Age-Related Effects of Task Difficulty on the Semantic Relevance of Query Reformulations

Saraschandra Karanam, Herre Oostendorp

Age-related effects of task difficulty on the semantic relevance of query reformulations

Saraschandra Karanam[1] and Herre van Oostendorp[2]

[1] Utrecht University, Utrecht, The Netherlands.
s.karanam@uu.nl
[2] Utrecht University, Utrecht, The Netherlands.
h.vanoostendorp@uu.nl

Abstract. This study examined the semantics of query reformulations in relation to age and task difficulty. Task difficulty was manipulated using a metric called task preciseness defined as the semantic similarity of the task description with the content of the target page(s) containing the answer. A behavioral experiment was conducted in which 24 younger adults and 21 older adults solved six low precise and six high precise information search tasks. The behavioral outcomes were found to be in line with preceding work indicating that the metric was successful in differentiating different levels of task difficulty. Analysis of the semantic relevance of queries showed that for low precise tasks, the queries generated by younger adults had significantly higher mean semantic relevance than that of older adults whereas for high precise tasks, it was the other way round. When analyzed across reformulations, it was found that the mean semantic relevance of queries generated by older adults, decreased for both low and high precise tasks. For younger adults, it remained constant for high precise tasks and even increased for low precise tasks. Implications of these findings for the design of information search systems are discussed.

Keywords: Information Search, Aging, Task Preciseness, Query Reformulations.

1 Introduction

Searching for information on the Internet is a complex cognitive activity involving a number of cognitive processes such as attention, comprehension, decision making and problem solving. Naturally, these cognitive processes are in turn affected by a number of cognitive factors such as age [8], domain knowledge [35], experience with the Internet [51] etc., leading to wide variations in the adoption of the Internet and its efficient use. The focus of this paper is particularly on the effect of aging on information search performance. Older adults are now one of the fastest growing users of the Internet [2, 10]. The Internet is known to decrease isolation by enabling alternate means of communication with near and dear, foster independence, enhance attention and memory [41], and keep the mind active, thereby increasing overall health and well-being of the elderly [33].

However, a number of barriers still exist preventing large scale adoption and usage of the Internet by the elderly. Older adults are known to be slow and less efficient when using the Internet because of their natural decline in motor skills and fluid intelligence

involving processing speed, cognitive flexibility or ability to switch processing strategies, attentional control and visuospatial span [16, 17, 48]. Crystallized intelligence, on the other hand, increases and/or becomes more stable with aging. Crystallized intelligence involves prior knowledge, experience and vocabulary skills. Therefore it is often higher for older adults compared to younger adults [16, 17, 48]. These cognitive abilities directly influence the cognitive processes underlying information search resulting in lower efficiency of older adults on information search tasks. For example, lower processing speed could lead to longer time in evaluating search results or hyperlinks on a website, difficulty in switching strategies could lead to difficulty in reformulating unsuccessful queries or difficulty in getting out of an unsuccessful path, lack of attentional control could lead to inefficient handling of relevant and irrelevant search results or hyperlinks and finally lower visuospatial span could mean less efficient exploration of the search result page or a website.

This is reflected in the outcomes of research from human factors and information science communities: older adults were found to generate less queries, use less keywords per query, reformulate less, spend longer time evaluating the search engine result pages (SERPs, henceforth), spend more time evaluating the content of websites opened from SERPs, switch fewer number of times between SERPs and websites and find it difficult to reformulate unsuccessful queries [8, 11, 24, 26, 37]. Studies that investigated the differences in search strategies employed by younger and older adults show that older adults follow a more structured and methodical approach such as careful selection of query terms and spending longer time evaluating the search results and younger adults are more impulsive which involves frequent switches and clicks on irrelevant search results [52]. Younger adults are able to adapt their strategy more often than older adults [8] and younger adults explore more (number of search results opened for any given query) and exploit less (number of websites and hyperlinks within those websites visited for any given query) whereas older adults exploit more and explore less [9].

The research on search strategies has, however, paid less attention to the impact of aging on query reformulation strategies and the actual content of the queries during reformulations. By actual content, we mean here the semantic aspects of the query terms. These, precisely, form the main focus of this paper. Researchers from the information retrieval community have studied query reformulations extensively [19, 21, 39, 44] and modeled query reformulation behavior based on certain structural patterns such as addition and deletion of terms, observed in them [20, 22, 42]. Researchers from the human factors and information science communities have examined a number of factors that influence query reformulations such as task type [31], cognitive style [28], query type [1], prior domain knowledge and familiarity with the topic of the search [14, 18]. For example, the study of [14] showed that participants with low familiarity tend to alter the spelling and use stemming for query reformulation whereas participants with medium or higher topic familiarity are inclined to add new terms and phrases to reformulate queries. Participants with a higher topic familiarity were also found to make less spelling errors and preferred to use specific terms or search from different aspects [18]. The focus of the above studies is largely on the structural aspects of query reformulations such as addition, retention and deletion of terms and not on the semantic

aspects. They also do not involve any age-related differences. However, it has been known since long that knowledge-based query reformulation strategies are more efficient, as shown by Shute and Smith in [40]. A recent study by [24] investigated both the structural aspects (that is, the type of reformulation: generalization vs. specialization) and the semantic relevance (that is, the semantic relatedness of queries with target information) of query reformulations. Their study found that younger adults used a specialization strategy to reformulate queries significantly more often than older adults. A generalization strategy was also used significantly more often by younger adults, especially for difficult tasks. The semantic relevance of queries with target information was found to be significantly higher for difficult tasks compared to simple tasks. When measured across reformulations, it showed a decreasing trend for older adults and remained constant for younger adults, indicating that as older adults reformulated, they produced queries that were further away from the target information.

It is useful to note from the above studies that the age-related differences in information search performance (in terms of time spent on search results vs. websites, number of clicks on search results, number of queries generated, number of reformulations), search strategies employed (exploration vs. exploitation, structured vs. impulsive) and structural aspects of query reformulations (generalization vs. specialization) were more prominent for difficult tasks. In the next section, we give a brief overview of how task difficulty was characterized in literature.

