components or the presentation of additional metadata). One study investigated effects of the device type on participants' viewing behaviour on SERPs.

3.2.1. Effects of the position of search results in SERPs

The probably most popular effect in web search research using eye-tracking is the effect of the position of search results in the SERP on individuals' viewing behaviour. As mentioned in the introduction, multiple studies have shown a position bias, such that most attention is directed to search results presented in the top of a SERP, whereas search results presented further down in the list are typically inspected later and to a lesser extent (e.g., Cutrell & Guan, 2007; Granka et al., 2004; Guan & Cutrell, 2007; Pan et al., 2007; Thomas et al., 2013; Turpin et al., 2006; Walhout et al., 2017).

In order to test the influence of the position of a search result on individuals' viewing behaviour against the actual relevance or the trustworthiness of a search result, a number of studies have systematically compared variations of the order in which search results are presented in the SERP or different positions of target search results in the SERP, respectively. In their seminal study, Pan et al. (2007) used a methodological paradigm in which they experimentally manipulated the relevance order of ten search results on a Google SERP by presenting search results either in the regular order (as provided by the search engine), in a swapped order with the first and the second search result being reversed, or in a

systematically reversed order with the search results ordered from top to bottom (i.e., the first search result was presented at the bottom of the SERP and the tenth search results at the top). Results demonstrated that in the reversed condition participants that were required to conduct a set of navigational and informational fact-finding tasks visually inspected more search results, more often reinspected already fixated search results, and had longer total fixation times and more fixations on the search results than in the normal and in the swapped condition. The latter two conditions, instead, did not differ regarding these measures. However, participants in the reversed condition generally still paid most attention to the search results on top of the SERP (and also selected these results most often), indicating that participants were heavily influenced by the position of the search result in the search engine results list. In a recent replication study, Schultheiß, Sünkler, and Lewandowski (2018) confirmed the results of Pan et al. (2007) concerning viewing behaviour. Overall, participants in the reversed condition had more fixations on SERPs than in the normal or swapped conditions. Moreover, in the reversed condition, the number of fixations was rather evenly distributed over the ten results of the SERP.

Kammerer and Gerjets (2014, Study 1) adapted the methodological paradigm used by Pan et al. (2007) by experimentally manipulating the trustworthiness order of the search results on a SERP for a search on a controversial medical issue. Nine search results, which were all of high topical relevance, were either presented in an optimal order, with the most trustworthy search results presented first and the least trustworthy ones presented last or in a reversed order, with the least trustworthy search results presented first. Results indicated that when the top search results were the least trustworthy ones in the list, participants visually inspected more search results before they accessed the first web page than when the top search results were the most trustworthy ones. However, it should be noted that no respective order effects were found when the search results were presented in a 3x3 grid instead of a list (Kammerer & Gerjets, 2014; Study 2). Likewise, Lu and Jia (2014), who also investigated the

effects of search results presentation order (normal or reversed) during image search in a 5x3 grid interface, also found no order effects.

To conclude, when search results are presented in the form of lists, search results at the top of the list receive the most attention. However, participants do not seem to blindly rely on the position of the search results, as it affects their viewing behaviour if the top search results are of low relevance or trustworthiness.

3.2.2. Effects of alternative SERP interface designs

Several studies examined the effects of alternative SERP interface designs (as compared to a standard Google-like list interface) on users' viewing behaviour on SERPs. One common alternative SERP interface is a grid interface, in which search results are presented in multiple rows and columns. Siu and Chaparro (2014) compared a 3x3 grid and a list interface, when participants completed a set of informational and navigational search tasks. Results revealed that fixation patterns within the grid were not as systematic as fixation patterns in the traditional list layout. Still, in general, the results in the upper left quadrant were fixated earlier than the results in the bottom right quadrant. Moreover, grid users looked at the three search results at the top row for a longer time than at the three search results of the left column, regardless of task type (informational tasks, navigational tasks). However, the order of the search results was not counterbalanced in this study. Therefore, it is unclear whether the longer fixation times on search results in the top row were really due to the position of the search results or whether these were simply the more relevant (or longer, more interesting, etc.) results.

Chen and Pu (2010), who examined a recommender system for product searches, found that as compared to a list interface, in which the products at the top received most attention, in a grid interface in which the products were grouped into four quadrants (according to four recommendation categories), the search results presented in quadrant 2 and

quadrant 3 received longer total fixation times than respective search results in the list interface (i.e., in the middle part of the list). Furthermore, in the grid interface more products were fixated than in the list interface. However, the products were also presented in larger size in the grid interface as compared to the list interface.

Xie et al. (2017) examined users' viewing behaviour in an image search engine that presented the images in a grid format (with 5 rows and 4 to 6 images per row). Other than a bias towards the top results in a list interface, they found a middle-position bias, such that images presented in the middle of the first row were fixated earlier and for a longer time than images presented in the left or the right corner. Scan path analyses further indicated that participants tended to scan the images horizontally rather than vertically or diagonally.

Two other studies compared a list interface with a tabular interface, in which search results are presented in multiple columns. Kammerer and Gerjets (2012) compared a Google-like list interface with an experimental SERP interface that presented search results in a tabular format with three labelled columns, in which search results were grouped according to objective information, subjective information, and commercial information. Concerning participants' gaze behaviour, results showed that users of the tabular interface paid less attention to commercial search results than users of the list interface. Likewise, while users of the list interface paid an equal amount of attention to all kinds of search results, users of the tabular interface paid more attention to objective search results than to search results of the two other categories. Thus, the authors concluded that the tabular interface was successful in guiding users' attention towards objective, that is, scholarly and neutral Internet resources.

Rele and Duchowski (2005) tested a conventional list interface against a tabular interface with four columns that corresponded to the different elements of the search results (title, description, URL, and metadata). Results showed that average fixation duration did not differ significantly across interfaces, but that the number of fixations on the URLs of the search results was significantly lower in the tabular interface than in the list interface (for the

other search result elements no differences were found) and that in the tabular interface due to its vertical arrangement participants tended to inspect the search result elements separately within columns (more transitions within an element-category than in the list interface), rather than moving between columns (i.e., horizontally).

To conclude, these studies provide quite consistent findings on how alternative search interface designs affect users' viewing behaviour. Specifically, as compared to a standard list interface, a grid interface seems to support a rather balanced exploration of all search results, thereby substantially reducing the position bias effect. Furthermore, a tabular interface has the potential to guide users to focus on specific kinds of search results or parts of search results, respectively. However, as a critical remark, it should be noted that all of these studies only examined short-term effects of these novel interface designs.

3.2.3. Effects of search result components or characteristics

Search results on the SERP (also called “snippets”) typically consist of a title, a description (also known as excerpt, respectively) of the content of the Web page, and its URL (uniform resource locator, i.e., the Web address). Several studies investigated how much attention is paid to these different search results components.

Balatsoukas and Ruthven (2012) asked participants to conduct a Web search on a self-chosen topic and analysed participants’ total fixation times and the number of fixations on titles, descriptions, and URLs of search results. No significant differences were found between the three components.

Hautala et al. (2018) examined the probability to which 6th graders inspected the different search results components (title, description, URL) when performing a set of informational fact-finding tasks, with a list of four search results provided for each task (differing in their relevance and trustworthiness, but being comparable in length, specifically in the number of characters). Results indicated that the probability of looking at the different

search result components depended on the relevance and trustworthiness of the search result, and whether or not a relevant and trustworthy result was presented early in the list: When a relevant and trustworthy search result appeared at the top of the list, students were less likely to look at the snippet or URL of the other search results. In contrast, when a search result with an untrustworthy URL address was presented at the top of the list, the snippet and URL components of the other search results were more likely to be looked at. Furthermore, a cluster analysis revealed three different viewing strategies: half of the students looked mainly at the search result titles and descriptions; one-third of the students with high probability examined all components; and one-sixth of the students mainly focused on titles, with the latter group frequently selecting irrelevant or untrustworthy search results.

Rovira, Capdevila, and Marcos (2014) investigated users' viewing behaviour on SERPs of Google Noticias (the Spanish edition of Google News). They focused on the elements of the snippets, i.e., title, source, date, description, and the image. They found that when controlling for the different sizes of the different elements (note that the description was four times larger than the source, and the title twice as large as the source) more visual attention was paid to sources than to descriptions. However, there were no significant differences in the time participants spent viewing sources and titles.

Muntinga and Taylor (2017) examined patients (from hospitals and a general practitioner's office) of a wide age range from 18 to over 60 years, regarding their ability to distinguish genuine, legitimate online pharmacies from rogue online pharmacies, when performing a series of predefined medical search tasks (i.e., navigational, transactional, and informational tasks) with Google. Results indicated that participants who selected genuine pharmacies, on average, had more fixations on both titles and URLs of search results than participants who selected rogue pharmacies. This finding was consistent across different kinds of tasks.

Other studies investigated the effects of particular search result characteristics, such as their length, their relevance, or the number of query terms included in the search results. Cutrell and Guan (2007) examined the effect of the length of the search result description on users' viewing behaviour and of task type (i.e., six navigational vs six informational fact-finding tasks). Specifically, they compared search results with short descriptions (a single line of words), medium-length descriptions (about two to three lines), and long descriptions (six to seven lines of words). Results revealed interaction effects between description length and task type. Total fixation times on descriptions increased with description length, with this effect being pronounced for navigational tasks. Moreover, for informational tasks on SERPs with long descriptions, titles and URLs were fixated shorter than on SERPs with shorter descriptions. This was not the case for navigational tasks.

Further results from Balatsoukas and Ruthven (2012, see above) showed that participants spent more fixations and longer fixations on search result snippets that they judged as not relevant in a post-search interview than on search result snippets that they judged as partially relevant or relevant. However, it is unclear whether snippets were comparable with respect to, for instance, their length or complexity.

