Choice Overload in Search Engine Use?

Pawitra Chiravirakul
Department of Computer Science,
University of Bath
Bath BA2 7AY, UK
p.chiravirakul@bath.ac.uk

Stephen J. Payne
Department of Computer Science,
University of Bath
Bath BA2 7AY, UK
s.j.payne@bath.ac.uk

ABSTRACT
Search engines typically return so many results that choosing from the list might be predicted to suffer from the effects of "choice overload". Preliminary work has reported just such an effect [12]. In this paper a series of three experiments was conducted to investigate the choice overload effect in search engine use. Participants were given search tasks and presented with either six or twenty-four returns to choose from. The results revealed that the choice behaviour was strongly influenced by the ranking of returns, and that choice satisfaction was affected by the number of options and the decision time. The main results, from the third experiment, showed that large sets of options yielded a positive effect on participants' satisfaction when they made a decision without time limit. When time was more strongly constrained, choices from small sets led to relatively higher satisfaction. Our studies show how user satisfaction with found information can be affected by processing strategies that are influenced by search engine design features.

Author Keywords
Choice satisfaction; search engines; decision behaviour.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Users of current web search engines are typically presented with a large number of returns after each query, ordered according to some obscure algorithm that makes it likely, to some extent, that valuable hits will appear early in the list. The length of the list means that users must choose returns to open and inspect without consideration of most of the alternatives.

As noted by Oulasvirta, Hukkinen & Schwartz [12], this situation may lead to "choice overload", i.e. negative psychological effects of being asked to choose from a large set of options (e.g. [1, 4, 9]), a phenomenon that has seen a great deal of empirical research, and some uneven conclusions since the classic work of Iyengar & Lepper [7], and the influential monograph by Schwartz [17].

Iyengar & Lepper's [7] first experiment involved food shopping. Either six different jams or twenty-four different jams were presented to buyers in an upscale retail outlet. After tasting, each buyer was given a discount coupon that could be used when buying a jam. These coupons were more used by those who tasted a chosen jam from a small set than those who had sampled from a large set.

In their second experiment [7] students were asked to choose an essay topic from either six topics or thirty topics and to write a two-page essay for additional course credit. The percentage of students who completed their essay and the quality of those essays were both higher for the students given the smaller set of options.

The final experiment [7] required participants to sample a chocolate that was chosen for them or was their own choice. Participants who could choose were presented with either six flavours or thirty flavours of chocolate. Having made the choice and eaten the chocolate, participants were then offered two options for payment, five dollars in cash or chocolates worth five dollars. Participants who had chosen from only six flavours were more likely to choose chocolates as compensation. Furthermore, participants who chose from the large set reported that the selection process was difficult and frustrating. Although at first the large number of options seemed attractive and enjoyable to consider, these participants felt regret and dissatisfaction with the final choice.

Such negative consequences of too many options have been called "The paradox of choice" [17], because in so many situations consumers and policy makers assume or report that more available choices is a Good Thing. Indeed, the empirical evidence concerning the choice overload effect is itself somewhat paradoxical, with as many studies reporting positive effects of large choice sets as those, since Iyengar & Lepper [7] that have been consistent with their findings. Scheibehenne, Greifeneder & Todd [16] reported a meta-analysis of fifty experiments in which number of alternatives was the major independent variable, combining the effect-size of measures such as unwillingness to choose or satisfaction with final choice. They found a mean effect size close to zero, and failed to identify any sufficient conditions for the choice overload effect.

One example of a positive effect of choice-set size was reported by Oppewal & Koelemeijer [11]. In their study, a set of either five or twelve flower photographs was sent to a florist's regular customers. The results demonstrated that more options had a positive effect, regardless of similarity of items and whether the options already contained a preferred item.

In many ways this study seems similar to [7], so the opposite finding illustrates the difficulty reported by Scheibhenne et al [15] in identifying important causal factors for choice overload effects. Perhaps one issue in the Oppewal & Koelemeijer [11] study is that neither set of options is "large enough" to create an overload effect, but finding an empirical basis for pinning down "large enough" is not yet possible.

It is a little easier to suggest *necessary* conditions for choice overload: in particular the non-familiarity of options seems necessary, because otherwise the decision-maker can fall back on simple recognition and preference judgments [15, 16] that make the size of the choice set less salient.

Returning to information foraging using a search engine, it seems clear that these necessary conditions for a choice overload effect are met – at least in many situations, especially novel searches, the returned choices will be unfamiliar to the searcher, and the choice set is often large by any standards.

However, some typical search engine design features may work against choice overload. First, as mentioned above the fact that the choice set is ordered (however unreliably relative to the searcher's needs) is a factor that may well affect the decision maker's response. If users are very confident that the ordering is a reliable guide to value, then the length of the list may seem irrelevant, and the tendency to select items from early in the list may mean that the list of options appears de facto small and manageable. This argument is supported by evidence that Google users typically select from very early in the list of returned pages [13] and further, by studies that show that judged relevance of documents decreases down lists of more than fifteen documents, even if the documents are in fact randomly ordered [6, 14].

Further, in many designs, the long list of returns is broken into separate pages, so that the length of the list is arguably less salient. Studies have shown that paginated lists lead to better performance and memory than uninterrupted, scrollable lists [2].

We conclude that on the basis of the existing empirical data it is not possible to predict, with any confidence, whether search engines will produce choice overload. For further guidance, we must look at explanations of the choice overload effect, and at the single published study of choice overload that we have found, which models a search engine scenario.