2 Task Difficulty

The effect of task difficulty on information search performance has been a focus of research for several decades [3, 6, 7, 15, 27, 30, 46, 49]. It is the most frequently incorporated attribute, to assess the performance of a user interacting with an information source (or a system). In spite of this, there is no consensus on how to operationalize task difficulty objectively and how to distinguish between different levels of task difficulty, as is evident from the number of different approaches that exist in literature: availability (simple tasks) or non-availability (difficult tasks) of keywords in the task description that could be used as queries [8,11], size (small for simple tasks vs. large for difficult tasks) and complexity (low for simple tasks vs. high for difficult tasks) of the search space [37], availability (simple tasks) or non-availability (difficult tasks) of the answer in the search snippets [24, 26]. Gwidzka and Spence [15] proposed to assess task difficulty objectively using three factors: path length (the length of the navigation path leading to the target information), page complexity (the complexity of navigation choices on each web-page) and page information assessment (the difficulty of relevance judgement on pages that contain the desired information). [36, 47] extended the work of Gwidzka and Spence by introducing a metric called path relevance which measures the degree to which the task description overlaps in meaning with the hyperlink text on the optimal path to the target page.

In this paper, we define an objective measure of task difficulty called *task preciseness*. Task preciseness measures the degree to which the task description overlaps in meaning with the content of the target page(s) containing the answer to the task. Let us

say, the user is searching for an answer to the following task: "What name is given to the valve that protects food from entering your lungs when you swallow?" and let us assume that the target page contains the following text "The epiglottis is a flap of soft cartilage, covered by a mucous membrane. It is attached to the back of the tongue and acts as a valve during swallowing to prevent food and liquid from entering the lungs." It is clear from the above example that the degree of overlap between the task description and the actual text on the target page is very high (swallow-swallowing, prevent-protect, food-food, lungs-lungs, valve-valve), almost directly giving the answer the user is searching for. If instead, the target page contains the following text: "Lips and tongue keep food in the mouth and in place prior to swallowing. The soft tissue created by the cricopharyngeus muscle, also called as epiglottis, at the top of the esophagus keeps air out of the digestive system during breathing", the degree of overlap between the task description and the actual text on the target page is very low (swallow-swallowing). The answer is not directly available and has to be indirectly inferred by the user. In order to validate the efficacy of task preciseness, we conduct an experiment with low and high precise information search tasks created using the new metric and check if the information search performance of real subjects on these tasks is indeed in line with known prior outcomes on task difficulty or not. We use Latent Semantic Analysis (LSA, henceforth) [29] to compute the degree of overlap in meaning or semantic similarity values. While we are aware of other methodologies to compute semantic similarity values such as LDA [4], HAL [32], PMI-IR [45] etc., we use LSA in our study because LSA scores have been shown to significantly overlap with those of human scores on synonym, antonym and subject matter tests. LSA has been successful in mimicking human sorting and category judgments, accurately estimating passage coherence, learnability of passages and the quality and the quantity of knowledge in an essay [12, 34, 50]. LSA is a machine-learning technique that builds a semantic space representing a given user population's understanding of words, short or whole texts by applying statistical computations, singular value decomposition and represents them as a vector in a multidimensional space of about 300 dimensions. The cosine value (+1 if identical and 0 if unrelated) between two vectors in this representation gives the measure of the semantic relatedness. It has been shown that higher semantic relatedness between two texts relates to higher overlap in the meanings associated with those two texts. Therefore, a high LSA value indicates a high overlap in the description of the task and the content pages containing the answer to the task.

As an example of a high precise task, for the task (presented in Dutch) "*Bij patient Jansen is waarschijnlijk sprake van een hersenbloeding omdat er een bloeding in en rondom de hersenen lijkt te zijn geweest. Op een CT-scan is een misvormd bloedvat te zien. Welke opties voor een operatieve behandeling heeft de neurochirurg? (Patient Jansen has probably a cerebral haemorrhage because of bleeding in and around the brain. A CT scan shows a malformed blood vessel. What options for a surgical procedure does a neurosurgeon have?)*", the target page containing the answer also contains words such as "*hersenbloeding (cerebral haemorrhage)*", "*misvormd bloedvat (malformed blood vessel)*", "*CT-scan*", "*neurochirurg (neurosurgeon)*" etc. leading to a high degree of overlap between the task description and the content of the target page (LSA value = 0.75).

Similarly, a low LSA value indicates a low overlap in the description of the task and the content pages containing the answer to the task. As an example of a low precise task, for the task (presented in Dutch), "*Fieke, pas 6 jaar, heeft behoefte aan veel water drinken en moet ook vaak plassen. Ook is ze vaak erg uitgeput. De arts stelt vast dat de waarde van haar glucose veel te hoog is. Wat zou er aan de hand kunnen zijn, wees specifiek. Welke behandeling zal de arts dan inzetten? (Fieke, 6 years old, needs to drink plenty of water and must urinate frequently. She is often very exhausted. The doctor notes that the value of its glucose is too high. What could be the problem, be specific. What treatment will deploy the doctor?)*", the words such as "*veel water drinken (drink plenty of water)*", "*vaak plassen (urinate frequently)*", "*uitgeput (exhausted)*", "*glucose veel te hoog (glucose level is too high)*" are not part of the target page containing the answer, leading to a low degree of overlap between the task description and the content of the target page (LSA value = 0.38).

Therefore, tasks with a high LSA value of task preciseness provide better, more precise contextual information pointing to the target information. Tasks with a low LSA value of task preciseness would, on the other hand, require the user to engage in higher level cognitive activities such as using his/her own knowledge to understand the task, generate relevant queries, examine search results and determine their usefulness etc. The advantage of using LSA is that it provides us an automatic and objective way of calculating overlap in meaning.

Overall, we make the following contributions in this paper:

1. We investigate the age-related differences in information seeking performance on varying levels of task difficulty.
2. We examine the age-related variations in the semantic aspects of query reformulations under varying levels of task difficulty.
3. We examine the age-related variations in the semantic aspects of queries measured across reformulations.

The remainder of this paper is organized as follows. Section 3 lists the research questions of this paper. Section 4 gives details of the experiment conducted and the results obtained. Section 5 concludes the paper with a discussion on future directions.

3 Research Questions

There were three main research questions that motivated this study:

1. What impact does task difficulty, operationalized as task preciseness, have on information search performance, measured in terms of task-completion time, number of clicks, task accuracy and number of reformulations? **(RQ1)**.
2. How would the information seeking performance vary between younger and older adults on different levels of task difficulty? **(RQ2)**.
3. How does the semantic relevance of a query with the target information sought vary in relation to age and task difficulty? **(RQ3a)**. How does the semantic relevance of

a query with target information sought vary in relation to age and task difficulty when analyzed at a more granular level: across reformulations? **(RQ3b).**

4 Experiment

4.1 Method

Participants. 24 younger adults (17 males and 7 females) ranging from 18 to 27 years (M = 21.08, SD = 1.9), and 21 older adults (11 males and 10 females) ranging from 66 to 88 years (M = 75.52, SD = 6.85) participated in the study.

Design. We followed a 2 (Age: Young vs. Old) X 2 (Task Preciseness: Low vs. High) mixed design with age as between-subjects variable and task preciseness as within-subjects variable.