In contrast, Oliveira, Aula, and Russell (2009), who examined participants' pupil dilation while inspecting relevant or irrelevant search results (with search results presented individually), found larger pupil sizes on relevant results as compared to irrelevant results (for text-based results in Study 1, and for image results in Study 2). However, again it is unclear whether relevant and irrelevant snippets and images were comparable with regard to, for instance, their contents and their saliency.

Turpin et al. (2006) correlated the number of query terms presented in search results (and search result components, respectively) with participants' total fixation time on the results (or components), based on five fact-finding and five navigational tasks. They found small but significant correlations for search results overall, as well as for all components, that

is, titles, descriptions, and URLs, separately. That is, the more query terms were presented within the results descriptions, the longer were users' fixation times.

Dickerhoof and Smith (2014) examined how much time participants spent on query terms presented in search results as compared to all other words contained in the search results. Participants had to formulate their own queries based on task descriptions that contained two nouns (e.g., oven and food). However, participants were not allowed to use these two words in their queries, so that it was made sure that participants used their own query words. Findings were that across SERPs, 13.5% of all fixations were on query terms, while 86.5% of fixations were on other terms and areas on the page. However, as the authors themselves also argued, it would be more meaningful to break down the rest of the SERP to a word level. This was done by Eickhoff, Dungs, and Tran (2015, Study 1), who indeed found significantly higher average fixation durations on query terms than on other words and also a higher relative number of fixations on the former (also when controlling for term length and complexity). However, the authors did not differentiate between query terms on SERP and on websites in their analyses. Thus, no conclusions can be drawn regarding query term fixation on SERPs only. Based on the fixation data on query terms they predicted which semantic cluster of terms (such as synonyms or antonyms of fixated terms) the user will use in query reformulations.

In summary, the results indicate that the titles of search results are fixated by most users, whereas the fixation time or likelihood to fixate descriptions and URLs seem to depend on the users, the concrete system, and the length of the descriptions, as well as the number of query terms provided. On a critical note, it should be mentioned that the number of words and the font size typically differ between the different search results components as well as between different snippets overall, which is likely also to affect the total fixation time and the number of fixations (as well as the pupil size).

3.2.4. Effects of social annotations

With the rise of social media, individuals increasingly produce and share information themselves. So-called social annotations are a type of metadata that provide information on who (of the user's personal network) shared, endorsed, or produced a particular content, by presenting a profile picture and the name of the sharing contact together with information about when and where the sharing happened. Two studies investigated the effects of such social annotations that were presented in addition to the standard search result components on users' gaze behaviour. Muralidharan, Gyongyi, and Chi (2012, Study 2) found that users increased their attention (i.e., number of fixations) to social annotations added to search results when the search result descriptions were shorter (more fixations for 1-line descriptions than for 4-line descriptions), when the pictures of the people were larger, and when the annotations were placed above the search result description rather than below. The authors conclude that users often seem to simply skip over social annotations and act as if they are not there because they are focused on performing their task and are not used to consider social annotations. Fernquist and Chi (2013) conducted a replication study of the study by Muralidharan et al. (2012). Whereas Muralidharan et al. (2012) used mockup SERPs, Fernquist and Chi (2013) examined the likelihood for looking at social annotations for different types of search topics (e.g., shopping, local services, how-to, news, etc.). They categorized the search questions into three semantic categories (Local/Shopping; How-to/Fact-finding/Navigation; and Entertainment/News), but found only marginally significant differences between the three categories.

To conclude, according to the results of these two studies, users' attention to social annotations seems to depend on the salience of the annotations as well as on a user's concrete information need.

3.2.5. Effects of sponsored results

Search engines' predominant business model is selling advertisements on the SERPs. These so-called *sponsored results* are usually placed at the top or on the right-hand side of the SERP, respectively. *Sponsored results* look similar to the regular search results (so-called organic results; i.e., search results that were not paid for) in that they are composed of the same elements (i.e., title, description, URL) as the organic listings (see Lewandowski et al., 2018).

In the only large-scale study of our review, Lo, Hsieh, and Chiu (2014) examined the viewing behaviour of 451 participants while performing a web search in order to write a product review of a product (e.g., of an iPhone). For their task they were presented with a Yahoo mock-up SERP with two sponsored results shown at the top, two organic results below, and three sponsored results on the right-hand side of the SERP. Results revealed no differences in the number of times they looked at organic results vs at sponsored results shown at the top, which both were looked at more often than sponsored results on the right side of the SERP. However, participants fixated organic results for longer times than both sponsored results shown at the top and on the right-hand sight of the SERP, with the latter being fixated for the shortest time.

Buscher, Dumais, and Cutrell (2010) examined the amount of attention devoted to organic results and sponsored results, depending on the task type (informational or navigational, see above) and the relevance of the sponsored results (high vs low relevance, i.e. sponsored results that matched all query terms of the task at hand vs sponsored results that matched only a subset of the query terms). They found that sponsored results shown at the top received more attention (10.89% of fixations) than sponsored results on the right-hand side, which were mostly neglected by the participants. Moreover, sponsored results shown at the top received significantly more attention when they were of high as compared to low relevance. This also came at the cost of individuals' attention to organic results. Participants

paid significantly less attention to the organic results when high-relevant as compared to low-relevant sponsored results were shown at the top.

In a further analysis of the data of Buscher et al. (2010), Dumais, Buscher, and Cutrell (2010) examined individual differences in gaze patterns on SERPs containing organic results, sponsored results, and related searches (i.e., query suggestions generated by the search engine). They clustered users based on their distribution of attention to different areas of interest (i.e., the extent to which they had fixated sponsored results shown at the top, right-rail sponsored results, organic results, and search boxes). This resulted in three main clusters: The exhaustive searchers (32%), the economic results-searchers (39%), and the economic ads-searchers (29%). Users in the exhaustive cluster as compared to the other two clusters spent significantly more time fixating on organic search results across all positions and had more fixations before the first click. Moreover, they spent proportionally more time on sponsored results shown at the top than economic-results searchers, but proportionally less than economic-ads searchers. Users in the economic-results cluster spent significantly more time on organic results of position 1-3 and positions 7 and below than users of the economics-ads cluster. Finally, users in the economic-ads cluster spent significantly more time on the sponsored results than on lower-ranked organic results (i.e., results 4-6). Moreover, they spent more than twice as much time on the sponsored results shown at the top of the SERP than the economic-results group.

González-Caro and Marcos (2011) also examined organic versus sponsored results (and different task types). Their results showed that the total fixation times on top-listed sponsored results were higher than on side-listed sponsored results for all three types of tasks.

In summary, the studies show that advertisements in the form of sponsored links on SERPs receive visual attention (particularly those presented at the top), although they are often not clicked on. This indicates that users make an informed decision on whether to click on ads. Furthermore, the reported research shows that sponsored results shown at the top of

the SERPs receive more attention than sponsored results on the right-hand side, which rather seem to be ignored by most users. One should keep in mind, however, that in the studies reported, predominantly students were investigated. This user group might be especially ads-aware, and therefore, the results may not hold for a larger population.

3.2.6. Effects of vertical results (Universal Search)

In addition to organic search results and sponsored results, so-called vertical results (or Universal Search results) are integrated into the ranked list of organic search results. According to Lewandowski et al. (2018, p. 421), "*Universal Search results* are results generated from vertical search engine indexes, such as news, video, or images. Depending on the nature of the index, these results can either be generated similarly to organic results (as in the case of images) or be based on a certain collection of sources (as in the case of news, where a collection of trusted news sources is defined beforehand by the search engine vendor). Universal Search results can also come from document collections especially built by the search engine vendor (as opposed to the results from the web index that come from a multitude of sources distributed across the web)."

In a study by Liu, Liu, Zhou, Zhang, and Ma (2015), the influence of relevant as well as irrelevant vertical results on users' visual attention was tested. They found an *attraction bias*, i.e., a strong bias towards more attention to the vertical results when the vertical results were relevant. This means that when a relevant vertical result was shown, it attracted more attention than an organic result. When a vertical result had strong visuals in the snippet (e.g., image-only, application, or news verticals), users looked at this result immediately, which means they attended to the first organic result later than on SERPs that only have organic results. Concerning the interplay between vertical results and regular organic results, Liu et al. further found that when the vertical results were relevant, users paid less attention to the

organic results listed after the vertical result. In line with this finding, irrelevant vertical results increased the attention for organic results ("spill-over effect").

Marcos, Gavin, and Arapakis (2015) compared SERPs with plain textual search results with SERPs that also contained vertical search results, such as multimedia elements (from Google Images or Google Videos), recommendation search results (e.g., GooglePlus, Google Shopping), and geo-location search results (Google Places). In general, no differences between enriched snippets and plain textual snippets were found in terms of gaze behaviour. In case of both enriched snippets and plain textual snippets, top-ranked snippets were fixated earlier and received longer fixation times and more fixations than bottom-ranked snippets.

Lagun, Hsieh, Webster, and Navalpakkam (2014) examined the presence of relevant as well as irrelevant knowledge graphs (i.e., collections of basic facts describing an entity, shown in a separate box to the right of the actual search results list on the SERP) and instant answers (that are directly supplied by a search engine in response to a query without reference to a website, such as weather information or factual information) in SERPs presented on a mobile phone for a series of fact-finding and navigational tasks. However, results showed no significant effects of knowledge graph relevance or instant answer relevance on participants' total fixation times on search results.

In summary, results regarding the effects of vertical results on users' viewing behaviour are mixed, indicating that the effect depends on the type of vertical and also on the relevance of the vertical results, and maybe also on the device type.

