Visualisation to Aid Decision-Making for Time-Quality Tradeoffs In Search

Ryan Burton
University of Michigan School of Information
Ann Arbor, Michigan, USA
ryb@umich.edu

Kevyn Collins-Thompson
University of Michigan School of Information
Ann Arbor, Michigan, USA
kevynct@umich.edu

ABSTRACT

We present the results of two studies for an experimental setup that offers a tradeoff between time and result quality, and gives participants a visualisation of the expected quality of results in an optional sidebar before they begin their task. We investigate the effects of the type of visualisation we show to participants of a crowdsourced survey, and find that the type of visualisation presented had no significant effects on their perceptions of quality, but observed that there was value in offering a visualisation versus none at all. We also performed a pilot study that fixes the visualisation type and manipulates the estimation of future quality such that it conveys either an accurate estimate of quality, an overestimate, or an underestimate of quality. The results of clickstream and gaze tracking data analysis point towards potential effects on usage and attention to the optional sidebar, suggesting the utility of a full-scale study.

CCS CONCEPTS

• **Information systems** → *Search interfaces.*

KEYWORDS

quality, search user interfaces, information visualization, user behavior

ACM Reference Format:
Ryan Burton and Kevyn Collins-Thompson. 2022. Visualisation to Aid Decision-Making for Time-Quality Tradeoffs In Search. In *IQIIR 2022: Workshop on Information Quality in Information Interaction Retrieval, March 14, 2022, Regensburg, Germany.* , 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

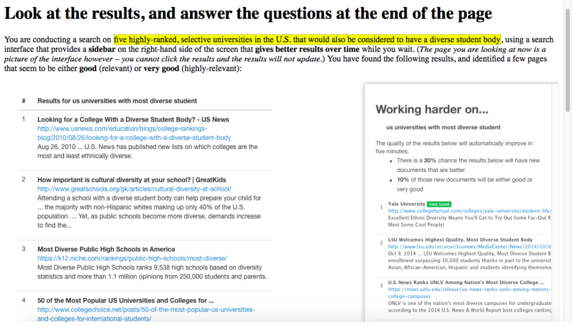
In the past decade, the concept of *algorithmic sensemaking* [18] has become a topic of increased research, particularly in the context of recommender systems [16]. Particularly, it has gained increased salience especially as algorithms have increasingly encroached on our interactions with computer systems. Previously, relatively "unsophisticated" users of shopping sites, content streaming services, and social networking services are exposed to the output of sophisticated and often opaque ranking and filtering algorithms. Not only is this a factor in terms of misaligned incentives between companies and individuals when used for product recommendation, but in the context of search and social media, there is also a risk of reinforcing bias and promoting misinformation. With the stakes increasing and the number of users exposed to sophisticated algorithms growing, there is value in helping users to form better mental models of the systems they use.

For designers of systems, there is also a very practical benefit of incorporating this idea. Explanations of output can build trust in recommender systems [14] and can be more persuasive, increasing the effectiveness of a system [8]. A study by Yeomans et al. [18] demonstrated the people are less averse to using a recommender system when an explanation is included. Even as experts increasingly use machine learning and artificial intelligence to assist them in decision-making, explanations have also been explored in conjunction with other classes of algorithms such as classification [13] and clustering [12].

A concept related to algorithmic sensemaking is that of *operational transparency* [5]. Buell and Norton [5] introduced this concept to motivate the *illusion of labour*, where a system may present a signal such as a spinner or progress bar to indicate that work more work is being performed on the backend of a system than is actually being performed. Although this kind of deception is not necessary, experiments by the authors show that merely presenting this illusion of labour increases the perceived value of a system. In a study by Tsekouras et al. [17], it was further shown that for product recommendation agents, users' perceived quality of the agent increased when the perceived user effort decreases and the perceived agent effort increases. In the interest of applying such benefits to search, we incorporated these principles into a system that we designed, which revolves around search with an interface that gives users the option of using a sidebar on the result page of a conventional Web search engine. In our setup, the sidebar presents additional, high-quality results that are relevant to the task that participants are currently focused on. Unlike the static main results for a query, the sidebar results are updated progressively over time as the system 'works harder' to find better results, throughout the duration of the task. We incorporated operational transparency into our system design by presenting feedback on how the sidebar is dynamically working to deliver these results. We believe that, in addition to other benefits, this signal will engender trust and encourage the use of our optional sidebar.