Material. *Websites.* Five mockup websites based on material from popular Dutch medical and health websites were built. The URLs of the real websites from which the content of our mockup websites was sourced and adapted, the number of topics, the number of pages and the maximum depth of each website are presented in Table 1.

Table 1. Details of mockup websites used.

Website	Number of Topics	Number of Pages	Maximum Depth
Website 1[1]	6	37	6
Website 2[2]	11	194	6
Website 3[3]	10	124	5
Website 4[4]	11	38	4
Website 5[5]	10	65	5

It was ensured that the websites look realistic both in terms of content (text and pictures) and the visual layout and information architecture. These five websites were indexed using Google's custom search engine[6]. We ran our experiment on mockup material that was designed by us because of the following reason: we need to know the target pages in advance in order to be able to compute task preciseness.

Material. *Information Search Tasks.* The experiment was conducted with twelve simulated information search tasks [5], all from the domain of health (because it is interesting for older adults), divided into six low precise and six high precise tasks based on

[1] https://www.hartstichting.nl/
[2] https://www.dokterdokter.nl/
[3] https://www.gezondheidsplein.nl/
[4] http://www.gezondheid.nl/
[5] https://gezondnu.nl
[6] https://cse.google.com

the semantic similarity between the task description and the content of the target page(s). We used LSA [29] to compute the similarity value between a task description and the content of its corresponding target page(s). An independent samples t-test between the semantic similarity values obtained for low and high precise tasks showed a significant difference $t(10) = -2.2$, $p<.05$. The mean semantic similarity between the task description and the content of the target page(s) was significantly higher for high precise tasks (M=0.7, SD=0.058) compared to low precise tasks (M=0.54, SD=0.15), indicating that the amount of overlap between the task description and the corresponding target page(s) is significantly higher in high precise tasks compared to low precise tasks. The tasks were all presented in Dutch.

4.2 Procedure

Participants first filled out a demographic questionnaire in which they were asked details about their age, gender, number of hours spent on the Internet per week and number of years of experience with computers. Based on the self-reported answers, older adults (M=20.73, SD=14.6) were found to be spending significantly less number of hours on the Internet compared to younger adults (M=37.8, SD=13.14) $t(20) = 3.78$, $p<.001$. Older adults were significantly longer experienced with computers (M=24.7 years, SD=10.92) than younger adults (M=13.0 years, SD=2.16) $t(20)=-4.78$, $p<.001$.

They were next presented with two tests: a computerized version of a Dutch vocabulary test, adapted from Hill Mill Vocabulary (HMV) test [38], and a fluid intelligence test: a computerized version of the Trail Making Test (TMT Part B) [43]. The score on the vocabulary test gives us an indication of the amount of crystallized intelligence and the score on the trail making test gives us an indication of the amount of fluid intelligence. For the vocabulary test, participants were presented with 24 Dutch keywords (one followed by the other) along with six other keywords presented as multiple choice options. For each test keyword, the participants had to choose an option from the six alternatives that is closest in meaning to it. There was only one possible correct answer for each test keyword. Correct choices were scored 1 and wrong choices were scored 0. Thus the maximum possible score on this test is 24 and the minimum possible score is 0. In line with the traditional cognitive aging literature, it was found that the scores of older adults (M=18.52, SD=2.7) on the vocabulary test were significantly higher than that of the scores of younger adults (M=15.3, SD=2.45) $t(20)=-3.7$, $p<.001$ indicating that they had significantly higher crystallized knowledge compared to younger adults.

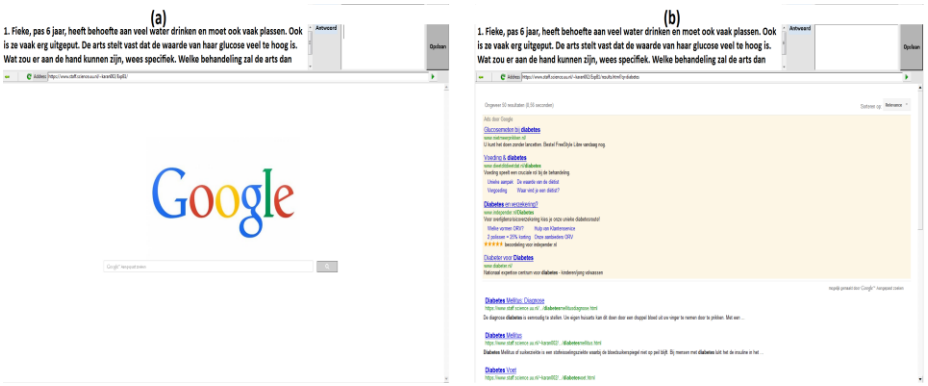

Fig. 1. Interface showing the (a) Google custom search interface and the (b) main screen in which the information search tasks are solved by participants

For the trail making test, participants were shown 25 circles containing both numbers (1 to 13) and alphabets (A to L) on the computer screen. The participants had to click on the circles in ascending pattern alternating between numbers and alphabets (1-A-2-B-3-C and so on) starting from the number 1. If the circle clicked by a participant is right, it turned green, otherwise it turned red. We measured the time taken to finish the test correctly. In line with the traditional cognitive aging literature, significant differences were found in fluid abilities of younger and older adults. Older adults (M=80.75, SD=38.94) took significantly longer to finish this test ($t(20)$=-4.04, p<.001) than younger adults (M=49.37, SD=31.98) indicating that they had significantly lower fluid abilities compared to younger adults.

After the TMT test, participants were allowed a break of five minutes. They were then presented with the twelve information search tasks (six high precise and six low precise). The order of the presentation of tasks was counter balanced. Participants were first shown the task and then directed to the home page of Google's custom search engine. Participants were not allowed to use any other search engine. We show in Figure 1a the interface of Google custom search engine and in Figure 1b the main screen of our interface that participants used while solving the information search tasks. It was ensured the size and placement of Google logo and the search bar below it were exactly similar to the standard interface of Google search. Participants could enter queries as they normally would on any browser and the corresponding search results appeared on the next screen. Users had to always go back to the first screen (Figure 1a) in order to reformulate a query. The task description was made available to the participant at all times in the top left corner. An empty text box was provided in the top right corner for the participant to enter his/her answer. A time limit of 8 minutes was set for each task beyond which the interface automatically took the participant to the next task. All the queries generated by the users, the corresponding search engine result pages and the URLs opened by them were logged in the backend using Visual Basic. We first report the results of information seeking performance in terms of task-completion time, number of clicks, task accuracy and number of reformulations. Next, we report the results

of analyzing the semantic relevance of queries across reformulations with target information.