3.2.7. Effects of device type: Differences between desktop and mobile searches

Apart from desktop PCs and laptops, nowadays, individuals increasingly use mobile devices like smartphones and tablets to search for information on the Web. Search engine companies have reacted to this by adopting a "mobile-first" strategy with respect to their algorithms. Accordingly, a few studies examined users' viewing behaviour on SERPs displayed on mobile

devices (Djamasbi, Hall-Phillips, & Yang, 2013a; Djamasbi, Hall-Phillips, & Yang, 2013b Domachowski, Griesbaum, & Heuwing, 2016; Lagun et al., 2014). However, we did not come across any study that examined both users' viewing behaviour on SERPs displayed on a mobile device and a computer screen and compared these two conditions statistically (i.e., by reporting tests for statistical significance regarding the effects of device type in the paper). Yet, one study by Kim et al. (2015) statistically compared users' viewing behaviour on SERPs that were displayed on a small screen (i.e., a browser limited to a 320 × 480 pixel window showing three search results, mimicking the screen size of smartphones) versus on a large screen (i.e., a regular computer screen with a resolution of 1280 × 1024 pixels showing ten search results). Both types of SERPs, however, were presented on a 17" desktop PC monitor. For each screen size, participants completed five simple fact-finding tasks and five navigational tasks. Results showed that participants fixated the search results significantly longer on the small screen than on the large screen. Additional analyses revealed that this difference was due to longer fixation times on the first search result. Moreover, scan path analyses indicated that as compared to the large screen users on the small screen viewed the search results more linearly from top to bottom and with fewer skips and regressions or more frequently made a selection after looking at only a single search result, respectively. Besides, the authors analysed the relationship between total fixation time per search result and search speed (i.e., the time taken until the first click), showing similar positive relationships in both screen conditions.

To conclude, the results of this study indicate that in mobile SERPs, the first search result might receive even more attention than in SERPs presented on a laptop or desktop PC. However, as the authors themselves critically discuss, the study only examined the influence of the screen size without considering other differences such as the fact that mobile phones are held in hand and used via touch interactions. Thus, differences between mobile devices and desktop PCs might be underestimated in their study. Moreover, the screen size might

matter more when searching for more complex search topics that require comparing and integrating information from various websites rather than finding a single correct answer.

3.3. Individual factors

Several studies examined what role individual factors, that is, personal characteristics and (cognitive) prerequisites of the users, play in users' viewing behaviour on SERPs, such as their age, their prior knowledge or personal beliefs, or their gender.

3.3.1. The role of age

Two studies compared the search behaviour of students of different grade levels. Bilal and Gwizdka (2016) examined the effects of grade level (five 6th grade students vs eight 8th grade students) on SERP viewing behaviour when performing one fact-finding and one more complex informational search tasks. They differentiated between reading and scanning states. Results showed that eighth-graders had significantly more reading fixations on the SERPs than sixth-grade students, but there were no differences in respective total fixation times. Furthermore, eighth-graders initially looked more at the top search results of the SERP, whereas sixth-graders first looked more at lower ranks.

Dinet, Bastien, and Kitajima (2010) analysed SERP viewing behaviour of fifth-, seventh-, ninth-, and eleventh-grade students performing simple fact-finding tasks (history, geography). Their results indicated that the typographical cueing of boldfaced search words resulted in longer total fixation times on the search results, especially for fifth-grade students (Dinet et al., 2010). Older students were attracted by such cues only when the search task addressed unfamiliar topics, but not for familiar topics. Furthermore, the authors identified four different clusters of viewing strategies: 1) concentration on the upper part of the SERP, 2) exhaustive reading of the entire SERP, 3) visual jumps concentrated on the keywords, and 4) concentration on the lower part of the SERP. Whereas for the younger students, the third

strategy was predominant, for the older students, the exhaustive strategy was the predominant strategy, particularly for familiar topics.

To conclude, with increasing age, students seem to engage in a more systematic and sometimes also more intensive reading of the SERPs. These findings are corroborated by other studies that examined differences in viewing behaviour on SERPs between adults and children (Gossen, Höbel, & Nürnberger, 2014b, 2014a). However, these studies based their analyses on heatmaps rather than on statistical comparisons or only reported descriptive results. Therefore, they did not fulfil our inclusion criteria (see above).

3.3.2. The role of prior domain knowledge

Other studies examined the effects of prior domain knowledge on SERP viewing behaviour. Brand-Gruwel et al. (2017) compared the viewing behaviour of domain experts (psychology lecturers) and domain novices (first-year students) while performing two informational search tasks on two complex scientific issues from the field of psychology. Results revealed that domain experts fixated more search results than novices before a search result was clicked. Similarly, Kammerer and Gerjets (2013), who examined university students who were asked to conduct a web search on a complex health-related issue, found that the higher participants' self-reported prior domain knowledge, the longer they visually inspected the search results.

To conclude, similar to the results regarding age, increased prior knowledge about the search topic at hand (and also a sceptical stance towards the accuracy of knowledge on the Web) also seems to be related to a more systematic and more careful reading of the SERPs in order to decide which information sources to access.

3.3.3. Other user aspects

As an indication of carefully evaluating search results in terms of source information, Kammerer, Bråten, Gerjets, and Strømsø (2013) examined university students' total fixation times on the URLs of the search results while performing a Web search task on a

controversial medical issue. Specifically, the authors were interested in the role of individuals' beliefs whether or not the Web is a reliable resource of accurate academic knowledge in their thorough inspection of URLs. Results indicated that the more sceptical participants were about the Web being a reliable resource of accurate academic knowledge, the longer they fixated on the URLs of the search results.

Lorigo, Pan, Hembrooke, Joachims, and Granka (2006) examined the effects of gender on SERP viewing behaviour while performing a series of informational fact-finding and navigational tasks. Males were more likely to visually inspect search results further down in the result list (position 7-10), had more linear viewing sequences from top to bottom, and made fewer regressions to already inspected search results than female participants. Females, on the contrary, were more likely to visually inspect search results in the top of the list (i.e., positions 2 and 3 in the SERP). However, this difference might be the result of other (unknown) differences between male and female participants of this study. The large-scale study by Lo et al. (2014) (also see above), did not reveal any differences in viewing behaviour between male and female participants.

Finally, similar to Brand-Gruwel et al. (2017), Aula, Majaranta, and Räihä (2005) also examined the number of search results being fixated before performing the first action, when performing a series of fact-finding tasks in Google. Based on these data, the authors classified participants into exhaustive and economic evaluators of search results. Results showed that exhaustive evaluators (i.e., users who typically fixated more than half or even all of the search results) had longer average fixation durations on SERPs than economic evaluators (i.e., users who fixated only the first few search results). Unsurprisingly, the economic evaluation strategy was beneficial when one of the first search results was relevant for solving the task. In such cases, task times of economic evaluators were significantly shorter than those of exhaustive evaluators. To conclude, this finding once again illustrates that many effects depend on the concrete search results being presented in the SERPs.

4. DISCUSSION

4.1. Summary of findings

To summarize the findings of our narrative literature review comprising 41 papers, first, concerning the effects of *contextual factors*, neither the type of task (informational fact-finding vs informational complex task vs navigational task vs transactional task) nor the use of thinking-aloud instructions seem to have systematic effects on users' viewing behaviour on SERPs. A general problem with the investigation of effects of task types, however, is that findings might be heavily influenced by the concrete topics used for the different task types, which might differ, for instance, in their difficulty or interestingness, or regarding participants' familiarity with these topics, as well as by the concrete search results provided by the search engine for a particular topic (and particular queries entered). Thus, the search results can differ in their length, relevance, number of bold terms included, etc., which are all factors that can affect participants' viewing behaviour above and beyond the influence of the type of task.

Concerning instructions to evaluate the search results or to verbalize one's thoughts during task processing, there are first indications that such instructions might result in a more linear viewing behaviour from top to bottom. Yet, future research is needed to corroborate these findings, as are studies that examine the influence of other contextual factors, such as the presence or absence of time constraints during the search, the perceived personal relevance of the task, or the situation or setting in which the Web search is conducted (e.g., in the lab, at school or in the workplace, or at home, or alone vs collaboratively).

Second, with regard to effects of *resource factors* on users' viewing behaviour on SERPs, which according to our literature review are the most studied type of factors, two quite consistent findings across several studies are (1) that when search results are presented in the form of lists, search results at the top of the list receive most attention (i.e., the position

bias effect), and (2) that alternative SERP interfaces that present search results in a grid layout instead might support a rather balanced exploration of all search results. However, on a critical note, it should be mentioned that the latter effect might be due to users' unfamiliarity with grid interfaces. Thus, the effect might disappear with increased usage. Besides, tabular interfaces, that is, another type of alternative SERP interface, seem to have the potential to guide users to focus on specific kinds of search results or parts of search results, respectively. However, only two studies examined the effects of tabular as compared to list interfaces.

When studying attention to the different search results components (i.e., title, description, and URL of a search results), findings across studies indicate that most users fixate the titles of search results, whereas the fixation time on or fixation likelihood of descriptions and URLs seem to depend on the individual users, the concrete system, and the length of the summaries, as well as the number of query terms provided. Moreover, a direct comparison between search result components seems difficult to us, since the number and difficulty or relevance of words, the font size and colour, as well as the number of bold keywords typically differ between the different search results components, which are likely also to affect the total fixation time and the number of fixations directed to the components. With regard to additional metadata information provided for the search results, (e.g., social annotations), or other types of search results provided on the SERP (e.g., sponsored results or vertical results), findings indicate that users' attention to these elements seems to depend on their salience and position (e.g., elements at the top are likely to receive more attention), but also on their relevance for the search task at hand as well as on the relevance of other regular (i.e., organic) search results.