With this said, our primary concern in this particular paper is with users' perceptions of the expectation of quality of the results in the sidebar, and how their behaviours change as a result. To that end, our aim was to present users with a visualisation reflecting the sidebar's future expected performance over the course of a

IQIIR 2022, March 14, 2022, Regensburg, Germany
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



Look at the results, and answer the questions at the end of the page

You are conducting a search on five highly-ranked, selective universities in the U.S. that would also be considered to have a diverse student body, using a search interface that provides a **sidebar** on the right-hand side of the screen that **gives better results over time** while you wait. (*The page you are looking at now is a picture of the interface however – you cannot click the results and the results will not update*.) You have found the following results, and identified a few pages that seem to be either **good** (relevant) or **very good** (highly-relevant):

#	Results for us universities with most diverse student
1	**Looking for a College With a Diverse Student Body? - US News** http://www.usnews.com/education/blogs/college-rankings-blog/2010/08/26/looking-for-a-college-with-a-diverse-student-body Aug 26, 2010 ... U.S. News has published new lists on which colleges are the most and least ethnically diverse.
2	**How important is cultural diversity at your school? | GreatKids** http://www.greatschools.org/gk/articles/cultural-diversity-at-school/ Attending a school with a diverse student body can help prepare your child for ... the majority with non-Hispanic whites making up only 40% of the U.S. population. ... Yet, as public schools become more diverse, demands increase to find the...
3	**Most Diverse Public High Schools in America** https://k12.niche.com/rankings/public-high-schools/most-diverse/ Most Diverse Public High Schools ranks 9,538 high schools based on diversity statistics and more than 1.1 million opinions from 250,000 students and parents.
4	**50 of the Most Popular US Universities and Colleges for ...** http://www.collegechoice.net/posts/50-of-the-most-popular-us-universities-and-colleges-for-international-students/

Working harder on...

us universities with most diverse student

The quality of the results below will automatically improve in five minutes:

- There is a **30%** chance the results below will have new documents that are better
- **10%** of those new documents will be either good or very good

1 **Yale University** Very Good
http://www.collegefactual.com/colleges/yale-university/student-life/
Excellent Ethnic Diversity Means You'll Get to Try Out Some Far-Out R
Meet Some Cool People)

2 **LSU Welcomes Highest Quality, Most Diverse Student Body**
http://www.lsu.edu/ur/ocur/lsunews/MediaCenter/News/2014/10/it
Oct 9, 2014 ... LSU Welcomes Highest Quality, Most Diverse Student B
enrollment surpassing 30,000 students thanks in part to the universit
Asian, African-American, Hispanic and students identifying themselve

3 **U.S. News Ranks UNLV Among Nation's Most Diverse College ...**
https://news.unlv.edu/release/us-news-ranks-unlv-among-nations-college-campuses
UNLV is one of the nation's most diverse campuses for undergraduate
according to the 2014 U.S. News & World Report best colleges ranking

Figure 1: A screenshot of our Mechanical Turk questionnaire

search task, and to observe their choices of whether to wait for results in the sidebar, or to be more proactive and put in the immediate effort of finding results on their own. We will begin by presenting the results of a crowdsourced study to evaluate different types of visualisations in Section 2, and then show the results of a pilot experiment in Section 3 that uses one type of visualisation to investigate clickstream and eyetracking data.

2 COMPARISON OF VISUALIZATION TYPES FOR ESTIMATED RESULT QUALITY

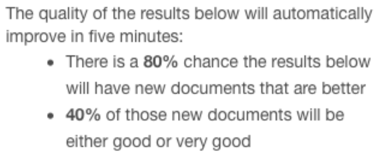
Before performing our user study, we sought to determine the effectiveness of various visualisation methods for conveying information on the state of the system, which incorporates an asynchronous 'slow' search process that continues to work in the background [6, 15]. In particular we focus on the system's estimate of the quality of the search results and how this quality is likely to evolve over time. For this, we employ a Bayesian framework that estimates a user's belief that 1. the system will provide better documents in its sidebar than on the main search page (probability of a 'win'), and 2. the expected utility of the better results (expected value given a 'win').