Three broad explanations have been put forward for choice overload. To simplify, we will couch these explanations in terms of a single effect – lower satisfaction with a chosen alternative when the set from which the alternative was chosen was larger rather than smaller. (We believe the arguments can readily be adapted to other dependent measures, such as motivation to consume.)

The first explanation, we will call process-product leakage. A large set of items to choose from is likely to make the process of choosing more problematic in several ways. Most obviously, if one assumes that time is constrained, it will mean less consideration must be given to each item (no consideration at all of some items in many cases of search engine use). More subtly, a bigger set of options makes it more likely that some items are hard to contrast (e.g. a second-best is likely to be closer in perceived quality to a best). If the decision process is difficult, the final choice may be viewed as unsatisfactory because the process that led to it is unsatisfactory in some way (and even if an experimental participant is responding to a question about their satisfaction with a consumed item, it seems plausible that their response may be less specific than the question's wording [15]).

The process-product leakage explanation predicts that decision time will be a moderator of choice overload effects, as has been proposed by Haynes [5]. Participants were asked to choose a prize to be entered in a drawing from a set of prize descriptions; either three prize descriptions or ten prize descriptions were presented. The result showed that under the time pressure participants felt the decision was difficult and reported less satisfaction, especially with a large set of options.

The second, related, explanation for choice overload is regret, i.e. regret about not consuming the un-chosen items [17]. The argument is that the more items rejected, the more likely that the decision-maker regrets doing without some of those items, i.e., a counterfactual, if-only, response negatively impacts the post-hoc evaluation of the chosen item.

Finally, the effect of choice-set size may be on expectations. If a choice set is larger, the decision maker may expect a better outcome, and this expectation works to set a standard against which post-choice comparisons are made. This explanation is in keeping with data that suggests participants have an a priori (pre-choice) more favourable reaction to large sets of alternatives (e.g. [7], experiment 3).

Intuitively, all three of these psychological processes seem plausible in the case of choosing information sources from the set of query returns by a search engine. On this basis, it seems reasonable to predict that choice overload is potentially an important issue for search engine design.

Oulasvirta, Hukkinen & Schwartz [12] conducted an experiment to investigate this prediction, using paper-based materials to model aspects of a typical search engine scenario. Because aspects of their study were an important guide to the design of our own studies, we will report them in some detail.

Participants were given three kinds of "realistic" search tasks: Simple facts, such as "Find out which country is located at the highest altitude"; Problems, with open-ended answers, such as "What determines the cost of railway

tickets in Europe?"; and Preferences, such as "Find your favourite novelist's homepage". For each search task a participant was provided with a printed page containing a query for that task and a set of results. Half of these pages were taken from Google, used Google formatting conventions, half used an invented search engine, with different terminology and layout but the same content. The main independent variable was the number of snippets returned, either 6 or 24. Independent variables were manipulated within-subjects. Participants had to select a single snippet from the returned set, within 30 seconds, and then (without consulting the actual web page) rate their satisfaction with the choice and their confidence that they had made the "correct" choice.

There were no significant differences between search engines, but a significant choice-overload effect, with participants reporting greater satisfaction and confidence when they chose from only 6 snippets. Post hoc analysis suggested that this effect was limited to simple fact and problem task types.

Several features of Oulasvirta et al's study limit the generalisation of the conclusions to real search engine use. Obviously, paper presentation is an approximate model in several ways, e.g., not allowing users to specify their own queries or to view any of the found web pages, both of which seem likely to be critical determinants of satisfaction in real scenarios. Further, limiting the judgments to snippets rather than the linked-to pages seems to us to make process-product leakage a very salient determinant of satisfaction. When no end product is actually experienced, what else can a satisfaction judgment rely on except process, or else the snippet information itself, the very basis of the choice. Furthermore, the time limit of 30 seconds seems quite severe, amounting to 1.25 seconds per item in the 24-item condition. Again, this seems likely to engender dissatisfaction with process. This account is supported by participants' report that when they chose from the six-option list they thought more carefully about their decision.

With these issues in mind, we sought to develop Oulasvirta et al's work with online experiments that more closely model real search-engine use, and which allow participants to choose and consult web pages from the sets of snippets their search returned before rating their satisfaction with those web pages.

EXPERIMENT 1: NUMBER OF OPTIONS AND ITERATION

Method

Design

In this experiment participants performed real-time Google searching, specifying their own search terms and choosing a web page to view before making any judgments. Only the Problem task type was used. Two separate sets of 10 questions were developed and utilised, but after preliminary analysis the question set was not treated as an independent variable in the main analyses. (Example problems are given in various method sections below, the full set is available from the authors.)

This experiment manipulated the number of options returned by Google: six options in a smaller set size condition (on a single Search Engine Results Page (SERP)) and twenty-four options in a large set size condition (on four SERPs, reached by pressing a page number button at the bottom of each page). Individual search terms were created by participants themselves. Half the participants were allowed to iterate or change their search terms after inspecting the returns whereas for the other half of the participants a web page had to be selected after the first search. Both independent variables were manipulated between subjects, giving a 2 (set-size: 6 vs. 24) x 2 (iterate vs. no-iterate) between-subjects design.

For each of 10 questions in their question set, participants were required to search for information using Google then choose a single web page from the SERPs. After each web page was chosen participants inspected the page and then rated their satisfaction with that page, their trust in its reliability and their judgment of the relevance of the web page for the current task.