Furthermore, with regard to the effects of the device type (or screen size, respectively), the only existing study indicates that in SERPs displayed on small screens (cf. on mobile devices) the first search result might receive even more attention than in SERPs displayed on

larger screens (cf. on laptops or desktop PCs). However, further research is needed to corroborate this finding.

Third, with regard to effects of *individual factors* on users' viewing behaviour on SERPs, both age and prior topic knowledge (as well as a sceptical stance towards the accuracy of knowledge on the Web) seem to be related to a more systematic and more careful reading of the SERPs in order to decide which information sources to access. Future research is needed that examines other cognitive factors such as reading skills or working memory capacity as well as motivational factors, such as personal interest in the topic or perceived relevance or importance of the task.

Furthermore, the majority of the reviewed studies comes with certain limitations that we will address in the following.

4.2. Theoretical considerations

The majority of the reviewed studies has used a largely data-driven rather than theory-driven research approach. A prominent theory to explain Web users' evaluation and selection of search results is the Information Foraging Theory by Pirolli (2007). According to this theory, the selection of search results is determined by the strength of a so-called "information scent". Information scent reflects the semantic similarity that a user perceives between information contained in link descriptions (i.e., proximal cues) and his or her search goal or information need (i.e., the desired information). A strong information scent indicates a high likelihood that the respective information source contains the desired information and thus increases the likelihood that the link will be selected. Thus, (seemingly) more relevant search results have a higher likelihood to be selected, as users make predictive relevance judgments based on the snippets (Rieh, 2002). Accordingly, more relevant search results, that is, search results with a higher information scent, should also be inspected in greater detail. Several findings reviewed

above (e.g. Buscher et al., 2010; Liu et al., 2015; Pan et al., 2007; Schultheiß et al., 2018) support this assumption.

A second assumption of the Information Foraging Theory (Pirolli, 2007) in line with the concept of bounded rationality (cf. Simon, 1955) is that Internet users apply a satisficing strategy (i.e., a blend of the words satisfy and suffice) that implies that they do not evaluate the relevance of all search results available, but sequentially evaluate them only until one is encountered that is "good enough". This suggests that the position of a search result in a SERP will also affect the likelihood that it will be selected. This is in line with the effects of the position or the order of search results in SERPs on users' viewing behaviour, as it has been typically found in the studies reported above (e.g., Cutrell & Guan, 2007; Granka et al., 2004; Guan & Cutrell, 2007; Pan et al., 2007; Thomas et al., 2013; Turpin et al., 2006; Walhout et al., 2017). However, one should keep in mind that the decision to stop searching for more information (i.e., selecting more results) presumably is strongly influenced by the task at hand and the user's motivation, as indicated by findings from transaction log analyses. Session length has been shown to vary greatly (e.g., Jansen, Spink, Blakely, & Koshman, 2007). For example, search sessions in medical tasks are considerably longer on average than in non-medical tasks (White & Horvitz, 2009).

Moreover, as proposed by Brumby and Howes (2008) in their interactive search framework (in the context of website menus), the evaluation of an individual item (i.e., in our case a search result) also depends on the similarity or distinctiveness of this item to other items (i.e., search results) encountered (see also Hautala et al., 2018). Accordingly, if a search result is sufficiently distinct (i.e., more salient or relevant than others) it might be selected without others being considered. This might even be the case for a search result that is positioned further down in the list. In contrast, if none of the search results stands out, according to the assumptions of the interactive search framework individuals should inspect a larger number of results. Similarly, according to the Prominence-Interpretation-Theory (Fogg,

2003) that addresses credibility evaluations on Websites (and that was adapted to evaluations of search results by Kammerer & Gerjets, 2014) assumes that it requires two constituent components for evaluations being made, that is, prominence and interpretation. In the context of evaluations of search results, prominence refers to the degree of saliency of cues in search results indicating the relevance or credibility of the Website (Kammerer & Gerjets, 2014). The more salient a cue is, the higher the likelihood to be noticed. Interpretation refers to an individual's personal interpretation of a cue that he or she has noticed. The quality of this interpretation is likely to depend on a user's prior knowledge and skills (cf. Fogg, 2003).

Information Foraging Theory, the interactive search framework, Prominence-Interpretation-Theory, as well as other theoretical approaches to users' interactions with information systems (cf. White, 2016) should be considered when formulating research questions and hypotheses in future research. Without such foundation, researchers might fail to interpret and contextualize their findings.

4.3. Stimulus material

As mentioned in the section on descriptive statistics, the majority of studies examined in our review used fabricated SERPs, most of the time presenting purely textual organic results only. Given that today's SERPs comprise many different elements (cf. Azzopardi, Thomas, & Craswell, 2018; Lewandowski et al., 2018), such as (a) conventional organic results, (b) sponsored results, (c) rich snippets (organic results with additional data such as rating stars or geo-location data, (d) vertical results from specialized search engines such as Google Video, Google Images, or Google News, (e) instant answers and featured snippets that are presented at the top of the organic results in a box and provide a direct answer to a user's question (with featured results references to a website, whereas instant answers do not), and (f) knowledge graphs presented as a box to the right of the list of search results, this is no longer a realistic study setting. While we are aware that stripped-down versions of SERPs allow for a larger

degree of experimental control, there still is a lack of studies using (more) realistic SERPs, as presented by contemporary commercial search engines. So far, only a few studies have included additional elements (e.g., ads, vertical results, knowledge graph results, see Section 3.2). Therefore, future research should examine the effects of these elements in more detail. Moreover, these elements should not only be considered separately but also in conjunction. This might be achieved by examining a research question both with an experimental design using fabricated SERPS and a complementary study using a setting with realistic SERPs.

4.4. Sample sizes

As already mentioned, sample sizes of the studies reviewed are rather small, with a median of 28 participants, and a mean of 41.5 participants. This does not seem to be specific to eye-tracking research on SERP viewing behaviour but to eye-tracking studies in general. For instance, Lund (2016) reported similar numbers for eye-tracking studies in the field of library and information science (mean: 34.21 participants, ranging from 5 to 96).

The low sample sizes might have led to a large proportion of non-significant results, but also might have increased the likelihood for Type II errors (i.e., that a found significant difference was a random effect or artefact). In other words, low sample sizes are a threat to the validity of the results. For future research, therefore, we encourage researchers to increase the sample sizes of their studies. Tools for calculating required sample sizes are readily available (e.g., G*Power[1]).

4.5. Threats to validity of eye-tracking measures related to SERPs

[1] http://www.gpower.hhu.de

Apart from the general problem of low sample sizes, a threat to the validity of studies using eye-tracking methodology in particular, is a lack of comparability of the stimuli that are compared against each other (Orquin & Holmqvist, 2018). This is also a central issue for eye-tracking research on SERP viewing behaviour, for instance, when comparing total fixation times of different search results. Search results typically vary in their content, length of text, visual characteristics, etc. Thus, it is challenging to interpret differences found in the eye-tracking data between search results (e.g., that some search results are fixated more or longer than others) as effects of the search result position or the type of search task, respectively, rather than as effects of, for instance, the relevance or comprehensibility of the different search results. In order to draw valid conclusions on the effects of search results positions, for instance, search results need to be counterbalanced across positions (cf. Guan & Cutrell, 2007), or different orders of search results need to be systematically varied between participants (cf. Kammerer & Gerjets, 2014; Pan et al., 2007; Schultheiß et al., 2018). However, when comparing different types of search tasks (and thus different topics and completely different search results), a controlled comparison seems difficult to accomplish, unless the search results are systematically designed with a comparable number and difficulty of words, comparable relevance, etc. Instead, another possibility to examine effects of task types would be to examine each task type with a large set of search topics and search results.

Another threat to the validity of eye-tracking results (Orquin & Holmqvist, 2018) is a too small size of the examined areas of interest (AOIs) and a small distance between them. Due to the limited accuracy and precision of eye-tracking systems, “fixations are never measured at their true location” (Orquin & Holmqvist, 2018, p. 1647), but with a certain offset. Thus, if stimulus objects are too small and too densely spaced, the risk of fixations falling outside the AOIs (i.e., false negatives) or being assigned to the wrong AOIs (i.e., false positives) increases, which would result in inaccurate (noisy) or false research findings. According to Orquin and Holmqvist (2018), even with today’s high precision eye trackers, a

minimum stimulus size of about 3 degrees of visual angle is required to achieve an accuracy of approximately 80%. However, many elements on real SERPs are smaller or more densely spaced than 3 degrees of visual angle. As a simple rule of thumb, the visual angle of the width of the thumb held at arm's length is about 2 degrees (cf. O'Shea, 1991). Moreover, frequently crucial information about characteristic of the study materials, about the average precision and accuracy of the eye-tracker used, and about the underlying fixation algorithm are missing from the papers examined. We, therefore, argue for authors to follow eye-tracking reporting standards (e.g., Fiedler, Schulte-Mecklenbeck, Orquin, & Renkewitz, 2019) to ensure that other researchers can replicate their work and also better assess the validity of results achieved.

4.6. Need for data triangulation

Most of the studies analysed time-based and/or count-based fixation measures for particular search results, sometimes complemented by analyses of viewing sequences across search results. However, one should keep in mind that it is often difficult to unambiguously interpret eye-tracking data taken by itself, such as, for what reasons an individual looks at presented stimuli for a certain amount of time or in a particular order (cf. Scheiter & van Gog, 2009). Therefore, methodological triangulation, that is, the combination with other types of data sources, such as, verbal protocols, logfile data, or outcome variables, is important to increase the external validity of the findings (cf. Orquin & Holmqvist, 2018; Scheiter & van Gog, 2009). In case of viewing behaviour on SERPs, for example, it can be examined, how total fixation times on particular search results are related to their selection (i.e., clicks), and to users' reasons to click or not to click on particular search results (as indicated in verbal protocols). While many of the reviewed papers in addition to eye-tracking also used other measures, such as, click data and outcome data, results from different data sources were only presented sequentially rather than in an integrated way. Therefore, we suggest future research

to analyse to what extent different process measures correlate (e.g., fixation data, click data, and verbal data) or to what extent they predict search success or degree of knowledge acquisition (i.e., outcome variables), respectively, when considered as predictors in a multiple regression analysis.