We conducted an online experiment with Amazon Mechanical Turk. In our setup, we presented an interactive HTML mockup of our interface with search results and sidebar. A screenshot of the interface can be seen in Figure 1.

In the description of the task, we told participants that they must read a brief tutorial with a screenshot that introduces a description of the sidebar and what it enables (namely, that it finds better results in the background during a search task and that the results are automatically updated over time). We recorded whether a user viewed the tutorial, which serves as a proxy for whether they read the instructions. Participants were paid 8 cents for completing the questionnaire.

We randomly showed each user one of the following choices of visualization in the sidebar:

(1) No visualization/information – we give the following explanation in place of the visualisation: "The quality of the results below will automatically improve over time. You may wait longer for better results."
(2) Numeric point estimates of success (see Figure 2)
(3) A CDF bar plot of success (see Figure 3)
(4) A CDF dot plot of success (see Figure 4)



The quality of the results below will automatically improve in five minutes:

- There is a **80%** chance the results below will have new documents that are better
- **40%** of those new documents will be either good or very good

Figure 2: Point estimates of ranking effectiveness at a particular time in the future (five minutes).

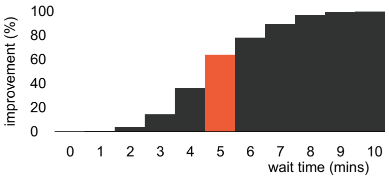


Figure 3: A cumulative bar chart showing the improvement in ranking effectiveness over time, with the same time as shown in the point estimate (five minutes; Figure 2) highlighted in orange.

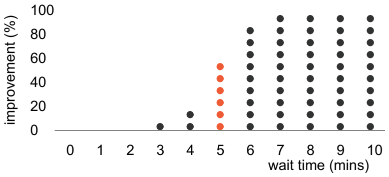


Figure 4: A cumulative dot plot showing the improvement in ranking effectiveness over time, with the same time as shown in the point estimate (five minutes; Figure 2) highlighted in orange. The dots are discrete, countable, and serve as frequency framing.

In the main search page, we highlight what we consider "good" and "very good" results. We also present highlighted good results in the sidebar. The results and relevance judgements are simulated and fixed.

We then asked the participant:

- *How likely is it that you will be able to find better results five minutes from now?* The participant is expected to read the information from the visualisation to answer this question correctly.
- *Just by looking at the results, are the links in the sidebar worth exploring **now**?*
- *Assuming you had to wait for **one** additional minute for better results in the sidebar, would you?*
- *Assuming you had to wait for **two** additional minutes for better results in the sidebar, would you?*
- *Assuming you had to wait for **five** additional minutes for better results in the sidebar, would you?*
- *How often do you perform searches online? Daily? Weekly? Monthly? Less often?*

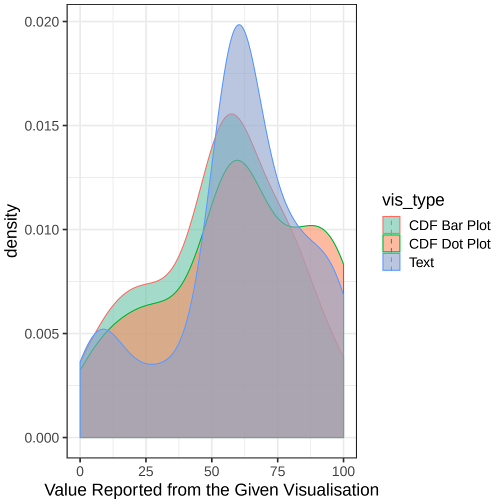


Figure 5: Of 265 users who saw the tutorial, most accurately estimated the likelihood of better results (60%) regardless of visualization type. Users who saw the text description reported more accurately than others.

- *In general, how confident are you in your ability to find the information you need from a Web search engine? (Select a value from a range of 1: Not at all confident, to 5: Very confident)*

2.1 Results

We performed multiple runs of our crowdsourcing experiment, with a total of 800 participants taking part. On average, 0.4 of our participants viewed our brief tutorial on the interface, which is the condition we use to filter out inattentive users who presumably did not read the instructions.

2.1.1 Effect of Visualization Type. One of our primary questions of this survey involved the effect of the type of visualization we presented to users. Is there a difference between users exposed to a particular visualization in their willingness to wait and their estimates of the parameters of the system? We investigate these below.