Participants

Thirty-two participants were students and academic staff from University of Bath, aged between 23 and 45, with the average age of 29.9. All participants reported commonly using Google in their everyday life. Participants were told that the quality of selected websites would be evaluated in order to allocate a cash prize. After finishing the experiment, each participant was given a box of chocolates as a compensation for the participation.

Materials

HCI Browser [3], an open-source extension to Mozilla Firefox 3, was modified to collect data and to guide participants through tasks. All interactions were logged. Google was used by participants to complete each task: SERPs were altered using Google's API.

After each task was completed participants rated their selected website. All ratings were done on 10-point Likert scales. Additionally, there was a single open-ended question inviting a brief typed answer, giving a questioning protocol as follows:

- How satisfied are you with your selected page?
- Why? (open-ended question – not analysed in this paper)
- To what extent was the information provided by the website relevant to your task?
- To what extent do you trust the information provided by the website?

After the set of 10 search tasks was completed, a final questionnaire asked participants about the strategy they used to choose web pages (an open-ended question), overall

satisfaction for their selected web pages and overall satisfaction for the list of options that the search engine returned (10-point Likert scales).

Procedure

Participants completed the experiment individually in a laboratory. The experiment was divided into a training session and an experimental session. On arrival at the laboratory, participants were introduced to the HCI Browser interface and instructed that their general task would be to search for web pages that provided valuable information for a series of separate questions or tasks. Then participants completed one sample task in the training session ("Why did ancient Egyptians mummify their dead?"), with the experimenter available to offer guidance as required. After each participant confirmed that they understood the general task and the user interface, the experimental session started. Each task question was presented at the top of the browser window and persisted during the search and the subsequent ratings. No time limit was specified.

Results

After preliminary comparisons showing no difference on any dependent measure, the distinction between question group A versus B was not considered. The analyses reported below are 2x2, Set Size x Iteration between-subjects ANOVAs.

Position of Selected Web pages

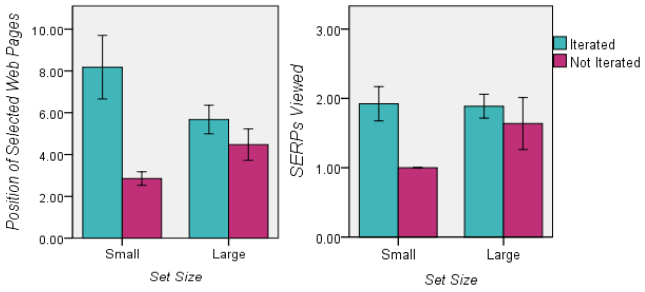

Figure 1: Mean position of selected web pages (left), and number of SERPs viewed before selecting (right). Error bars show standard errors. In the Iterated condition we accumulated the selected position and the SERPs viewed across iterations, e.g., choosing the second snippet after one iteration in the small set size would yield a Position of 8 and SERPs Viewed of 2.

Figure 1 (left) shows the mean ordinal position of selected web pages. Most of the selected web pages were chosen from near the top of the result lists in both the small set and the large set of options. Indeed there was no significant effect of Set Size on the ordinal position ($F(1,28)=0.2$, $p=.64$). However, participants who could iterate quite often changed their search term (see right hand panel, the mean of c. 2 SERPs viewed in the small Set Size means a mean of one iteration, i.e., two search-terms used). A simple main effect within the iteration condition showed a significant difference in selected web page positions between the small set size and the large set size ($F(1,28)=12.3$, $p=.002$, $\eta^2_p=.31$).

Figure 1 (right) shows the mean number of SERPs that participants viewed (out of a maximum of 4 separate SERPs for 24 items) before selecting a web page. Participants who were in the large set size condition only sometimes viewed beyond the first screen, and indeed there was no significant effect on the number of SERPs viewed across set size conditions ($F(1,28)=1.5$, $p=.22$). The number of SERPs that participants viewed was also updated by iteration, of course. The number of SERPs viewed between iteration and non-iteration condition were significantly different. ($F(1,28)=5.9$, $p=.021$, $\eta^2_p=.18$). Participants in the small set condition were more likely to iterate ($M=1.92$, $SD=.69$) than were participants in the large set condition ($M=1.26$, $SD=.36$). Independent t-test (among the 16 participants who could iterate) established this effect as significant: $t(14)=2.37$, $p=.038$.

Performance Time

Neither of the independent variables had a significant effect on time to select or view a web page, nor was there an interaction effect. The overall mean time to select a web page was 126 seconds; the overall mean time to view a web page before rating it was 34 seconds.

Rating of selected web pages

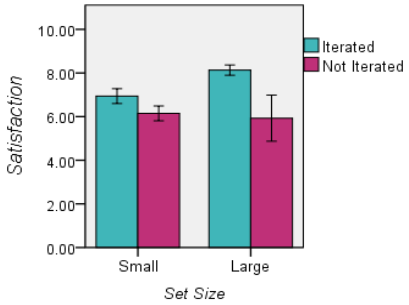

Figure 2: Mean satisfaction with selected web pages. Error bars show standard errors.

Figure 2 displays the participants' mean satisfaction with their chosen websites according to experimental condition. Participants given the large set size were more satisfied with their selected web pages but this effect failed to reach significance ($F(1,28)=3.9$, $p=.056$, $\eta^2_p=.12$). Participants who could iterate their search terms were reliably more satisfied with their selected web pages ($F(1,28)=9.9$, $p=.004$, $\eta^2_p=.26$). There was no significant interaction between Set Size and Iteration ($F(1,28)<1$).