4.7. Suggestions for future research

As can be seen from our review, not too much research has been conducted on factors affecting viewing behaviour on SERPs yet. First of all, this means that more research on all factors examined is more than welcome to strengthen the empirical basis and the evidential value of this line of research. However, researchers should make clear why they used eye-tracking and not other methodologies that could lead to comparable results. As found in our review, eye-tracking is often used as an additional method in laboratory settings, without making explicit what the eye-tracking data adds that could not be found out using other methods, either in or beyond the lab. For instance, large-scale click-through data of real users' natural information searches in many cases might produce more meaningful results than lab studies due to the usually much larger datasets used in this type of research (e.g., Goel, Broder, Gabrilovich, & Pang, 2010; Jansen & Spink, 2006; Wang et al., 2013). Furthermore, mouse cursor data might be used to complement or replace eye-tracking data, both with the aim of predicting users' relevance judgments or using the data for relevance feedback (e.g., Guo & Agichtein, 2010; Lagun & Agichtein, 2011), which is a vivid area of research in interactive information retrieval (White, 2016, p. 41ff.). For instance, Huang, White, and Dumais (2011), who examined correlations between eye-tracking data and mouse cursor position, found that the cursor position can be used to predict eye gaze. Therefore, they recommend using mouse-cursor data, which is much cheaper to acquire and easier to collect at large scale than eye-tracking data. These results are corroborated by a study by Liu et al. (2014). However, there is also some indication that eye-tracking data is richer than data from

other sources. Eickhoff et al (2015, Study 2) conducted a study (with 137 participants) using mouse-cursor data (relative number of hovers and time for which the cursor rests on a term) instead of eye-tracking data, revealing comparable results. Yet, the authors note that when comparing the findings obtained from eye-tracking data and from cursor-data, "it becomes obvious, that fixations are richer and more accurate predictors of user attention than cursor traces. The majority of users only occasionally use the mouse cursor in order to highlight text, mark their current reading position or follow textual hyperlinks." Nevertheless, researchers should explicitly point out why eye-tracking data has been used and how it contributes to answering the research questions. For example, eye-tracking data also provides insights into cognitive processes that do not lead to overt actions (e.g., when deciding not to click on a particular hyperlink). Furthermore, some SERPs do not lead to users' clicks at all. There are various reasons for this, including (1) users' not being satisfied with the results or (2) users finding the desired information already on the SERP. In the latter case, the required information can be included, e.g., in the description of organic search results, in the knowledge graph, or in instant answers or featured snippets (see above). For these SERPs where no click-interaction can be recorded, eye-tracking together with verbal protocols can provide important insights into users' behaviour on SERPs.

Second, some of the methodological weaknesses of previous studies revealed in our review lead to suggestions for future research. As nearly all studies were small-scale and, therefore, their evidential value has to be questioned, we recommend to increase sample sizes and sample diversity. Probably most of the reviewed studies are underpowered. Thus, we recommend conducting power analyses already when planning a study. An area where the problem of homogeneous samples is particularly evident is when studying effects of sponsored results. As results from lab studies using student samples often contradict findings from quantitative studies conducted outside the lab conducted with more diverse samples, we highly recommend diversifying samples in this area. As outlined above, further

methodological issues of many studies are (1) problems related to the fact that stimuli that are compared to each other often lack comparability, thus, confounding the results, (2) problems resulting from the small size of the examined areas of interest (AOIs) and/or the small distance between them, and (3) problems resulting from the use of stripped-down SERPs limiting ecological validity. Future studies should aim at overcoming these problems.

Third, our review identified a need for more a priori hypothesis-based and theory-driven research. To make results on users' viewing behaviours on SERPs more meaningful, we highly recommend basing future eye-tracking studies on theory-driven hypotheses. For instance, when an effect of an independent variable on some of multiple dependent variables is found, authors often cannot explain why the effect held for these particular variables but not for others. If such studies were properly grounded in theory, researcher would be better able to predict in advance which effects should be likely and which variable should not be affected by the independent variable.

Following these three recommendations would already result in a robust body of research and would increase evidential value. Nevertheless, from our results, we can also derive some areas where we think further research is particularly fruitful.

Regarding research on contextual factors, we found that effects of task types might be influenced by the concrete topics, by participants' familiarity with these topics, and by the concrete individual search results provided by the search engine. Future research should take measures to control for these factors, e.g., through using standardized tasks/topics or reusing search tasks from prior studies. Researchers should make clear how findings from their studies can be generalized to all search tasks having the same attributed (see the Project "Systematic Review of Assigned Search Tasks", which also provides a "Repository of

Assigned Search Tasks (RepAST)"[2]) Also, further studies should examine the influence of other contextual factors, such as the presence or absence of time constraints during the search, the personal relevance of the task, or the situation in which the Web search is conducted, e.g., in the lab, at school or in the workplace, or at home).

Regarding research on resource factors, more research on alternative SERP interface designs and the more complex designs of current commercial search engines is needed. As for alternative designs, only short-term effects have been tested; we recommend conducting studies examining longer-term effects, as novelty may play a huge part in the effects already examined. Regarding search results components, future research should consider that the number of words and the font size typically differ between the different search results components, which is likely also to affect the total fixation time and the number of fixations directed to the components. A solution to this problem would be to control for the number of words and font sizes, which would, however, result in less realistic SERPs.

In our review, we had only one study systematically measuring the effect of device type, in this case, desktop vs mobile. Given the wide variety of devices used today, we highly recommend conducting further studies on the effect of device type on users' SERP viewing behaviour. This should not only include smartphones of varying screen sizes but also tablets.

Regarding individual factors, the role of age in users' viewing behaviour on SERPs has not been researched extensively yet. This again may have to do with the samples usually used in eye-tracking studies. However, as the research conducted so far suggests an effect of age, we see considering this factor thoroughly as a fruitful way for future research. The same holds for the role of prior domain knowledge about the search topic at hand. Finally, future research could also consider motivational and emotional aspects of the searchers, such as, to

[2] https://ils.unc.edu/searchtasks/index.php

what extent and how individuals' personal attitudes or feelings towards the search topic affect their SERP viewing behaviours.

5. CONCLUSION

There is a large interest in using eye-tracking to study users' viewing behaviour on SERPs. The present paper provides a comprehensive narrative literature review of the respective body of research comprising 41 papers that examined influencing factors on users' SERP viewing behaviour. To synthesize the findings of these papers, factors influencing users' viewing behaviour were grouped into contextual factors, resource factors, and individual factors (cf. Lazonder & Rouet, 2008).

The most important finding related to *contextual factors* is that task type (informational fact-finding vs informational complex task vs navigational task vs transactional task) does not seem to have a systematic effect on users' viewing behaviour, or that it is at least difficult to systematically investigate such task effects. Regarding *resource factors*, the most crucial overall finding still is that when search results are presented in the form of lists, search results at the top of the list receive most attention. Concerning effects of *individual factors* on users' viewing behaviour on SERPs, both age and prior topic knowledge seem to be related to a more systematic and more careful reading of the SERPs.

Through reviewing the available research, we identified three major limitations of the published studies: Firstly, nearly all studies are small-scale and, therefore, their evidential value has to be questioned. Secondly, stimuli that are compared to each other often lack comparability. Thirdly, problems result from the small size of the examined areas of interest (AOIs) and/or the small distance between them.

To conclude, a fruitful way forward for examining users' viewing behaviour through eye-tracking would be (1) to conduct larger and more theory-driven studies, (2) to couple controlled eye-tracking laboratory studies using fabricated (i.e., stripped-down) SERPs with

data from more realistic settings, by having the same participants conduct both web searches in an experimental setting with fabricated materials and in a realistic setting using the open Web, and (3) to triangulate eye-tracking results with data obtained from other data sources, such as clickthrough data.

ACKNOWLEDGEMENTS

REFERENCES

Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers and Education*, *125*, 413–428. https://doi.org/10.1016/j.compedu.2018.06.023

Athukorala, K., Głowacka, D., Jacucci, G., Oulasvirta, A., & Vreeken, J. (2015). Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology*, *14*(4). https://doi.org/10.1002/asi.23617

Aula, A., Majaranta, P., & Räihä, K. J. (2005). Eye-tracking reveals the personal styles for search result evaluation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *3585 LNCS*, 1058–1061. https://doi.org/10.1007/11555261_104

Azzopardi, L., Thomas, P., & Craswell, N. (2018). Measuring the utility of search engine result pages. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18* (pp. 605–614). New York, New York, USA: ACM Press. https://doi.org/10.1145/3209978.3210027

Balatsoukas, P., & Ruthven, I. (2012). An eye-tracking approach to the analysis of relevance judgments on the Web: the case of Google search engine. *Journal of the American Society for Information Science and Technology*, *63*(9), 1728–1746. https://doi.org/10.1002/asi

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Bilal, D., & Gwizdka, J. (2016). Children's eye-fixations on google search results. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–6.

https://doi.org/10.1002/pra2.2016.14505301089

Bradley, M. B., Miccoli, L. M., Escrig, M. a, & Lang, P. J. (2008). The pupil as a measure of emotional arousal and automatic activation. *Psychophysiology*, *45*(4), 602. https://doi.org/10.1111/j.1469-8986.2008.00654.x.The

Brand-Gruwel, S., Kammerer, Y., van Meeuwen, L., & van Gog, T. (2017). Source evaluation of domain experts and novices during Web search. *Journal of Computer Assisted Learning*, *33*(3), 234–251. https://doi.org/10.1111/jcal.12162

Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum*, *36*(2), 3–10. https://doi.org/10.1145/792550.792552

Buscher, G., Dumais, S. T., & Cutrell, E. (2010). The good, the bad, and the random: an eye-tracking study of ad quality in web search. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 42–49). ACM. Retrieved from http://portal.acm.org/citation.cfm?id=1835459&dl=ACM

Chen, L., & Pu, P. (2010). Eye-tracking study of user behavior in recommender interfaces. In D. De Bra, P.; Kobsa, A.; Chin (Ed.), *UMAP 2010* (pp. 375–380). Berlin, Heidelberg: Springer.