In our first run of the experiment, we fixed the parameters that we showed to participants to be constant: 60 percent for the likelihood of getting better results in five minutes. We would therefore expect that users, in giving their response to this confirmatory question, would give this estimate or a number close to it. This was indeed the case, as can be seen in Figure 5. In order of the visualisations where users where correct most often, the text description performed the best, then the bar plot, and finally the dot plot.

When asked about their willingness to wait one minute for better results, we found that those with a visualization were slightly more likely to wait than those with no visualization or feedback, but this difference was not statistically significant ($\chi^2 = 2.86$, $p = 0.41$). When asked about waiting five minutes, we found that users without a visualization were much less likely to wait than those with a visualization ($\chi^2 = 18.56$, $p = 0.0003$). This suggests that conveying the expected benefit of the system has value over not conveying this information at all. However, we did not see a statistically significant difference in how we conveyed this information.

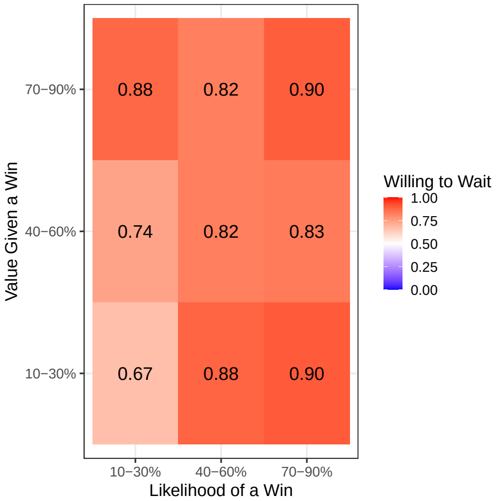


Figure 6: Heatmap of the proportion of users who indicated that they would be willing to wait one minute for better results when presented with a particular likelihood of a win (i.e., better results), and the value given a win. The combination of a low likelihood and value results in the lowest proportion (0.67, lower left) who were willing to wait. The next lowest proportion (0.74) was seen at the lowest likelihood (10-30%), and a medium value given a win (40-60%).

2.1.2 Effect of Varying Parameters. In our previous section, we presented findings based on static parameters. However, we were also interested in the effect of varying the parameters: the probability of a win and the value given a win. For each participant, we randomly chose a value for each of the parameters as a number between 10% and 90%. Because there was no significant effect on visualization type, we stuck to showing a text description of these two parameters (Figure 2).

For ease of analysis, we binned each of these parameters into bins of 10-30%, 40-60%, and 70-90%. We show the proportion of users willing to wait one minute based on the parameters in Figure 6. From this plot, it would seem that showing a lower likelihood of a win and value given a win result in users showing less interest in waiting. Performing a Chi-square test however, we found that there was no significant difference between our parameter values ($\chi^2 = 3.03$, $p = 0.55$) in a willingness to wait one minute. Similarly, there was not a significant difference ($\chi^2 = 4.61$, $p = 0.33$) in a willingness to wait five minutes.

2.1.3 Summary of Findings. We began by looking at the effects of different types of visualisations on the willingness of crowdsourced workers to wait on better results, if the visualisations presented the probability seeing better results in the sidebar as well as the expected value of the better results. The expectation was that showing the uncertainty of these aspects over time provides a more suitable basis for decision making about the future. In this vein, a study by Kay et al. [10] found that users trust measurements more when they are provided with information about uncertainty, and Kay et al. [9] showed that frequency framing – a method of discretising a probability distribution to emphasise particular outcomes rather than the distributions themselves, made them easier to reason about. We found however that there was no difference in visualization type for users making their decisions for their willingness to wait. We nonetheless carried over the dotplot-style of visualisation into the user study, which we will focus on next.

3 ESTIMATES OF QUALITY

We performed the following user study-style experiment, which repurposed the use of a sidebar on a conventional search tool (i.e., a Google-style Web search). The system's sidebar acts as an assistive agent to provide better results for some measure than the main results in some circumstances.

Inspired by Pastor-Bernier et al. [11], we formulated our experiment design in terms of revealed preference theory, which provides the means for determining if the decisions made by agents are such that they seem to be maximizing some underlying utility function. We are interested in looking at users' revealed preferences when exposed to our sidebar that is capable of providing better results; when does he or she prefer one "system" versus the other?