5 Analysis

5.1 Analysis of Search Performance

In this section we will examine the impact of task difficulty on search performance measured in task-completion time, number of clicks, task accuracy and number of reformulations (RQ1), as well as the impact of age (RQ2).

Measures. We used the following metrics to analyze search performance: task-completion time, number of clicks, accuracy and number of reformulations.

Task-completion time. Task-completion time is computed from the moment of opening a browser and typing in the first query to the moment of answering the question. This includes the time it takes in typing queries, evaluating search results, clicking on one of the search results, evaluating the content of the websites opened from the search results and finally typing the answers.

Number of clicks. Number of clicks is the total number of clicks made by a participant for each task. This includes the clicks made on the search results as well as the clicks made on websites opened from the search results.

Accuracy. Accuracy is measured as 0, 0.25, 0.5, 0,75 or 1 depending on whether the participant's answer was correct (in which case the score is 1) or partially correct (in which case the score is 0.25, 0.5 or 0.75) or wrong (in which case the score is 0). Two researchers scored the participant's answers for their accuracy using the above definition and the inter-scorer reliability (Cronbach's alpha) was found to be very high (> 0.8).

Number of reformulations. Number of reformulations is the total number of unique queries that a user could come up with for each task in the process of answering it (e.g., if participant added, deleted keywords or created new ones, we counted them as reformulations of query).

Results. Data of only those tasks was included in the analysis for which the participants successfully completed the tasks. 14 data points out of a total number of (12 tasks X 45 participants) = 540 data points (2.6%) were therefore dropped. For all the four dependent variables, a 2 (Age: Young vs. Old) X 2 (Task Preciseness: Low vs. High) mixed ANOVA was conducted with age as between-subjects variable and task preciseness as within-subjects variable.

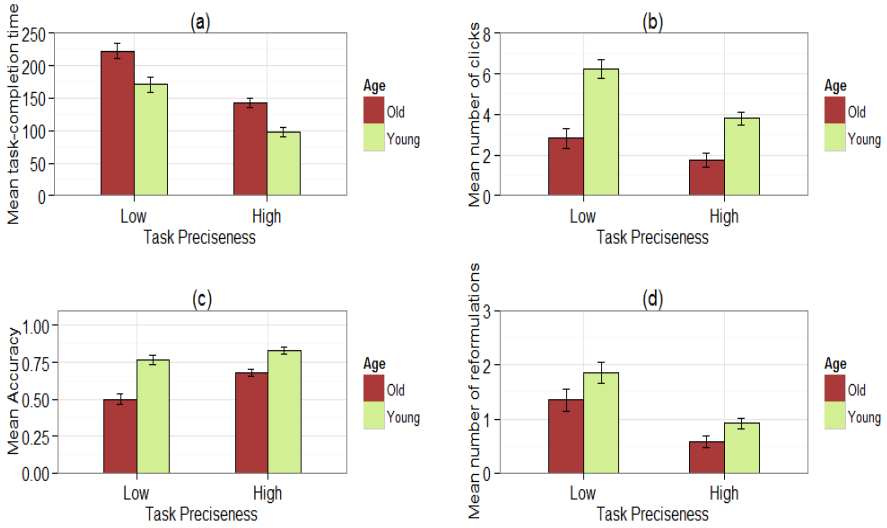

Fig. 2. Analysis of search performance in terms of (a) task-completion time, (b) clicks, (c) accuracy and (d) number of reformulations in relation to age and task preciseness

Task-completion time. The main effect of task preciseness was significant $F(1,43) = 84.76$, $p<.001$. Low precise tasks demanded significantly more time than high precise tasks. See Fig 2a. The main effect of age was significant $F(1,43) = 18.46$, $p<.001$. Older adults took significantly longer to complete tasks compared to younger adults. The interaction of task preciseness and age was not significant ($p>.05$).

Number of clicks. The main effect of task preciseness was significant $F(1,43) = 29.62$, $p<.001$. See Fig 2b. Participants clicked significantly more often for low precise tasks compared to high precise tasks. The main effect of age was significant $F(1,43) = 31.67$, $p<.001$. Younger adults clicked significantly more often than older adults. The interaction of task preciseness and age was significant $F(1,43) = 4.33$, $p<.05$. Younger adults clicked significantly more often than older adults, especially for low precise tasks.

Accuracy. The main effect of task preciseness was significant $F(1, 43) = 19.85$, $p<.001$. See Fig 2c. Accuracy on high precise tasks was significantly higher than accuracy on low precise tasks. The main effect of age was significant $F(1,43) = 37.41$, $p<.001$. Younger adults found significantly more accurate answers than older adults. The interaction of age and task preciseness was also significant $F(1,43) = 4.45$, $p<.05$. Post-hoc tests showed that there was no effect of task preciseness on the accuracy of answers found by younger adults. The accuracy of answers found by older adults, however, dropped significantly for low precise tasks compared to high precise tasks.

Number of reformulations. The main effect of task preciseness was significant $F(1,43) = 40.0$, $p<.001$. See Fig 2d. Queries corresponding to low precise tasks were reformulated significantly more often than the queries corresponding to high precise tasks. The main effect of age was significant $F(1,43) = 5.2$, $p<.05$. Younger adults reformulated significantly more than older adults. The interaction of task preciseness and age was not significant ($p>.05$).

Summarizing the outcomes, low precise tasks demanded significantly more time, significantly more clicks and significantly more reformulations than high precise tasks. Furthermore, the accuracy of low precise tasks was significantly lower than that of high precise tasks. Younger adults were significantly faster in completing tasks compared to older adults, clicked significantly more often, especially for low precise tasks and reformulated significantly more than older adults. The accuracy of older adults was significantly lower than that of younger adults, especially for low precise tasks. These results are in-line with prior outcomes [8, 11, 24, 26, 37] and provide evidence to its validity. The objective metric (task preciseness) we introduced to compute task difficulty was successful in differentiating different levels of task difficulty. We next analyse the queries for their semantic relevance with the target information.

5.2 Analysis of Semantic Relevance of Queries

We will examine in this section whether the semantic relevance of a query does vary in relation to age and task difficulty (RQ3a).

Measures. Semantic relevance was used in the past as a metric to evaluate the content of hyperlink texts (to predict navigation behavior on websites) [13, 23, 25] or the snippets of the search results on the SERPs (to predict interaction behavior with search engines) [26]. However, the semantic aspects of queries are not well studied. In this section, we analyse the semantic aspects of queries using a metric called semantic relevance of query.