Cutrell, E., & Guan, Z. (2007). What are you looking for? An eye-tracking study of information usage in Web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007)* (pp. 407–416). New York: ACM.

Dickerhoof, A., & Smith, C. L. (2014). Looking for query terms on search engine results pages. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–5. https://doi.org/10.1002/meet.2014.14505101156

Dinet, J., Bastien, J. M. C., & Kitajima, M. (2010). What, where and how are young people looking for in a search engine results page? In *Conference Internationale Francophone sur I'Interaction Homme-Machine on - IHM '10* (p. 105). New York, New York, USA: ACM Press. https://doi.org/10.1145/1941007.1941022

Djamasbi, S., Hall-Phillips, A., & Yang, R. (2013a). SERPs and ads on mobile devices: an eye tracking study for Generation Y. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8010 LNCS, pp. 259–268). User experience and decision making research laboratory, Worcester Polytechnic Institute, United States. https://doi.org/10.1007/978-3-642-39191-0_29

Djamasbi, S., Hall-Phillips, A., & Yang, R. R. (2013b). SERPs and ads on mobile devices: An eye tracking study for generation Y. In *International Conference on Universal Access in Human-Computer Interaction* (pp. 259-268). Springer, Berlin, Heidelberg.

Domachowski, A., Griesbaum, J., & Heuwing, B. (2016). Perception and effectiveness of search advertising on smartphones. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–10. https://doi.org/10.1002/pra2.2016.14505301074

Dumais, S. T., Buscher, G., & Cutrell, E. (2010). Individual differences in gaze patterns for web search. In *Proceeding of the third symposium on Information interaction in context - IIiX '10* (p. 185). New York, New York, USA: ACM Press. https://doi.org/10.1145/1840784.1840812

Eickhoff, C., Dungs, S., & Tran, V. (2015). An eye-tracking study of query reformulation. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15*, (2), 13–22. https://doi.org/10.1145/2766462.2767703

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal report as data*. Cambridge, Massachusetts, London, England: MIT Press.

Fernquist, J., & Chi, E. H. (2013). Perception and understanding of social annotations in Web search. In *WWW '13 Proceedings of the 22nd international conference on World Wide Web* (pp. 403–412). New York. https://doi.org/10.1145/2488388.2488424

Fiedler, S., Schulte-Mecklenbeck, M., Renkewitz, F., & Orquin, J. L. (2019). Increasing

reproducibility of eyetracking studies: The EyeGuidelines. In M. Schulte-Mecklenbeck, A. Kühberger, & J. Johnson (Eds.), *A handbook of process tracing methods: 2nd Edition*. New York: Routledge.

Fogg, B. J. (2003). Prominence-interpretation theory. In *CHI '03 extended abstracts on Human factors in computing systems - CHI '03* (p. 722). New York, New York, USA: ACM Press. https://doi.org/10.1145/765891.765951

Gerjets, P., Kammerer, Y., & Werner, B. (2011). Measuring spontaneous and instructed evaluation processes during Web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. *Learning and Instruction*, *21*(2), 220–231. https://doi.org/10.1016/j.learninstruc.2010.02.005

Goel, S., Broder, A., Gabrilovich, E., & Pang, B. (2010). Anatomy of the long tail. In B. D. Davison, T. Suel, N. Craswell, & B. Liu (Eds.), *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10* (p. 201). New York, New York, USA: ACM Press. https://doi.org/10.1145/1718487.1718513

González-Caro, C., & Marcos, M.-C. (2011). Different users and intents: An eye-tracking analysis of web search. In *Proceedings of Web Search and Data Mining (WSDM 2011)* (pp. 9-12). New York: ACM.

Gossen, T., Höbel, J., & Nürnberger, A. (2014a). A comparative study about children's and adults' perception of targeted web search engines. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (pp. 1821–1824). New York, New York, USA: ACM Press. https://doi.org/10.1145/2556288.2557031

Gossen, T., Höbel, J., & Nürnberger, A. (2014b). Usability and perception of young users and adults on targeted web search engines. In *Proceedings of the 5th Information Interaction in Context Symposium on - IIiX '14* (pp. 18–27). New York, New York, USA: ACM Press. https://doi.org/10.1145/2637002.2637007

Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in

WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 478–479). New York: ACM.

Granka, L., Feusner, M., & Lorigo, L. (2008). Eye monitoring in online search. In *Passive eye monitoring* (pp. 347–372). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-75412-1_16

Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07* (p. 417). New York, New York, USA: ACM Press. https://doi.org/10.1145/1240624.1240691

Guo, Q., & Agichtein, E. (2010). Towards predicting web searcher gaze position from mouse movements. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 3601-3606). New York: ACM.

Hautala, J., Kiili, C., Kammerer, Y., Loberg, O., Hokkanen, S., & Leppänen, P. H. T. (2018). Sixth graders' evaluation strategies when reading Internet search results: an eye-tracking study. *Behaviour and Information Technology*, *37*(8), 761–773. https://doi.org/10.1080/0144929X.2018.1477992

Jansen, B.J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, *42*(1), 248–263.

Jansen, Bernard J., Spink, A., Blakely, C., & Koshman, S. (2007). Defining a session on Web search engines. *Journal of the American Society for Information Science and Technology*, *58*(6), 862–871. https://doi.org/10.1002/asi.20564

Jiang, J., He, D., & Allan, J. (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* (pp. 607–616). New York: ACM. https://doi.org/10.1145/2600428.2609633

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Kammerer, Y, & Gerjets, P. (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), 177–191. https://doi.org/10.1080/10447318.2013.846790

Kammerer, Y., Bråten, I., Gerjets, P., & Strømsø, H. I. (2013). The role of Internet-specific epistemic beliefs in laypersons' source evaluations and decisions during Web search on a medical issue. *Computers in Human Behavior*, *29*(3), 1193–1203. https://doi.org/10.1016/j.chb.2012.10.012

Kammerer, Y., & Gerjets, P. (2013). The role of thinking-aloud instructions and prior domain knowledge in information processing and source evaluation during Web search. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 716–721.

Kellar, M., Watters, C., & Shepherd, M. (2006). A goal-based classification of Web information tasks. *Proceedings of the American Society for Information Science and Technology*, *43*(1), 1–22. https://doi.org/10.1002/meet.14504301121

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., & Yoon, H.-J. (2015). Eye-tracking analysis of user behavior and performance in web search on large and small screens. *Journal of the Association for Information Science and Technology*, *66*(3), 526–544. https://doi.org/10.1002/asi.23187

Lagun, D., & Agichtein, E. (2011). Viewser: Enabling large-scale remote user studies of web search examination and interaction. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 365-374). New York: ACM.

Lagun, D., Hsieh, C.-H., Webster, D., & Navalpakkam, V. (2014). Towards better measurement of attention and satisfaction in mobile search. *Proceedings of the 37th*

International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '14 (pp. 113–122) New York: ACM. https://doi.org/10.1145/2600428.2609631

Lazonder, A. W., & Rouet, J. F. (2008). Information problem solving instruction: Some cognitive and metacognitive issues. *Computers in Human Behavior*, *24*(3), 753–765. https://doi.org/10.1016/j.chb.2007.01.025

Lewandowski, D. (2015). Evaluating the retrieval effectiveness of web search engines using a representative query sample. *Journal of the Association for Information Science and Technology*, *66*(9), 1763–1775. https://doi.org/10.1002/asi.23304

Lewandowski, D., Kerkmann, F., Rümmele, S., & Sünkler, S. (2018). An empirical investigation on search engine ad disclosure. *Journal of the Association for Information Science and Technology*, *69*(3), 420–437. https://doi.org/10.1002/asi.23963

Liu, Z., Liu, Y., Zhou, K., Zhang, M., & Ma, S. (2015). Influence of vertical result in Web search examination. In R. Baeza-Yates, M. Lalmas, A. Moffat, & B. Ribeiro-Neto (Eds.), *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15* (pp. 193–202). New York, New York, USA: ACM Press. https://doi.org/10.1145/2766462.2767714

Lo, S.-K., Hsieh, A.-Y., & Chiu, Y.-P. (2014). Keyword advertising is not what you think: Clicking and eye movement behaviors on keyword advertising. *Electronic Commerce Research and Applications*, *13*(4), 221–228. https://doi.org/10.1016/j.elerap.2014.04.001

Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., … Pan, B. (2008). Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, *59*(7), 1041–1052. https://doi.org/10.1002/asi.20794

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., & Gay, G. (2006). The influence of task and gender on search and evaluation behavior using Google.

Information Processing and Management, *42*(4), 1123–1131. https://doi.org/10.1016/j.ipm.2005.10.001

Lu, W., & Jia, Y. (2014). An eye-tracking study of user behavior in web image search. In *Pacific Rim International Conference on Artificial Intelligence* (pp. 170–182). Cham: Springer.

Lund, H. (2016). Eye tracking in library and information science: a literature review. *Library Hi Tech*, *34*(4), 585–614. https://doi.org/10.1108/LHT-07-2016-0085

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, *49*(4), 41–46.