With our optional component, the sidebar, preference is reflected in the choice between searching or waiting. These are activities with potentially different components: searching may correspond to 1. browsing the results of one's current query, or 2. continuing one's task with another query. Similarly, waiting may correspond to either 1. doing nothing (i.e., no interaction with the system), or 2. browsing our set of "slow" results in the sidebar. We hence specify what it means to search or wait to the participants of our study in a brief tutorial preceding their use of the system.

3.1 Method

We performed a pilot for an in-lab experiment to understand user indifference and tradeoffs between risk and value. In this pilot, we provided users with a choice between spending time and effort searching on their own and time spent waiting for better results from an asynchronous system in a sidebar. Our pilot consisted of six graduate students within our department, none of whom specialise in information retrieval.

The system's design was conveyed to users to indicate that the main search interface was akin to a typical Web search system, using the same algorithms and giving a similar expectation in quality. However, the sidebar results were intended to be more volatile, and as such may give much better or much worse results. We controlled the production of results in both cases, and associated the characteristics of the simulated results with the user's selection. Therefore, we could potentially see the effect of variance on preference. We give users a visualization that conveys this variance in a way similar to that shown in Figure 7.

This visualisation reflects the sidebar's expected performance during the course of the user's entire task. Because of our primary interest in decision making, we separate concerns of generating results in the sidebar from showing results in the sidebar, and assume that we have a system with enough spare resources to run the necessary background processing to find the documents to populate the sidebar even if the sidebar is never used. The benefit is that the implications of the visualisation is easily conveyed, and users can make an informed decision of when it is worthwhile to pay the cost of looking at or interacting with the sidebar.

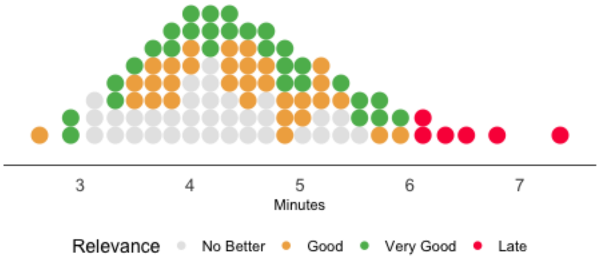


Figure 7: An example of the distribution(s) shown on the time selection screen. One dot is meant to represent a 1% chance of a document with the colour corresponding to its relevance (no better than without the sidebar, good, and very good) being shown in the sidebar. Documents that are expected to be seen after the task completion time of six minutes are coloured red to signify that they are late.

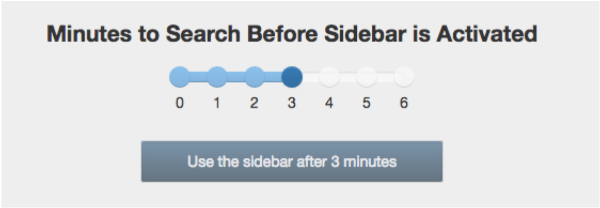
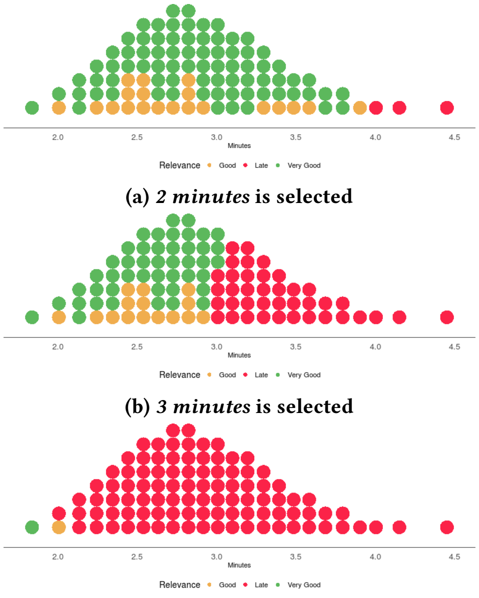

Minutes to Search Before Sidebar is Activated
0
1
2
3
4
5
6
Use the sidebar after 3 minutes

Figure 8: The slider that participants use to signify how much time they would like to spend searching on their own in comparison to using the sidebar.