No significant effects were found for Trust judgments (Trust was significantly correlated with Satisfaction, $r=.68$, $p<.001$). Relevance judgments behaved very similarly to Satisfaction judgments, and these judgments were very strongly correlated with each other ($r=.83$, $p<.001$). The only significant effect was a main effect of Iteration on judgments of relevance, ($F(1,28)=4.3$, $p=.046$, $\eta^2_p=.14$). All other main effects and interactions failed to reach significance, all Fs close to 1.

Overall rating of selected web pages

Participants were asked to rate their overall satisfaction for selected web pages and for lists of options returned at the end of the experiment and presumably reflecting memory for the overall experience across 10 tasks. Participants were marginally more satisfied with the selected web pages in the large set condition than were participants in the small set condition ($F(1,28)=3.8$, $p=.059$, $\eta^2_p=.12$). In addition participants who could iterate their search term were reliably more satisfied with their selected web pages ($F(1,28)=7.9$, $p=.009$, $\eta^2_p=.22$). The interaction effect between Set Size and Iteration condition was not significant ($F(1,28)<1$). There was no effect of Set Size on overall satisfaction with the lists of options returned ($F(1,28)=2.1$, $p=.155$).

Discussion

This experiment found no evidence for a choice overload effect. Indeed, the effect of Set Size on satisfaction was in the opposite direction, although failed to reach significance. This result contrasts with the findings from Oulasvirta et al [12], despite the overlap in task context and the identical set sizes in the two experiments. We believe that some of the issues reviewed above may explain the contrast. First, Oulasvirta et al asked participants to report satisfaction with a chosen snippet without actually consulting the web page to which the snippet related. This, it seems to us, makes it even more likely that satisfaction judgments will be judgments of process (because snippets contain far less information and therefore are less differentiated than are full web pages). Second, Oulasvirta et al imposed a very strict time limit, which again, as argued above, is likely to make for an unsatisfactory decision process in the case of large sets: participants in the large set condition in their experiment were allowed just over 1 second per snippet to make their choice. Third, our large sets were paginated, whereas Oulasvirta et al presented them as a single list. Pagination makes the size of the set less salient and more manageable.

The other very important difference in our experimental method was that participants constructed their own search terms, and half the participants were allowed to iterate over search terms if dissatisfied with the initial set of returns.

It is unsurprising that there was a main effect of Iteration on satisfaction judgments. This shows that participants were able to judge the quality of web pages from the snippets and were able to improve their search terms if initial results were disappointing. Participants who received a small set of options iterated more than did those who received a large set of options, which suggests that the small option set was more likely to be judged as unsatisfactory, adding additional support to the conclusion that, in this experiment, larger sets of returned snippets were judged more satisfactory than smaller sets.

It is striking that most participants in the large option set condition usually selected a web page from the first page, often without bothering to iterate their search terms or browse options on later result pages. This is consistent with the finding from [13], which showed that users' choice of a particular web page was mostly based on its position on the result list. It is also consistent with our suggestion above that the size of the returned set was not psychologically salient (although of course this suggestion cannot explain the advantage we observed for the large set).

EXPERIMENT 2: NUMBER OF OPTIONS AND RANKING

Method

Design

The second experiment was designed to further investigate the role of ordering of the choice set. Participants' expectation about the ranking algorithm was manipulated as a within-subjects independent variable. Participants were informed that one of three processes was used to rank the returns of a search (although in fact, in all cases the Google ranking was preserved, as explained to participants during post-experiment debriefing):

- Expert ranking: Participants were informed that the web pages linked to in the returns-list had been ranked by an expert in the appropriate topic area, according to how well they answer the question.
- Novice ranking: Participants were informed that the web pages had been ranked according to how well they answer the question, but by someone with no particular knowledge of the fields.
- Random ranking: The order of the list of links had been randomised.

To make this manipulation plausible we constrained the design of the study so that pre-specified search terms were used and indicated to participants (in this respect, this second study moves closer to the study by Oulasvirta et al [12]). Search terms were abbreviated versions of the problem specification. For example, when the task question was "What determines the cost of living in the UK?" the search term was "cost of living in UK". Consequently, there was no iterated search in this experiment.

Three separate question sets were used (four questions per set). Participants received all three question sets, with each question set being associated with a particular ranking type. Ranking types were assigned to question sets, and ordered so that each ranking type was associated with each question set and appeared in each ordinal position equally across participants.

For each of 12 tasks, participants were required to select a single web page that provided valuable information from the list of search results. After each web page was selected participants consulted the web page and then rated their satisfaction for the selected web page, their trust in its

content, the relevance of the page for the task, and their familiarity with the task question. This last question was added as a check that one of the main necessary factors for choice overload was not compromised by participants' prior experience.

One final change to the procedure was a minor redesign to the user interface, so that the page numbers of search returns beyond the first page in the large set was more salient (the page number was increased in size and each returns page was additionally labelled with "This page is the <Nth> page of 4."

Participants

Twenty-four participants from University of Bath were recruited via online ad and posters on notice boards. No participant had taken part in Experiment 1. The participants aged between 19 and 39, with an average of 27.9. Each participant was paid five pounds in cash for participation. Participants were instructed that the quality of the selected web pages would be evaluated in order to allocate the prizes. Three cash prizes were used as a motivation in searching for high quality web pages.

Materials

The HCI browser, as used in Experiment 1, was modified, as above, to make the size of larger sets more salient. A single question about familiarity was added to the 10-point Likert ratings requested after each web page was selected, i.e., "To what extent is the task question familiar to you".