Marcos, M.-C., Gavin, F., & Arapakis, I. (2015). Effect of snippets on user experience in Web search. In *Proceedings of the XVI International Conference on Human Computer Interaction - Interacción '15* (pp. 1–8). New York, New York, USA: ACM Press. https://doi.org/10.1145/2829875.2829916

Muntinga, T., & Taylor, G. (2017). Information-seeking strategies in medicine queries: a clinical eye-tracking study with gaze-cued retrospective think-aloud protocol. *International Journal of Human–Computer Interaction*, (1), 1–13. https://doi.org/10.1080/10447318.2017.1368949

Muralidharan, A., Gyongyi, Z., & Chi, E. H. (2012). Social annotations in web search. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (pp. 1085–1094). New York, New York, USA: ACM Press. https://doi.org/10.1145/2207676.2208554

O'Shea, R. P. (1991). Thumb's rule tested: visual angle of thumb's width is about 2 deg. *Perception*, *20*(3), 415–418. https://doi.org/10.1068/p200415

Orquin, J. L., & Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. *Behavior Research Methods*, *50*(4), 1645–1656. https://doi.org/10.3758/s13428-017-0998-z

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, *12*(3), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

Papoutsaki, A., Laskey, J., & Huang, J. (2017). SearchGazer: Webcam eye tracking for remote studies of web search. In *2nd ACM SIGIR Conference on Information Interaction and Retrieval, CHIIR 2017* (pp. 17–26). New York: ACM. https://doi.org/10.1145/3020165.3020170

Pirolli, P. (2007). *Information Foraging Theory: Adaptive Interaction with Information*. Oxford, UK: Oxford University Press.

Poole, A., & Ball, L. J. (2006). Eye Tracking in HCI and Usability Research. In *Encyclopedia of Human Computer Interaction* (pp. 211–219).

Purcell, K., Brenner, J., & Raine, L. (2012). *Search Engine Use 2012*. Washington, DC. Retrieved from http://pewinternet.org/~/media/Files/Reports/2012/PIP_Search_Engine_Use_2012.pdf

Rayner, K. (1998). Eye movements in reading and information processing. 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rele, R. S., & Duchowski, A. T. (2005). Using eye tracking to evaluate alternative search results interfaces. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, pp. 1459–1463). Orlando: SAGE Publications. Retrieved from http://andrewd.ces.clemson.edu/research/vislab/docs/Final_HFES_Search.pdf

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology*, *53*(2), 145–161.

Rovira, C., Capdevila, J., & Marcos, M.-C. (2014). The importance of sources in the selection of online newspaper articles: A study of Google Noticias using eye-tracking.

Investigacion Bibliotecologica, *28*(63), 15–27. Retrieved from http://rev-ib.unam.mx/ib/index.php/ib/article/viewFile/57780/51693

Saito, H., Terai, H., Egusa, Y., Takaku, M., Miwa, M., & Kando, N. (2009). How task types and user experiences affect information-seeking behavior on theWeb: using eye-tracking and client-side search logs. In *Proceedings of the Workshop on Understanding the User - Logging and Interpreting User Interactions in Information Search and Retrieval* (pp. 19–22).

Scheiter, K., & van Gog, T. (2009). Using eye tracking in applied research to study and stimulate the processing of information from multi-representational sources. *Applied Cognitive Psychology*, *23*(9), 1209–1214. https://doi.org/10.1002/acp.1524

Schultheiß, S., Sünkler, S., & Lewandowski, D. (2018). We still trust in google, but less than 10 years ago: An eye-tracking study. *Information Research*, *23*(3).

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Siu, C., & Chaparro, B. S. (2014). First look: examining the horizontal grid layout using eye-tracking. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1119–1123). Software Usability Research Lab., Wichita State University, United States. https://doi.org/10.1177/1541931214581234

Sullivan, D. (2016). Google now handles at least 2 trillion searches per year. Retrieved from http://searchengineland.com/google-now-handles-2-999-trillion-searches-per-year-250247

Tavani, H. (2012, August 27). Search engines and ethics. Retrieved August 12, 2015, from http://plato.stanford.edu/entries/ethics-search/

Thomas, P., Scholer, F., & Moffat, A. (2013). What users do: The eyes have it. In Asia Information Retrieval Symposium (pp. 416-427). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-45068-6_36

Turpin, A., Scholer, F., Billerbeck, B., & Abel, L. A. (2006). Examining the pseudo-standard web search engine results page. In *Proceedings of the Eleventh Australasian Document Computing Symposium, ACDS 2006*.

Varian, H. (2006). The economics of Internet search. *Rivista Di Politica Economica*, (November-December), 177–191.

Walhout, J., Oomen, P., Jarodzka, H., & Brand-Gruwel, S. (2017). Effects of task complexity on online search behavior of adolescents. *Journal of the Association for Information Science and Technology*, *68*(8), 1449–1461.

Wang, C., Liu, Y., Zhang, M., Ma, S., Zheng, M., Qian, J., & Zhang, K. (2013). Incorporating vertical results into search click models. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13* (pp. 503–512). New York: ACM. https://doi.org/10.1145/2484028.2484036

Westerwick, A. (2013). Effects of sponsorship, Web site design, and Google ranking on the credibility of online information. *Journal of Computer-Mediated Communication*, *18*(2), 80–97. https://doi.org/10.1111/jcc4.12006

White, R. W. (2016). *Interactions with search systems*. New York: Cambridge University Press.

White, R. W., & Horvitz, E. (2009). Cyberchondria. *ACM Transactions on Information Systems*, *27*(4), Article No. 23. https://doi.org/10.1145/1629096.1629101

Xie, X., Liu, Y., Wang, X., Wang, M., Wu, Z., Wu, Y., … Ma, S. (2017). Investigating examination behavior of image search users. In *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275–284). New York: ACM. https://doi.org/10.1145/3077136.3080799

APPENDIX: EXCLUSION TABLE

No.	Reference	Reason for exclusion
1	Al-Rashid, W., Al-Dosary, B., & Al-Salloom, R. (2010). Visual Attention on Search Engine Results: Information Retrieval of Arabic Web Content. In *IADIS International Conferences Informatics 2010, Wireless Applications and Computing 2010 and Telecommunications, Networks and Systems 2010* (pp. 247–249).	No tests for statistical significance regarding the factors under examination
2	Alsaffar, M. (2017). Online visual search behaviour, performance and satisfaction: A comparative study of English, Arabic and Chinese users. In Proceedings of the *11th International Conference on Research Challenges in Information Science (RCIS)*, Brighton, 2017, pp. 408-413.	No empirical result reported
3	Alsaffar, M., Pemberton, L., Echavarria, K. R., & Sathiyanarayanan, M. (2017). Visual behaviour in searching information: A preliminary eye tracking study. In *Proceedings of the 11th International Conference on Research Challenges in Information Science (RCIS)* (pp. 365–370). University of Brighton, United Kingdom: IEEE. https://doi.org/10.1109/RCIS.2017.7956560	No tests for statistical significance regarding the factors under examination
4	Al-Samarraie, H., & Al-Hatem, A. I. (2018). The effect of web search result display on users' perceptual experience and information seeking performance. *Reference Librarian, 59(1)*, 10-18. https://doi.org/10.1080/02763877.2017.1399849	No tests for statistical significance regarding the factors under examination
5	Berget, G., & Sandnes, F. E. (2016). Do autocomplete functions reduce the impact of dyslexia on information-searching behavior? The case of Google. *Journal of the Association for Information Science and Technology*, *67*(10), 2320-2328.	Focus on query entries, not on search results
6	Beckers, T., & Korbar, D. (2011). Using eye-tracking for the evaluation of interactive information retrieval. In *INEX 2010* (pp. 236–240). https://doi.org/10.1007/978-3-642-23577-1_21	No empirical results reported
8	Buscher, G., Van Elst, L., & Dengel, A. (2009). Segment-level display time as implicit feedback: a comparison to eye tracking. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 67-74). New York: ACM.	Eye-tracking used as input data
8	Chen, Y., Liu, Y., Zhou, K., Wang, M., Zhang, M., & Ma, S. (2015, October). Does vertical bring more satisfaction?: Predicting search satisfaction in a heterogeneous environment. In *Proceedings of the 24th ACM International on Conference on Information and*	Eye-tracking used as input data

	Knowledge Management (pp. 1581-1590). New York: ACM.	
9	Chizari, S. (2016). Exploring the role of culture in online searching behavior from cultural cognition perspective. *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval - CHIIR '16*, 349–351. New York: ACM. https://doi.org/10.1145/2854946.2854954	No empirical result reported
10	Djamasbi, S., Hall-Phillips, A., & Yang, R. (2013a). An examination of ads and viewing behavior: An eye tracking study on desktop and mobile devices. In *19th Americas Conference on Information Systems, AMCIS 2013 - Hyperconnected World: Anything, Anywhere, Anytime* (Bd. 1, S. 350–355). User Experience and Decision Making Lab, School of Business, Worcester Polytechnic Institute, United States.	No tests for statistical significance regarding the factors under examination
11	Djamasbi, S., Hall-Phillips, A., & Yang, R. R. (2013b). SERPs and ads on mobile devices: An eye tracking study for generation Y. In *International Conference on Universal Access in Human-Computer Interaction* (pp. 259-268). Springer, Berlin, Heidelberg.	No tests for statistical significance regarding the factors under examination
12	Domachowski, A., Griesbaum, J., & Heuwing, B. (2016). Perception and effectiveness of search advertising on smartphones. In *Proceedings of the Association for Information Science and Technology, 53(1)*, 1–10. https://doi.org/10.1002/pra2.2016.14505301074	No tests for statistical significance regarding the factors under examination
13	Egusa, Y., Takaku, M., Terai, H., Saito, H., Kando, N., & Miwa, M. (2008). Visualization of user eye movements for search result pages. In Proceedings of the 2nd International Workshop on Evaluating Information Access (EVIA 2008) (pp. 42-46). National Center of Sciences, Tokyo, Japan	Data visualization only
14	Feild, H., White, R. W., & Fu, X. (2013). Supporting orientation during search result examination. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13* (S. 2999). New York: ACM. https://doi.org/10.1145/2470654.2481416	Eye-tracking results address behavior on landing pages, not on SERPs
15	Garkavijs, V., Okamoto, R., Ishikawa, T., Toshima, M., & Kando, N. (2014). GLASE-IRUKA: gaze feedback improves satisfaction in exploratory image search. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 273-274). New York: ACM.	Eye-tracking used as input data
16	Golenia, J. E., Wenzel, M. A., Bogojeski, M., & Blankertz, B. (2018). Implicit relevance feedback from electroencephalography and eye tracking in image search. *Journal of Neural Engineering*, *15*(2).	Eye-tracking used as input data