Before users begin a task, we ask them to choose along a slider the time they would like to spend searching vs. the time they would like to spend waiting. They are able to choose a time from zero (none of the time on the task) to six minutes (all of the time on the task) of time they would like to spend searching.

To assist in decision making, the visualisation is updated as the sidebar is adjusted to reflect the changes in when "late" results arrive in relation to the time they would have left to complete the task. By having this cut-off point in the visualisation, a user will be able to readily see if enough "good" or "very good" documents will be expected to pop up in the sidebar before time runs out to make its usage worthwhile.

They are then presented with a search screen, on which they are reminded of their task and can enter their query. When they are brought to the search engine results page (SERP), they can refine their query, explore results on the left of the screen, or use the sidebar on the right if it is activated. Figure 10 shows the parts of SERP interface labelled along with the sidebar. The results on the left are retrieved using Google's Custom Search API and shows up to 50 results matching the query. The sidebar on the right shows up to five results which were pre-selected by one of the authors to



(a) *2 minutes* is selected

(b) *3 minutes* is selected

(c) *4 minutes* is selected

Figure 9: Variations of the dotplot visualisation are shown when the time preference slider is adjusted.

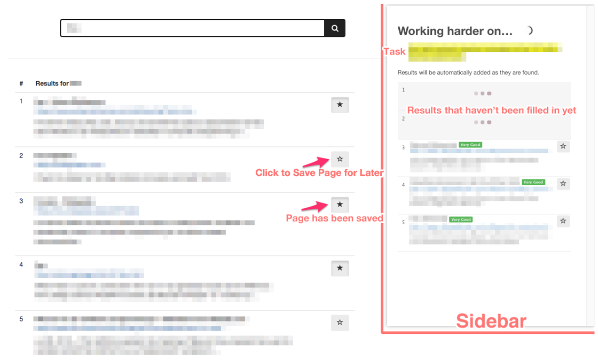

Working harder on...
Task
Results will be automatically added as they are found.
Results for
Results that haven't been filled in yet
Click to Save Page for Later
Page has been saved
Sidebar

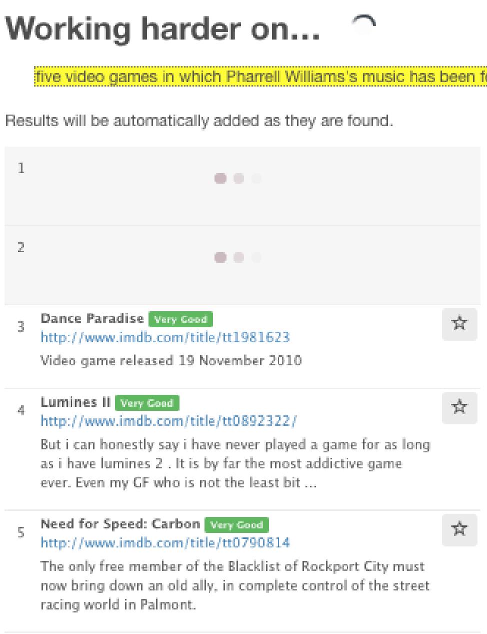
Figure 10: Parts of the search engine results page. Labelled are the sidebar with partially-filled results, task, and buttons to save and unsave results for later.

partially answer the question given in the task. Results that may show up in the sidebar are excluded from the main search results on the left.

When the sidebar is activated, either manually by the user or automatically as time elapses during the task, five results will be populated dynamically over time starting from the lowest position at 5. This is shown in Figure 11.

The user is able to save pages for later using the *star button* on the right of each result, which can be revised when they are entering their answer to the question asked for their task.

For this study, we separate users into one of three conditions randomly for the duration of the study based on the *estimate of future sidebar quality* we present in the visualisation:



Working harder on...

five video games in which Pharrell Williams's music has been featured

Results will be automatically added as they are found.

1

2

3 Dance Paradise Very Good
http://www.imdb.com/title/tt1981623
Video game released 19 November 2010

4 Lumines II Very Good
http://www.imdb.com/title/tt0892322/
But i can honestly say i have never played a game for as long as i have lumines 2 . It is by far the most addictive game ever. Even my GF who is not the least bit ...

5 Need for Speed: Carbon Very Good
http://www.imdb.com/title/tt0790814
The only free member of the Blacklist of Rockport City must now bring down an old ally, in complete control of the street racing world in Palmont.