Procedure

The procedure was unchanged from Experiment 1 except as required by the changes to the Experimental Design. Three types of ranking algorithm were described to participants in written instructions and questions were answered by the experimenter. During the main part of the experiment, participants were informed by a screen message about the type of ranking used for each task: the browser window displayed the task question, the search term and the ranking type at the top of the window. This information persisted during the search task and the subsequent ratings.

Results

The main analyses reported below are 2x3, Set Size (6 vs. 24) x Ranking Type (expert vs. novice vs. random) mixed ANOVAs.

Position of Selected Web pages

Figure 3 (left) shows the mean ordinal position of selected web pages. The results indicate that participants selected web pages from further down the list when they were provided with the large set of options. The main effect of Set Size on the ordinal position of selected web pages was significant ($F(1,22)=11.2$, $p=.003$, $\eta^2_p=.34$). In the Random ranking condition, the selected web pages were located considerably further down the search result list. The main effect of Ranking Type was significant ($F(2,21)=4.9$, $p=.018$, $\eta^2_p=.32$). Further, the interaction effect between Ranking Type and Set Size was significant ($F(2,21)=4.1$, $p=.031$, $\eta^2_p=.28$).

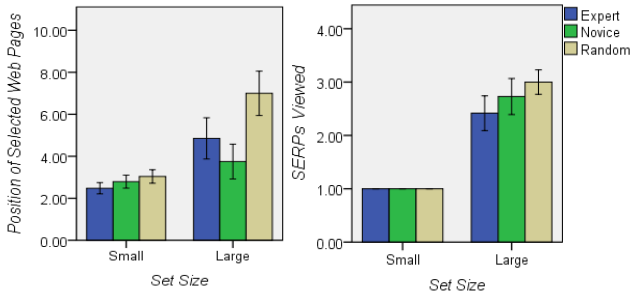

Figure 3: Position of selected web pages (left), and number of SERPs viewed before selecting (right). Error bars show standard errors.

Although it remains true that in the large set condition the majority of the selected web pages were located in the first SERP (Figure 3, left), participants mostly checked the search results provided in at least one other SERP before selecting the web page (Figure 3, right). According to a one-sample t-test, the number of SERPs viewed in the large Set Size is significantly greater than 1 ($t(11)=6.4$, $p<.001$). The main effect of Ranking Type on the number of SERPs viewed in the large-set condition was not significant ($F(2,22)=2.98$, $p=.07$).

Performance Time

Neither of the independent variables had a significant effect on time to select or view a web page, nor was there an interaction effect. The overall mean time to select a web page was 67 seconds; the overall mean time to view a web page before rating it was 35 seconds.

Rating of selected web pages

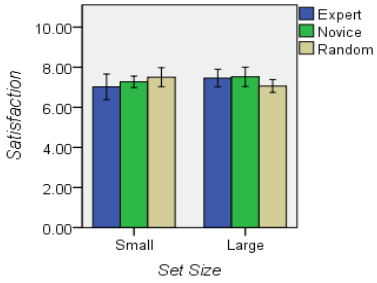

Figure 4: Mean satisfaction with selected web pages. Error bars show standard errors.

Figure 4 shows participants' rated satisfaction with chosen web pages. There was no significant main effect of Ranking Type or Set Size and no interaction effect (all Fs < 1).

Similarly, there were no main effects or interactions on judgments of Trust, Relevance, and Familiarity, all Fs close to 1. Trust was significantly correlated with Satisfaction ($r=.52$, $p<.001$). Relevance judgment was even more strongly correlated with Satisfaction ($r=.87$, $p<.001$). Familiarity judgments averaged around 5, suggesting that participants were not over-familiar with the choices they

were asked to make. The correlation between Familiarity and Satisfaction was not significant (r=.19).

Overall Rating of selected web pages

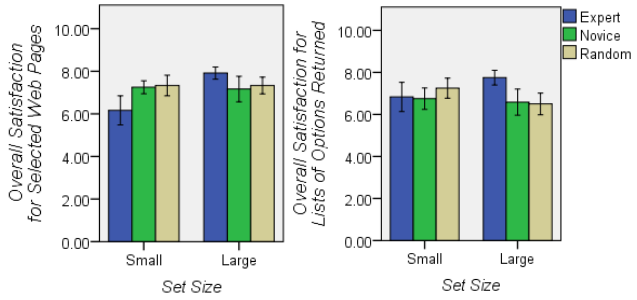

Figure 5: Overall satisfaction with selected web pages (left), Overall satisfaction with lists of options returned by search engine (right). Error bars show standard errors.

Figure 5 shows participants' satisfaction ratings at the end of each task-block (i.e. after each ranking type). There was no significant main effect of Ranking Type or of Set Size, but the interaction effect was reliable ($F(2,21)=3.9$, $p=.036$, $\eta^2_p=.27$).

The interaction effect between Ranking Type and Set Size on overall satisfaction for lists of options returned by search engine was not significant ($F(2,21)=1.98$, $p=.16$). Neither Ranking Type nor Set Size yielded the main effect on overall satisfaction for lists of options returned (Fs close to 1). Participants did not perceive any difference between the overall result lists from different ranking algorithms.

Discussion

Participants' selection behaviour confirms that they are influenced by what they believe to be true about the ranking of returns. Participants do not typically simply accept the Expert ranking (in that condition, the average rank of the chosen website is 2 in the small set and about 4 in the large set). Nevertheless, participants choose form further down the set of returns when they believe the ranking is by a novice or random, and when more returns are available (i.e., the Set Size is large).