Semantic Relevance of Query (SRQ). For each task and each query corresponding to that task, semantic relevance was computed between the query and the target information sought using LSA. [29]. We compute SRQ in the following way: we used 65,000 Dutch documents (consisting of 60% newspaper articles and 40% medical and health related articles) as a corpus to create first a semantic space in Dutch. The LSA values were then computed between a query and the target information for each task. This is repeated for all queries of the task and a mean LSA value is computed. This is repeated again for all the tasks of a participant and finally for all the participants. This metric gives us an estimate of how close in semantic similarity the queries generated by the participants are to the target information. So in general, the higher the SRQ value is, the more relevant the query is.

Results. Data of two older participant had to be dropped for this analysis due to some technical registration problems.

Semantic Relevance of Query (SRQ): A 2 (Age: Young vs. Old) X 2 (Task Preciseness: Low vs. High) mixed ANOVA was conducted with age as between-subjects variable and task preciseness as within-subjects variable. The main effects of task preciseness and age were not significant (p>.05). However, interestingly, the interaction of age and task preciseness was highly significant $F(1,41) = 15.26$, p<.001. Post-hoc tests showed that for low precise tasks, the mean SRQ was significantly higher for younger adults compared to older adults. For high precise tasks, it was the other way round.

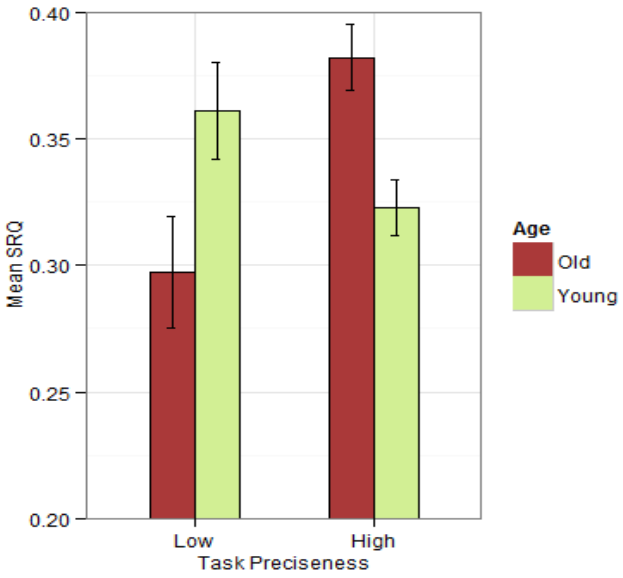

Fig. 3. Mean SRQ of queries with target information in relation to age and task preciseness

Based on the literature, there can be two possible reasons that can explain this interaction effect: differences in fluid abilities and differences in crystallized knowledge between younger and older adults. We examined each possibility further.

1) *Fluid ability*: We saw earlier in Section 4.2 that younger adults reformulated significantly more often than older adults. We could not observe this main effect when the scores on the fluid intelligence test were included as a covariate in the analysis, which indicates that, after controlling for fluid abilities, there was no significant age-related difference in the number of reformulations. This increases the possibility that fluid intelligence is also the reason behind the interaction effect observed between age and task preciseness for mean SRQ. To examine this, we included once more, the scores on the fluid intelligence test as a covariate and repeated the ANOVA analysis with mean SRQ as dependent variable. The interaction of age and task preciseness was still significant $F(1,40) = 13.53$, p<.001. Together with the outcome on the number of reformulations,

it indicates that after controlling for the differences in fluid abilities between younger and older adults, there is no significant difference in the number of reformulations performed by younger and older adults, however, there is still a strong interaction between age and task preciseness on the mean SRQ. Therefore, fluid ability cannot explain the interaction effect.

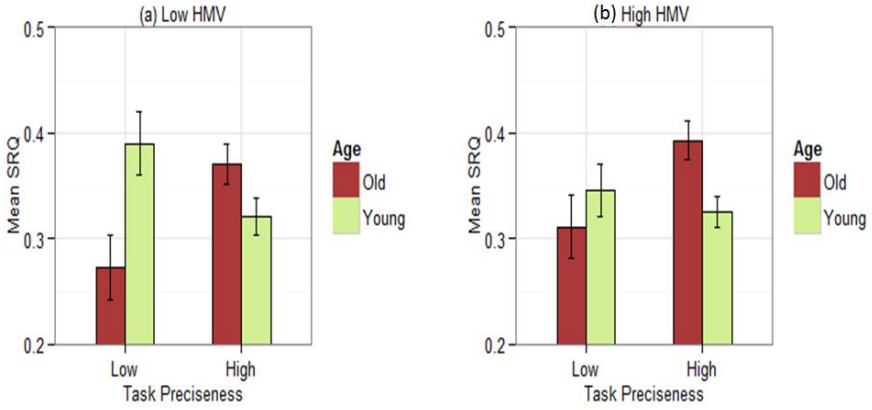

Fig. 4. Mean semantic relevance of queries with target information for participants with a) low crystallized knowledge and b) high crystallized knowledge (HMV) in relation to age and task preciseness

2) Crystallized intelligence: We divided both younger and older adults into two groups of high and low crystallized intelligence based on the median scores on the HMV test. In the high HMV group, we had 14 younger adults and 10 older adults. In the low HMV group, we had 10 younger adults and 9 older adults. For each group, separately, we conducted a 2 (Age: Young vs. Old) X 2 (Task Preciseness: Low vs. High) mixed ANOVA with age as between-subjects variable and task preciseness as within-subjects variable. For the high HMV group, only the interaction of age and task preciseness was significant $F(1,22) = 6.72$, $p<.05$ and no other effects were significant ($p>.05$). See Fig. 4b. Post-hoc tests showed that for low precise tasks, there was no significant difference in the mean SRQ between younger and older adults ($p>.05$). Whereas, for high precise tasks, the mean SRQ was significantly higher ($p<.05$) for older adults compared to younger adults.

For the low HMV group, also, only the interaction of age and task preciseness was significant $F(1,17) = 8.32$, $p<.01$ and no other effects were significant ($p>.05$). See Fig. 4a. However, post-hoc tests revealed an interesting difference with the high HMV group. For high precise tasks, there was no significant difference in the mean SRQ between younger and older adults ($p>.05$). Whereas, for low precise tasks, the mean SRQ was significantly ($p<.05$) lower for older adults compared to younger adults.

Summarizing the outcomes relevant to RQ3a, for high precise tasks, the mean SRQ was significantly higher for older adults compared to younger adults only in the high HMV group and not in the low HMV group. Whereas for low precise tasks, the mean

SRQ was significantly lower for older adults compared to younger adults only in the low HMV group and not in the high HMV group. These outcomes indicate that, older adults, due to their higher crystallized intelligence, are able to utilize the higher contextual information present in the high precise tasks much better than younger adults. But, when it comes to low precise tasks, which demand generating own queries using one's own knowledge and understanding of the task, older adults perform poorly compared to younger adults.