Figure 11: The sidebar, activated for use. The results arrive and are populated in reverse order starting with the lowest ranked. In this figure, the lowest three ranked have been found and are shown; the top two results have not yet arrived. Animations in the form of a spinning indicator beside the title ("Working harder on...") and throbbers in the unfilled slots of the results provide a degree of operational transparency to show that more results are coming.

- *Accurate*: A condition in which the visualisation is meant to reflect the relative likelihood of the sidebar displaying "good" or "very good" documents
- *Underestimate*: The visualisation's relative likelihood of "good" and "very good" documents are downweighted, and the likelihood of "no better" documents are increased
- *Overestimate*: The visualisation's relative likelihood of "good" and "very good" documents are increased, and the likelihood of "no better" documents are decreased.

The effects of the *underestimate* and *overestimate* conditions are as a form of deception, which is meant to be reminiscent of the type of benevolent deception described in [4]. In this case, a robot therapist gives stroke patients feedback based on the amount of force they exert which is *less* than the actual amount applied. This helps patents overcome their limits. In this context, the *overestimate* condition is meant to encourage participants to use the sidebar more than they might with a more accurate representation. In [1], the authors outline different forms of benevolent deception and their intended benefits – one reason that fits our use is to increase users' comfort with the interface. In a sense this is self-serving because as experimenters it is also to our benefit that users interact with the system. However, for the task and its difficulty, it is indeed more helpful for users to have the assistance of the sidebar, as we expect it will lead to better task performance and thus a higher payout for participants. To overestimate performance, we consider the proportion of "very good", "good", and "no better" documents that will appear in the sidebar and re-weights the proportions so that participants will expect to see many more "very good" documents than the others by applying the following formula:

$$w^{(k)} = \frac{\lambda_k w^{(k)}}{\sum_{k=1}^{K} \lambda_k w^{(k)}}$$

where w is the proportions for each of the document types that will be shown in the sidebar ("very good", "good", and "no better"), K is the number of types of documents (three, in this case, for "very good", "good", and "no better"), and λ_k is the factor we want to re-weigh the proportions by (0.7 for "very good", 0.2 for "good", and 0.1 for "no better").

In comparison, the *underestimate* condition serves as somewhat of a manipulation check – if participants are encouraged to use the sidebar with an overestimate, we would expect participants to be similarly discouraged from using the sidebar when performance is underestimated. Similar to the process taken in the *overestimate* condition, we apply the same to the proportions of "very good", "good", and "no better" documents that are represented in the dotplot visualisation, flipping the λ_k factors so that they become 0.1 for "very good", 0.2 for "good", and 0.7 for "no better".

We compared the users' interaction and gaze behaviours to determine if these conditions had the intended effect on user choice.

We randomly exposed each user to only a single treatment condition, leading to a between-subjects design. Because of our interest in investigating user adjustment to the system (the sidebar and its characteristics), we kept the major aspects of the system constant for a user to be able to observe how their behaviour changes as they continue using the system.

The study procedure was structured as follows:

(1) *Background Questionnaire.* Participants complete a questionnaire that gathers information about their education level, online search experience, and confidence in their search ability.
(2) *Tutorial.* Participants are given a primer on the system's functionality, how to use the interface, and how to read the visualisation. Participants must complete a test task that does not affect their final performance, but that requires them to use functionality such as searching and saving web pages.
(3) *Task completion.* Users complete six tasks in succession, where before each task they are shown the visualisation in Figure 7 and select how much time they would like to spend searching before being able to use the sidebar (Figure 8).
(4) *Post-task Questionnaire.* After each task, the participant gives their answer to the task prompt and fills in survey questions on their experience completing the task with or without the sidebar.

Background and Pre-task Questionnaire. To gather information about user demographics and search expertise, we asked users the following questions before the they are exposed to the system tutorial:

- How would you describe your English language fluency?
  - Non-native English speaker with a near-native understanding of English
  - Non-native English speaker with a very good understanding of English
  - Non-native English speaker with a good understanding of English
  - Non-native English speaker with a limited understanding of English
- What is the highest degree or level of schooling that you have completed?
- How often do you perform searches online? (Select the most specific frequency)
- On average, when you perform a search online, how often have you *not* been able to find what you were looking for? (1: Very rarely; 5: Very often)
- If you needed to perform an online search, how confident are you that you could... (1: Not at all confident; 5: Very confident)
  - Find the information you need?
  - Find pages or articles similar in quality to those obtained by a professional searcher?
  - Create a query that would return every useful page?
  - Create a query that would return only a few *very* useful pages?