As in Experiment 1, there is no hint of a choice overload effect in the satisfaction ratings of chosen web pages. This is true even when participants believe the ranking of returns is random, showing that participants' belief that the rank ordering is helpful is not necessary to explain the relatively positive reaction to large set sizes. (That the rank ordering IS typically helpful may still be important in this respect.) Participants in the Random, large Set-Size condition certainly process the size of the set (often inspecting all four SERPs and choosing on average the web page ranked seventh), but this does not lead to significantly less satisfaction. As it happens, of course, these participants are still choosing early items more than late items, and this in itself may mitigate any negative effects of the large set.

This experiment does not replicate the marginal satisfaction advantage of large set size reported in Experiment 1, except when overall ratings of chosen web pages in a task block are considered, in the Expert ranking condition (see Figure 5).

One important difference that remains between our first two experiments and the typical choice overload experiment, and especially Oulasvirta et al [12], is that our participants were not given any time constraints. We have suggested that such constraints may be operational in choice overload contexts, because they lead to dissatisfaction with choice process and this "leaks" into judgments of chosen items. The third experiment explores this issue as well as seeking further evidence concerning the importance of the ranking of returns on participants' search behaviour and subjective judgments.

EXPERIMENT 3: NUMBER OF OPTIONS, RANKING AND TIME PRESSURE

Method

Design

Time limitation was manipulated in this experiment. Participants were given either 45 seconds or 90 seconds to finish each task.

To further study the effect of ranking, participants were provided with search results ordered by one of the two ranking algorithms, explained as follows.

- Google ranking: All the web pages that are linked to in the returns-list have been ranked by Google.
- Random ranking: The order of the list of links has been randomised.

In the Google ranking condition, SERPs were generated by Google. Broken links were eliminated from the lists to reduce noise in the experiment. In the Random ranking condition, the search results in SERPs were those from Google-ordered list. However, the order of those search results was in fact random (a single random order was used for each task/set-size).

As in Experiments 1 and 2 a between-subjects independent variable was the Set Size of a result list returned by the search engine, which was either 6 items (on a single SERP) or 24 items (on four separate SERPs, each reached by pressing a page number button at the bottom of each SERP).

There were 16 tasks in total, which were divided into four blocks (four tasks per block). Each block had a different combination of Ranking and Time Limit. Half the participants performed the two Google blocks before the two Random blocks, whereas for the other half this was reversed. Similarly, half the participants performed the 90s block before the 45s block for each Ranking, whereas for the other half this was reversed.

After each web page was selected participants consulted the web page and then rated their satisfaction for the selected web page, their trust in its reliability, the relevance of the web page for the task, and their confidence that their selected web page was the best in the search result list (10-point Likert Scales).

The use of a confidence rating is novel in this Experiment. It seemed to us that it might tap satisfaction with the decision making process, e.g., confidence would be low if only poor consideration of alternatives was possible. The additional question was "How confident are you that your selected website is the best in the set you choose from?"

After each single session was completed, a questionnaire asked participants about the strategy they used to choose web pages (an open-ended question), overall satisfaction for their selected pages and overall satisfaction for the list of pages that the search engine returned (10-point Likert scales).

Participants

Twenty-four participants were students and staff from University of Bath, who were recruited via online advertisement and posters on notice boards. No participant had taken part in the earlier studies. The participants were aged between 21 and 41, with an average of 29.7. All participants use a search engine in their everyday life. They were given five pounds in cash as compensation for their time. Participants were motivated to search for a good quality web page in order to compete for two cash prizes of 30 pounds.

Materials

The version of the HCI browser used in Experiment 2 was altered only by the addition of a digital clock in the top right-hand corner, counting down the time remaining for each task.

Procedure

The procedure was unchanged from Experiments 1 and 2 except as required by the changes to the Experimental Design. Participants had two training tasks to complete, so as to experience both time conditions. The questions for the training were "Why did ancient Egyptians mummify their dead?" and "How does economics affect our daily life?" with time limit of 45 seconds and 90 seconds respectively. Pre-specified search terms, which were abbreviated versions of the problem specification, were used and indicated to participants. For example, when the task question was "Why is meditation sometimes recommended for managing stress?" the search term was "meditation for managing stress".

The browser window displayed a task question, a search term, a ranking type and a timer at the top of the window. This information persisted during the search task and the subsequent ratings. During each task a beep sounded and a visual sign appeared if and when ten seconds remained.

Results

The main analyses reported below are 2x2x2, Set Size (6 vs. 24) x Ranking Type (Google vs. Random) x Time Limit (45s vs. 90s) mixed ANOVAs.

Position of Selected Web pages

Figure 6 (left) shows the mean ordinal position of selected web pages. There were significant main effects of Set Size ($F(1,22)= 20.3$, $p=.000$, $\eta^2_p=.48$), Ranking Type ($F(1,22)= 13.5$, $p=.001$, $\eta^2_p=.38$), and Time Limit ($F(1,22)=5.14$, $p=.034$, $\eta^2_p=.19$). The interaction effect between Time Limit and Set Size on the position of the selected web pages was significant ($F(1,22)=6.7$, $p=.017$, $\eta^2_p=.23$). The interaction effect between Ranking Type and Set Size on the position of selected web pages was also significant ($F(1,22)=12.9$, $p=.002$, $\eta^2_p=.37$). Participants in the large set condition selected web pages from further down the list than did participants in the small set condition, especially when they selected from the Random-ordered SERPs and when they had 90 seconds available to make their selection.