17	Gossen, T., Höbel, J., & Nürnberger, A. (2014). A comparative study about children's and adults' perception of targeted web search engines. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (S. 1821–1824). New York: ACM. https://doi.org/10.1145/2556288.2557031	No tests for statistical significance regarding the factors under examination
18	Gossen, T., Höbel, J., & Nürnberger, A. (2014). Usability and perception of young users and adults on targeted web search engines. In *Proceedings of the 5th Information Interaction in Context Symposium on - IIiX '14* (S. 18–27). New York ACM. https://doi.org/10.1145/2637002.2637007	No tests for statistical significance regarding the factors under examination
19	Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (S. 478–479). New York: ACM.	No tests for statistical significance regarding the factors under examination
20	Grauer, L., & Lomakina, A. (2015). On the Effect of "Stupid" Search Components on User Interaction with Search Engines. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, 1727–1730. https://doi.org/10.1145/2806416.2806601	No tests for statistical significance regarding the factors under examination
21	Guo, Q., & Agichtein, E. (2010). Towards predicting web searcher gaze position from mouse movements. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 3601-3606). New York: ACM.	Approach to substitute eye-tracking recordings
22	Huang, J., White, R. W., & Dumais, S. (2011). No clicks, no problem. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11* (p. 1225). New York: ACM. https://doi.org/10.1145/1978942.1979125	Eye-tracking used as input data
23	Joachims, T., Granka, L. A., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 154-161). New York: ACM.	No tests for statistical significance regarding the factors under examination; For statistical effects see Pan et al. (2007)
24	Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, *25*(2), 1-27.	No tests for statistical significance regarding the factors under examination; For statistical effects see Pan et al. (2007)
25	Kammerer, Y., & Gerjets, P. (2010). How the interface design influences users' spontaneous trustworthiness evaluations of web search results. In *Proceedings of the 2010 Symposium on Eye-Tracking Research &*	Preliminary results, see Kammerer & Gerjets (2014)

	Applications - ETRA '10 (p. 299). New York: ACM. https://doi.org/10.1145/1743666.1743736	
26	Kim, J., Thomas, P., Sankaranarayana, R., & Gedeon, T. (2012). Comparing scanning behaviour in web search on small and large screens. In *Proceedings of the Seventeenth Australasian Document Computing Symposium* (pp. 25-30). New York: ACM.	Preliminary results, see Kim et al. (2015)
27	Lagun, D., & Agichtein, E. (2011). Viewser: Enabling large-scale remote user studies of web search examination and interaction. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 365-374). New York: ACM.	Approach to substitute eye-tracking recordings
28	Liu, Z., Liu, Y., Zhang, M., & Ma, S. (2014). How do sponsored search results affect user behavior in web search? In *Lecture Notes in Computer Science* (Vol. 8870, pp. 73–85). Heildberg: Springer.	No tests for statistical significance regarding the factors under examination
29	Liu, Y., Liu, Z., Zhou, K., Wang, M., Luan, H., Wang, C., ... & Ma, S. (2016, July). Predicting search user examination with visual saliency. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 619-628). New York: ACM.	Eye-tracking used as input data
30	Liu, Y., Wang, C., Zhou, K., Nie, J., Zhang, M., & Ma, S. (2014, November). From skimming to reading: A two-stage examination model for web search. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 849-858). New York: ACM.	Eye-tracking used as input data
31	Liu, Y., Xie, X., Wang, C., Nie, J. Y., Zhang, M., & Ma, S. (2017). Time-aware click model. *ACM Transactions on Information Systems (TOIS)*, *35*(3), 1-24.	Eye-tracking used as input data
32	Lund, H. (2016). Eye tracking in library and information science: a literature review. *Library Hi Tech*, *34*(4), 585–614. https://doi.org/10.1108/LHT-07-2016-0085	Review article
33	Luo, C., Liu, Y., Zhang, M., & Ma, S. (2016, November). Search success evaluation with translation model. In *Asia Information Retrieval Symposium* (pp. 251-266). Cham: Springer.	Eye-tracking used as input data
34	Mao, J., Liu, Y., Zhang, M., & Ma, S. (2014). Estimating credibility of user clicks with mouse movement and eye-tracking information. In *Natural Language Processing and Chinese Computing* (pp. 263-274). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-45924-9_24	Eye-tracking used as input data
36	Matsuda, Y., Uwano, H., Ohira, M., & Matsumoto, K. (2009). An analysis of eye movements during browsing multiple search results pages. In *Proceedings of the*	No tests for statistical significance regarding the factors under examination

	International Conference on Human-Computer Interaction (pp. 121-130). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-02574-7_14	
36	Ostergren, M., Yu, S., & Efthimiadis, E. N. (2010). The value of visual elements in web search. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10* (pp. 867–868). New York, New York, USA: ACM Press. https://doi.org/10.1145/1835449.1835656	No tests for statistical significance regarding the factors under examination
37	Pan, L., & Yang, L. (2015). A review of eye-movement research developments. In *Conference Proceedings of the 3rd International Symposium On Project Management, Ispm 2015* (pp. 216–220).	Review article
38	Papoutsaki, A., Laskey, J., & Huang, J. (2017). SearchGazer: Webcam eye tracking for remote studies of web search. In *2nd ACM SIGIR Conference on Information Interaction and Retrieval, CHIIR 2017* (pp. 17–26). New York: ACM. https://doi.org/10.1145/3020165.3020170	No tests for statistical significance regarding the factors under examination
39	Phillips, A. H., Yang, R., & Djamasbi, S. (2013). Do ads matter? An exploration of web search behavior, visual hierarchy, and search engine results pages. In *2013 46th Hawaii International Conference on System Sciences* (pp. 1563-1568). IEEE.	No tests for statistical significance regarding the factors under examination
40	Poole, A., & Ball, L. J. (2006). Eye Tracking in HCI and Usability Research. In *Encyclopedia of Human Computer Interaction* (pp. 211–219).	
41	Régis, L., Maio, D., Wagner, N., & Cavalcante, F. (2014). High-literate and lLow-literate user interaction: a comparative study using eyetracking in an emergent economy. In *Proceedings of the International Conference on Universal Access in Human-Computer Interaction* (pp. 89–100). Springer.	No tests for statistical significance regarding the factors under examination
42	Şendurur, E., & Yildirim, Z. (2015). Students' web search strategies with different task types: an eye-tracking study. *International Journal of Human-Computer Interaction*, *31*(2), 101–111. https://doi.org/10.1080/10447318.2014.959105	No tests for statistical significance regarding the factors under examination
43	Shah, C., Liu, J., González-Ibáñez, R., & Belkin, N. (2012). Exploration of dynamic query suggestions and dynamic search results for their effects on search behaviors. In *Proceedings of the American Society for Information Science and Technology, 49(1),* 1-10. https://doi.org/10.1002/meet.14504901135	No tests for statistical significance regarding the factors under examination
44	Terai, H., Saito, H., Egusa, Y., Takaku, M., Miwa, M., & Kando, N. (2008). Differences between Informational and Transactional Tasks in Information Seeking on the Web. In *IIiX '08 Proceedings of the second international symposium on Information interaction in*	Preliminary results, see Saito, Terai, Egusa,Takaku, Miwa, & Kando (2009)

	context (pp. 152–159). https://doi.org/10.1145/1414694.1414728	
45	Thomas, P., Scholer, F., & Moffat, A. (2013). What users do: The eyes have it. *In Asia Information Retrieval Symposium* (pp. 416-427). Springer, Berlin, Heidelberg.	No tests for statistical significance regarding the factors under examination
46	Vrochidis, S., Patras, I., & Kompatsiaris, I. (2011). An eye-tracking-based approach to facilitate interactive video search. In *Proceedings of the 1st ACM international conference on multimedia retrieval* (Article Nr. 43). New York: ACM.	Eye-tracking used as input data
47	Wang, C., Liu, Y., Zhang, M., Ma, S., Zheng, M., Qian, J., & Zhang, K. (2013). Incorporating vertical results into search click models. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13* (pp. 503–512). New York: ACM. https://doi.org/10.1145/2484028.2484036	No tests for statistical significance regarding the factors under examination
48	Zhang, Y., & Gwizdka, J. (2016). Rethinking the cost of information search behavior. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 969-972). New York: ACM.	Eye-tracking used as input data
49	Zhu, J., Han, X., Ma, R., Li, X., Cao, T., Sun, S., & Hu, B. (2016). Exploring user mobile shopping activities based on characteristic of eye-tracking. In *Proceedings of the International Conference on Human Centered Computing* (pp. 556-566). Cham: Springer. https://doi.org/10.1007/978-3-319-31854-7_50	No tests for statistical significance regarding the factors under examination