5.3 Analysis of SRQ across reformulations

In this section we will examine the impact of age and task difficulty on the semantic relevance of a query across reformulations (RQ3b). For this we analyzed the mean semantic relevance of the queries with the target information at a more granular level by looking at each reformulation cycle separately. The first cycle corresponds to the first query, the second cycle corresponds to the second subsequent query, and the third cycle corresponds to the third query and so on. The mean semantic relevance was computed for all the queries of all the tasks of a particular type (high precise and low precise separately), generated by young and old participants *in each reformulation cycle*. To achieve higher reliability, only those cycles were considered for which there were at least 4 queries (per reformulation cycle). By doing so, only 3.2% of data was excluded from the analysis. The resulting graphs are shown in Figure 5. It is clear from Figure 5 that younger adults reformulated much longer, that is, more successive queries than older adults.

We tried to answer the following questions in relation to Figure 5:

a) How does the mean SRQ vary across reformulations for younger and older adults? And what effect does task preciseness have on it? For high precise tasks, there was no significant difference in the mean SRQ across reformulations for younger adults whereas for older adults, it decreased as they reformulated (the mean semantic relevance of the ending queries was lower than that of the starting queries, $t(23) = 1.88$, $p=.07$). For low precise tasks, the mean SRQ increased across reformulations for younger adults (the mean semantic relevance of ending queries was significantly higher than that of the starting queries, $t(26) = -3.5$, $p<.005$), whereas it decreased for older adults (the mean semantic relevance of the ending queries was significantly lower than that of the starting queries, $t(20) = 2.15$, $p<.05$).

b) Do younger and older adults start a search task with a similar SRQ? What happens to this difference as they reformulate? And what is the effect of task preciseness in this context? For high precise tasks, there was no significant age difference in the semantic relevance of either the starting or the ending queries. For low precise tasks as well, there was no significant age difference in the semantic relevance of the starting queries. However, the ending queries of younger adults had a higher mean semantic relevance than the ending queries of older adults and this difference was close to conventional significance, $t(5) = 2.2$, $p=.07$. Because of the small number of observations for low precise tasks, we examined the difference between the first four and the last four queries of a session and found no significant age difference between the first four queries. However,

for the last four queries, the mean SRQ was significantly higher for younger adults compared to older adults, $t(18) = -3.22$, $p<.001$.

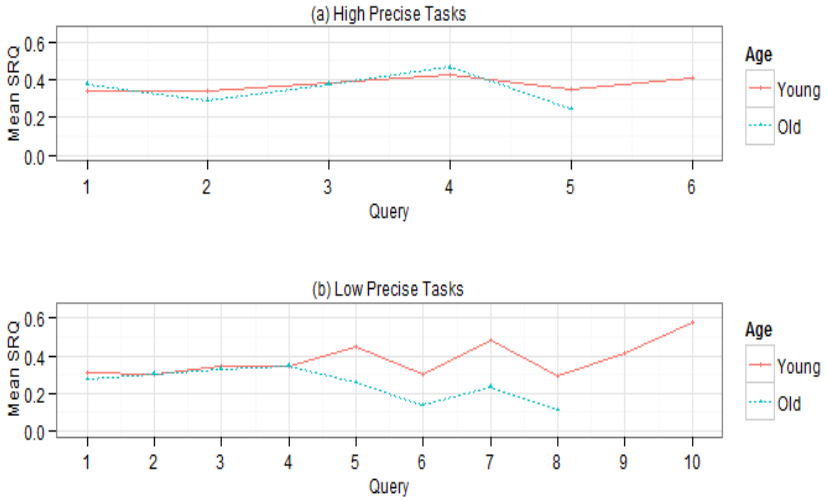

Fig. 5. Mean semantic relevance of queries with target at each reformulation cycle for (a) high precise and (b) low precise tasks

6 Conclusions and Discussion

This study focused on the semantics of query reformulations in relation to age and task difficulty. Task difficulty was manipulated using a metric called task preciseness defined as the degree to which the task description overlaps with the content of the target page(s) containing the answer. It is computed as the semantic similarity between the task description and the content of the target pages using LSA. We conducted an experiment in which 24 younger adults and 21 older adults solved twelve information retrieval tasks divided into two levels of task difficulty based on our task preciseness metric making use of LSA. For low precise tasks, the semantic similarity value was low and for high precise tasks, it was high. Analysis of search performance on these two groups of tasks shows significant differences: low precise tasks demanded significantly more time, significantly more clicks and significantly more reformulations than high precise tasks. Furthermore, the accuracy of low precise tasks was significantly lower than that of high precise tasks. These outcomes are in-line with outcomes reported in prior work [24] (RQ1). It indicates that our objective measure was valid and sensitive enough in differentiating two levels of task difficulty.

We next examined the effects of age (RQ2), younger adults were, as expected, significantly faster in completing tasks compared to older adults, clicked significantly more often, especially for low precise tasks and reformulated significantly more than older adults. The accuracy of older adults in task performance was significantly lower

than that of younger adults, especially for low precise tasks. These outcomes are also in-line with aging-related literature [8, 11, 24, 26, 37] (RQ2). It indicates that our task preciseness metric is able to successfully simulate the effects of task difficulty on aging.

We next analysed the age-related differences in the semantic relevance of queries with target information. For low precise tasks, the mean SRQ was significantly higher for younger adults compared to older adults. For high precise tasks, it was the other way round (RQ3a). We ruled out the possibility that fluid abilities could be the underlying cause for the interaction effect by including the score on our fluid intelligence test as a covariate. The interaction effect was still found to be significant. We then checked if differences in crystallized intelligence could explain the interaction effect. We divided both younger and older adults into two groups of high and low crystallized intelligence based on the median scores on our HMV test and repeated the analysis on each group separately. For high precise tasks, the mean SRQ was significantly higher for older adults compared to younger adults only in the high HMV group and not in the low HMV group. Whereas for low precise tasks, the mean SRQ was significantly lower for older adults compared to younger adults only in the low HMV group and not in the high HMV group. These outcomes indicate that, older adults, due to their higher crystallized intelligence, are able to utilize the better specified contextual information present in the high precise tasks more efficiently than younger adults. But, when it comes to low precise tasks, which demand generating own queries using one's own knowledge and understanding of the task, older adults perform poorly compared to younger adults.

Lastly, we examined the age-related differences in the mean SRQ *across reformulations*. Younger adults were found to reformulate much longer than older adults. For older adults, the mean SRQ decreased for both low and high precise tasks. For younger adults, the mean SRQ remained constant for high precise tasks and even increased for low precise tasks. Furthermore, both younger and older adults start a session with queries of similar SRQ value. They also end a session with queries of similar SRQ value for high precise tasks. For low precise tasks at the end of a session, the mean SRQ was found to be significantly higher for younger adults compared to older adults (RQ3b). It is important to note that these outcomes are largely in line with [24].