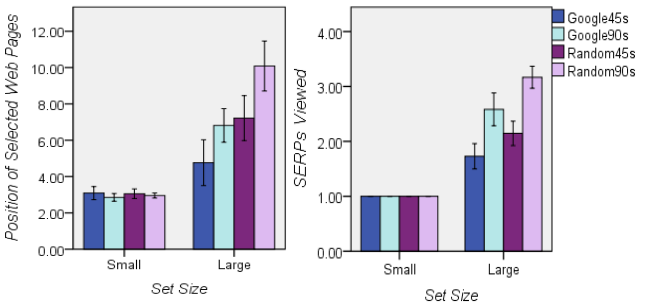

Figure 6: Position of selected web pages (left), and number of SERPs viewed before selecting (right). Error bars show standard errors.

According to a one-sample t-test, the number of SERPs viewed in the large set is significantly greater than 1 ($t(11)=12.76$, $p<.001$). The ANOVAs revealed that the main effects of the independent variables on the number of SERPs viewed in the large-set condition were significant, Ranking Type ($F(1,22)=6.77$, $p=.025$, $\eta^2_p=.38$), and Time Limit ($F(1,22)=25.34$, $p<.001$, $\eta^2_p=.68$). There was no significant interaction effect between Ranking Type and Time Limit ($F<1$). Participants in the large set condition browsed through more SERPs when they had to select a web page from the Random-ordered list. With the time limit of 90 seconds, participants in the large set browsed significantly more SERPs more than they did when they had 45 seconds to select a web page.

Rating of selected web pages

Figure 7, left, shows judged satisfaction with chosen web pages in each cell of the experiment. The main effect of Ranking Type on the participants' mean satisfaction was significant ($F(1,22)=6.31$, $p=.02$, $\eta^2_p=.22$). Participants were more satisfied with pages from Google-ordered lists than from Random-ordered lists. There was no significant main effect of either Time Limit or Set Size on the participants' mean satisfaction ($F(1,22)=.75$, $p=.39$ and

$F(1,22)=.05$, $p=.81$ respectively). However, the interaction effect between Time Limit and Set Size was significant ($F(1,22)=5.13$, $p=.034$, $\eta^2_p=.19$). In both the Google and the Random conditions, satisfaction in small set sizes relative to large set sizes increases as time pressure increases.

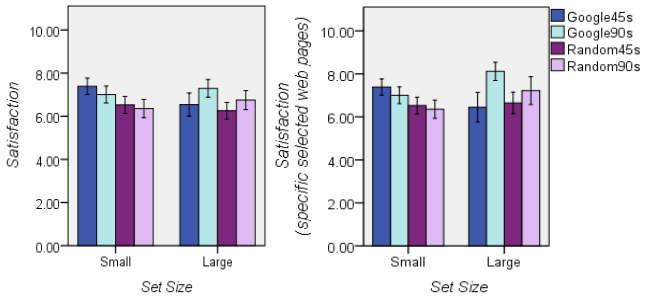

Figure 7: Mean satisfaction with selected web pages (left), mean satisfaction with selected web pages that are present in the small set (right). Error bars show standard errors.

The data were pruned so that only chosen web pages that are present in the small set are considered in all cells of the experiment (otherwise, reduced satisfaction in selections from the large set could be a peculiarity of the actual web pages that were chosen). Every participant contributed at least one such judgment in all large set conditions (Figure 7, right). There was now no significant main effect of Ranking Type on the participants' mean satisfaction ($F(1,20)=3.42$, $p=.079$). Nevertheless, importantly, the significant interaction effect between Set Size and Time Limit on participants' mean satisfaction ($F(1,20)=10.04$, $p=.005$, $\eta^2_p=.33$) was maintained.

Ranking Type, Time Limit and Set Size did not have significant effects on Trust judgments. There was no significant interaction effect, all Fs close to 1 (Trust was significantly correlated with Satisfaction $r=.60$, $p<.001$). Relevance judgments was strongly correlated to Satisfaction ($r=.78$, $p<.001$). The main effect of Ranking Type on Relevance was significant ($F(1,22)=6.18$, $p=.021$, $\eta^2_p=.22$). Participants thought that their selected web pages in Google-ordered lists were more relevant to the task questions than their selected web pages in Random-ordered lists. There was no main effect of Time Limit or Set Size and no interaction effect on Relevance (all Fs<1).

There was no significant main effect or interaction effect on Confidence judgments, all Fs close to 1. However Confidence was significantly correlated with Satisfaction ($r=0.86$, $p<.001$). This indicated that if the participant was satisfied with the selected option he/she was more likely to be confident that the selected option was the best in the set that available.

Overall Rating of selected web pages

A main effect of Ranking Type on overall satisfaction for selected pages was significant ($F(1,22)=11.42$, $p=.003$, $\eta^2_p=.34$). In both large and small set sizes, participants were significantly more satisfied with their selected web pages from Google-ordered lists more than from Random-ordered lists. However, the main effect of Set Size was not reliable ($F(1,22)=.032$, $p=.86$). There was no significant interaction effect.

A main effect of Ranking Type on overall satisfaction for lists of options returned by the search engine was significant ($F(1,22)=23.38$, $p<.000$, $\eta^2_p=.52$). Participants were satisfied with the result lists ranked by Google more than Random lists. There was no significant main effect of Set Size or interaction effect. (both Fs<1)

Discussion

This experiment suggests a crucial role for time pressure in determining the effect of set size on satisfaction with the results of a choice process. When time is more strongly constrained, choices from small sets led to relatively more satisfaction.

The main effect of Ranking Type on participants' satisfaction was significant, yet its interactions with other independent variables were insignificant. Participants were more satisfied with the selected web pages from Google-ordered lists than from Random-ordered lists regardless of time pressure or the number of options provided. This may simply be because better web pages were selected, enabled by Google's ranking.

GENERAL DISCUSSION

In contrast with [12], there is no suggestion from our studies that choice overload will typically affect satisfaction with web pages found through keyword search. This is not just a matter of failure to replicate through null effects – in most conditions of our studies, the tendency was for larger numbers of returns to be associated with better subjective outcomes, although this tendency was only marginally reliable for judgements of individual web pages in Experiment 1 and for overall satisfaction in Experiment 2.

We had predicted that the ordering of search engine returns would mitigate choice overload effects. We confirmed that ordering does indeed affect selection processes. When participants believed the results had been ordered by Experts they were more likely to choose a web page from earlier in the list, compared to a Novice ranking or Random ordering – even when the order of items was the same. Similarly, participants chose earlier items from Google-ranked lists than from randomly ordered lists (Expt. 3). These ordering effects confirm and extend previous work which has shown that Google users typically select from the first page of returns [13] and that order of a document set affects judgments of relevance [14]. However, these effects of ordering do not seem to affect Choice Overload – the effects of set size on subjective judgments did not interact with beliefs about order, or actual order of search returns.

Instead, we have found evidence that time pressure is an important determinant of choice overload. In Experiment 3, the relative satisfaction with web pages chosen from larger

versus smaller sets interacted significantly with the time available to choose.

Our findings challenge search engine design to be sensitive to time pressure. Of course the way time pressure affects users' needs and satisfaction is likely to vary across search contexts, how important it is to find excellent rather than good-enough sources, etc. One important limit of our studies is that participants were asked to choose single web pages. In many contexts multiple information sources will be sought, and in that case we suspect that search results should be diversified; search results from different domain categories could be presented early on the result list, possibly in hierarchical structure. Indeed such a design illustrates one way in which search interfaces could be sensitive to time pressure: users could consider in turn each specific category where the number of options is considerably fewer than the entire result list [see also 12]. On the other hand, with less time pressure, a user could browse across categories to explore and make comparisons between options.

The role of time pressure is also important for theoretical reasons, because it supports the "process-product leakage" account of choice overload, rather than regret or expectation-setting. Our studies confirm that interface design can affect process and strategy in such a way that user satisfaction with retrieved information is affected.

ACKNOWLEDGMENTS

We thank all the volunteers, and all publications support and staff, who wrote and provided helpful comments on previous versions of this document.

REFERENCES

1. Arunachalam, B., Henneberry, S. R., Lusk, J. L., & Norwood, F. B. An empirical investigation into the excessive-choice effect. *American journal of agricultural economics, 91* (2009), 810-825.
2. Bernard, M., Baker, R., & Fernandez, M. Paging vs. scrolling: Looking for the best way to present search results. *Usability News, 4* (2002).
3. Capra, R. HCI browser: A tool for studying web search behavior. *Proceedings of the American Society for Information Science and Technology, 47* (2001), 1-2.
4. Fasolo, B., McClelland, G. H., & Todd, P. M. Escaping the tyranny of choice: When fewer attributes make choice easier. *Marketing Theory, 7* (2007), 13-26.
5. Haynes, G. A. Testing the boundaries of the choice overload phenomenon: The effect of number of options and time pressure on decision difficulty and satisfaction. *Psychology & Marketing, 26* (2009), 204-212.
6. Huang, M. H., & Wang, H. Y. The influence of document presentation order and number of documents judged on users' judgments of relevance. *Journal of the American Society for Information Science and Technology, 55* (2004), 970-979.
7. Iyengar, S. S., & Lepper, M. R. When choice is demotivating: Can one desire too much of a good thing?. *Journal of personality and social psychology, 79* (2000), 995-1006.
8. Iyengar, S. S., Wells, R. E., & Schwartz, B. Doing Better but Feeling Worse Looking for the "Best" Job Undermines Satisfaction. *Psychological Science, 17* (2006), 143-150.
9. Loewenstein, G. Is more choice always better?. *Social Security Brief, 7* (1999), 1-8.
10. Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., Granka, L., Pellacini, F., & Pan, B. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology, 59* (2008), 1041-1052.
11. Oppewal, H., & Koelemeijer, K. More choice is better: Effects of assortment size and composition on assortment evaluation. *International Journal of Research in Marketing, 22* (2005), 45-60.
12. Oulasvirta, A., Hukkinen, J. P., & Schwartz, B. When more is less: the paradox of choice in search engine use. *Proc. SIGIR 2009.* ACM Press (2009), 516-523.
13. Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication, 12* (2007), 801-823.
14. Purgalis, Parker, L. M., & Johnson, R. E. Does order of presentation affect users' judgment of documents?. *Journal of the American Society for Information Science, 41* (1990), 493-494.
15. Scheibehenne, B., Greifeneder, R., & Todd, P. M. What moderates the too-much-choice effect?. *Psychology & Marketing, 26* (2009), 229-253.
16. Scheibehenne, B., Greifeneder, R., & Todd, P. M. Can There Ever Be Too Many Options? A Meta-Analytic Review of Choice Overload. *Journal of Consumer Research, 37* (2010), 409-425.
17. Schwartz, B. The paradox of choice: Why less is more. *New York: Ecco* (2004).