One of the main limitations of our work is that the task preciseness metric we defined can be used only for those types of tasks for which there is a known target answer page(s). Therefore, it is necessary to know the target page(s) in advance to compute the task preciseness. Though this limits the applicability of the metric in real environments where neither the user intent nor the target answer are known before hand, it can be very useful in providing training and support to users with low information search skills. We describe how this training and support can be provided in the next section.

7 Design Implications

Based on the behavioral outcomes and the analysis of the content of search queries during reformulations, we come up with the design and methodology of constructing two types of automatic tools that can support interaction with a search engine.

7.1 Support Tool 1

We saw in the analysis of search outcomes that older adults take much longer time to finish tasks than younger adults (See Figure 2). One of the possible reasons could be that they are unable to differentiate between a relevant and a non-relevant search result as efficiently as the younger adults do. We propose a support tool that visually highlights the most relevant search results for a given query, as shown in Figure 6. This methodology was successfully used in the past to provide navigation support within websites [23]. We propose to extend the same methodology to generate support for interaction with a search engine.

Given a query, semantic relevance is computed between the query and each of the search results on the basis of LSA. The search result with the maximum semantic relevance is highlighted with a green arrow as shown in Figure 6. This form of support would enable older adults to spend less mental resources in differentiating between a relevant and a non-relevant search result which in turn would lead to better accuracy.

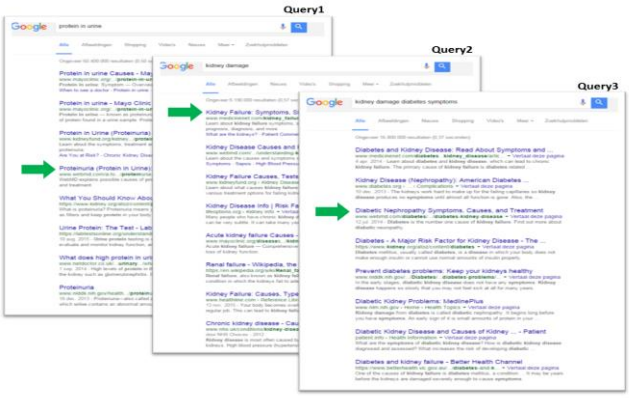

Fig. 6. Design of a support tool for interacting with a search engine that highlights the most relevant search result for a given query

7.2 Support Tool 2

We saw in the analysis of semantic relevance of queries across reformulations that the SRQ of younger adults remained constant and that of the older adults decreased as they reformulated further. In other words, older adults were going further away in semantic distance from the target information as they reformulate, which probably could be one of the reasons for their lower accuracy (See Figure 2). This support tool is intended to ensure that a user does not digress too far away from the goal information in the form of irrelevant queries. To address this problem, we propose a second support tool that monitors - based on the LSA value - the average semantic relevance of the SERPs with the goal information derived from the query and warns the user when it falls below a threshold as shown in Figure 7. This form of support would indicate to

the users that their search results on a page are not relevant enough and they can use this information to take corrective actions such as generating a more relevant query.

Fig. 7. Design of a support tool for interacting with a search engine that monitors the average semantic relevance of the user's queries to the goal information

The study of [24] and the results of our study indicate that such a support system would be of immense help to older adults in improving their overall search performance. We envision both tools to be used for training purposes with a large collection of tasks and their corresponding expected answers. The efficacy of such a training and support mechanism in improving the semantic relevance of search queries of older adults in real contexts needs to be empirically verified.

Acknowledgements. This research was supported by Netherlands Organization for Scientific Research (NWO), ORA Plus project MISSION (464-13-043).

References

1. Aloteibi, S., Sanderson, M. 2014. Analyzing geographic query reformulation: An exploratory study. Journal of the Association for Information Science and Technology, 65(1), 13-24.
2. Aula, A. 2005. User study on older adults' use of the Web and search engines. Universal Access in the Information Society, 4(1), 67-81.
3. Bell, D. J., Ruthven, I. 2004. Searcher's assessments of task complexity for web searching. In European Conference on Information Retrieval (pp. 57-71). Springer Berlin Heidelberg.
4. Blei, D. M., Ng, A. Y., Jordan, M. I. 2003. Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.
5. Borlund, P., Ingwersen, P. 1997. The development of a method for the evaluation of interactive information retrieval systems. Journal of Documentation, 53(3), 225-250.

6. BystrÖm, K. 2002. Information and information sources in tasks of varying complexity. Journal of the American Society for Information Science and Technology, 53(7), 581-591.
7. Campbell, D. J. (1988). Task complexity: A review and analysis. Academy of Management Review, 13(1), 40-52.
8. Chevalier, A., Dommes, A., Marquié, J. C. 2015. Strategy and accuracy during information search on the Web: Effects of age and complexity of the search questions. Computers in Human Behavior, 53, 305-315.
9. Chin, J., Anderson, E., Chin, C. L., Fu, W. T. 2015. Age differences in information search: An exploration-exploitation tradeoff model. In Proceedings of the Human Factors and Ergonomic Society (HFES 2015). 85–89.
10. Dinet, J., Brangier, E., Michel, G., Vivian, R., Battisti, S., Doller, R. 2007. Older people as information seekers: Exploratory studies about their needs and strategies. In International Conference on Universal Access in Human-Computer Interaction (pp. 877-886). Springer Berlin Heidelberg.
11. Dommes, A., Chevalier, A., Lia, S. 2011. The role of cognitive flexibility and vocabulary abilities of younger and older users in searching for information on the web. Applied Cognitive Psychology, 25, 5 (2011), 717–726.
12. Foltz, P. W., Kintsch, W., Landauer, T. K. 1998. The measurement of textual coherence with latent semantic analysis. Discourse Processes, 25(2-3), 285-307.
13. Fu, W. T., Pirolli, P. 2007. SNIF-ACT: A cognitive model of user navigation on the World Wide Web. Human–Computer Interaction, 22(4), 355-412.
14. Ghosh, D. 2016. Effects of Topic Familiarity on Query Reformulation Strategies. In Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (pp. 261-264). ACM.
15. Gwizdka, J., Spence, I. 2006. What can searching behavior tell us about the difficulty of information tasks? A study of Web navigation. Proceedings of the American Society for Information Science and Technology, 43(1), 1-22.
16. Horn, J. L. 2012. The Theory of Fluid and Crystallized Intelligence in Relation to Concepts of Cognitive Psychology and Aging. In Aging and Cognitive Processes. Vol. 8. Springer Science & Business Media, 237–278.
17. Horn, J. L., Cattell, R. B. 1967. Age differences in fluid and crystallized intelligence. Acta Psychologica, 26 (1967), 107–129.
18. Hu, R., Lu, K., Joo, S. 2013. Effects of topic familiarity and search skills on query reformulation behavior. Proceedings of the American Society for Information Science and Technology, 50(1), 1-9.
19. Huang, J., Efthimiadis, E. N. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In Proceedings of the 18th ACM Conference on Information and knowledge Management (pp. 77-86). ACM.
20. Jansen, B. J., Booth, D. L., Spink, A. 2009. Patterns of query reformulation during Web searching. Journal of the Association for Information Science and Technology, 60(7), 1358-1371.
21. Jansen, B. J., Spink, A., Narayan, B. 2007. Query modifications patterns during web searching. In Information Technology, 2007. ITNG'07. Fourth International Conference on (pp. 439-444). IEEE.
22. Jiang, J., Ni, C. 2016. What Affects Word Changes in Query Reformulation During a Task-based Search Session? In Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (pp. 111-120). ACM.
23. Juvina, I., Van Oostendorp, H. 2008. Modeling semantic and structural knowledge in Web navigation. Discourse Processes, 45, 4-5 (2008), 346–364.

24. Karanam, S., Van Oostendorp, H. 2016. Age-related differences in the content of search queries when reformulating. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5720-5730). ACM.
25. Karanam, S., Van Oostendorp, H., Fu, W. T. 2016. Performance of computational cognitive models of web-navigation on real websites. Journal of Information Science, 42, 1, (2016), 94–113.
26. Karanam, S., van Oostendorp, H., Sanchiz, M., Chevalier, A., Chin, J., Fu, W. T. 2015. Modeling and predicting information search behavior. In Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, Article Number 7. ACM.
27. Kim, J. 2006. Task as a predictable indicator for information seeking behavior on the Web. ProQuest.
28. Kinley, K., Tjondronegoro, D., Partridge, H., Edwards, S. 2012. Human-computer interaction: the impact of users' cognitive styles on query reformulation behaviour during web searching. In Proceedings of the 24th Australian Computer-Human Interaction Conference (pp. 299-307). ACM.
29. Landauer, T. K., McNamara, D. S., Dennis, S., Kintsch, W. 2007. Handbook of Latent Semantic Analysis. Mahwah, NJ: Erlbaum.
30. Li, Y. 2009. Exploring the relationships between work task and search task in information search. Journal of the American Society for Information Science and Technology, 60(2), 275-291.
31. Liu, C., Gwizdka, J., Liu, J., Xu, T., Belkin, N. J. 2010. Analysis and evaluation of query reformulations in different task types. Proceedings of the American Society for Information Science and Technology, 47(1), 1-9.
32. Lund, K., Burgess, C. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers, 28(2), 203-208.
33. Mellor, D., Firth, L., Moore, K. 2008. Can the Internet Improve the Well-being of the Elderly? Ageing International, 32(1), 25-42.
34. Millis, K., Magliano, J., Wiemer-Hastings, K., Todaro, S., McNamara, D. S. 2007. Assessing and improving comprehension with latent semantic analysis. Handbook of latent semantic analysis, 207-225.
35. Monchaux, S., Amadieu, F., Chevalier, A., Mariné, C. 2015. Query strategies during information searching: Effects of prior domain knowledge and complexity of the information problems to be solved. Information Processing & Management, 51(5), 557-569.
36. Puerta Melguizo, M. C., Vidya, U., Can Oostendorp, H. 2012. Seeking information online: the influence of menu type, navigation path complexity and spatial ability on information gathering tasks. Behaviour & Information Technology, 31(1), 59-70.
37. Queen, T. L., Hess, T. M., Ennis, G. E., Dowd, K., Grühn, D. 2012. Information search and decision making: effects of age and complexity on strategy use. Psychology and Aging, 27(4), 817.
38. Raven, J. C., Court. J. H. 1998. Raven's progressive matrices and vocabulary scales. Oxford, UK: Oxford Psychologists Press.
39. Rieh, S. Y. 2006. Analysis of multiple query reformulations on the web: The interactive information retrieval context. Information Processing & Management, 42(3), 751-768.
40. Shute, S. J., Smith, P. J. 1993. Knowledge-based search tactics. Information Processing & Management, 29(1), 29-45.
41. Slegers, K., Van Boxtel, M. P., Jolles, J. 2012. Computer use in older adults: determinants and the relationship with cognitive change over a 6year episode. Computers in Human Behavior, 28(1), 1-10.

42. Sloan, M., Yang, H., Wang, J. 2015. A term-based methodology for query reformulation understanding. Information Retrieval Journal, 18(2), 145-165.
43. Strauss, E., Sherman, E. M., Spreen, O. 2006. A compendium of neuropsychological tests: Administration, norms, and commentary. American Chemical Society.
44. Teevan, J., Adar, E., Jones, R., Potts, M. A. 2007. Information re-retrieval: repeat queries in Yahoo's logs. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 151-158). ACM.
45. Turney, P. D. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In European Conference on Machine Learning (pp. 491-502). Springer Berlin Heidelberg.
46. Vakkari, P. 2003. Task-based information searching. Annual Review of Information Science and Technology, 37(1), 413-464.
47. Van Oostendorp, H., Madrid, R., Melguizo, M. C. P. 2009. The effect of menu type and task complexity on information retrieval performance. Ergonomics Open Journal, 2, 64-71.
48. Wang, J. J., Kaufman, A. S. 1993. Changes in fluid and crystallized intelligence across the 20-to 90-year age range on the K-BIT. Journal of Psychoeducational Assessment, 11(1), 29-37.
49. Wildemuth, B., Freund, L., G. Toms, E. 2014. Untangling search task complexity and difficulty in the context of interactive information retrieval studies. Journal of Documentation, 70(6), 1118-1140.
50. Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., Landauer, T. K. 1998. Learning from text: Matching readers and texts by latent semantic analysis. Discourse Processes, 25(2-3), 309-336.
51. Wood, E., De Pasquale, D., Mueller, J. L., Archer, K., Zivcakova, L., Walkey, K., Willoughby, T. 2016. Exploration of the relative contributions of domain knowledge and search expertise for conducting internet searches. The Reference Librarian, 57(3), 182-204.
52. Youmans, R. J., Bellows, B., Gonzalez, C. A., Sarbone, B., Figueroa, I. J. 2013. Designing for the wisdom of elders: age related differences in online search strategies. In International Conference on Universal Access in Human-Computer Interaction (pp. 240-249). Springer Berlin Heidelberg.