Some of these questions were informed by prior work on eliciting and measuring search experience and expertise [2, 3].

After viewing the tutorial and selecting a task, users are asked the following each questions between task selection and performing the task:

- How familiar are you with the subject that you selected?
- How relevant is this subject to your life?
- How much are you personally interested in this subject?
- How often have you searched for something similar before?
- How much effort do you think it would take you to find an answer to this question?
- How quickly do you think it would take you to find enough relevant pages to answer this task's question?

Post-task Questionnaire. After users completed each task, we provided them with a questionnaire to fill out to not only enter their answers for the task and to give their selection of helpful pages for arriving at their answer, but also to evaluate their performance and experience with the task.

As has been noted, we asked participants about their estimate of effort as well as time for the task before they perform the task. Post-task, we then asked participants to re-evaluate their estimate: whether it was about right, too optimistic, or too pessimistic. If their estimate was an underestimate, we then asked how much time they would need to fully complete the task. We think that it is worthwhile to ask about their time estimates because time estimates are a less-biased metric for ascertaining subjective task difficulty than asking users directly [7]. We do still ask users for their estimate of *effort* however, both as a failsafe and to verify the relationship between the two. We note that we are aware of the difference in method between that in [7] and our own – whereas Czerwinski et al. look at *duration* estimates of tasks, we asked users to estimate task completion time directly. In order to capture this

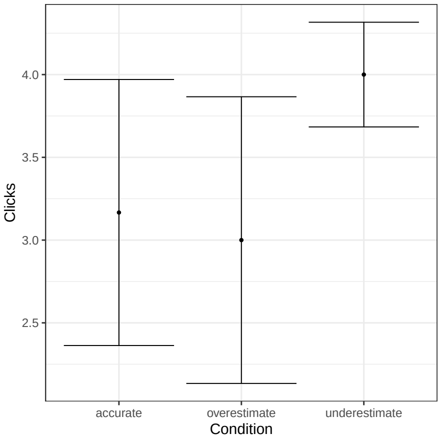


Figure 12: Number of Sidebar Clicks Per Task. Error bars show the standard deviation.

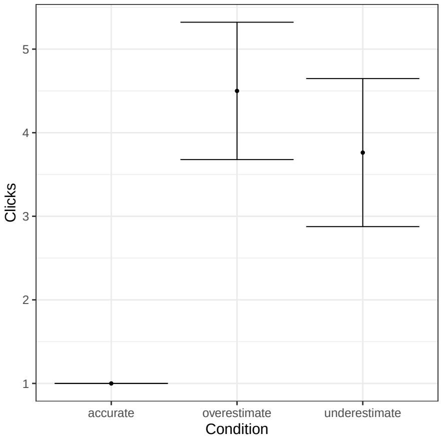


Figure 13: Number of Clicks in the Main Ranking Per Task. Error bars show the standard deviation.

element of subjective time duration, we also asked users who used the full six minutes if they were surprised when they ran out of time.

Data Collection. We used our interaction log data to compute a list of features to characterize search behaviour, including task time and performance, query features such as character and word length, interaction features such as clicks and scrolling as a proxy for reading, eye-tracking data to characterise interface element gaze (i.e., focus on the sidebar, progress bar, etc.), and relevance features such as precision. When performing our experiment, we controlled the system and results in such a way that we could simulate the interaction needed to arrive at the best results, so we could compare these simulated results to actual user behaviour during the progression of the experiment.

3.2 Results

We use this section to analyse users' interactions with the system depending on whether they were given a visualisation that accurately portrayed an estimate of future quality, gave an overestimate of quality, or an underestimate of quality. We first turn our attention to the amount of effort that users expend in completing their tasks, which we may operationalise as the number of clicks and queries submitted.

Click Behaviour. We show the average number of times that a user clicks a result in the sidebar per task in Figure 12. We see that users in the condition for an "overestimate" of future sidebar result quality for our pilot tended to click the sidebar the least, whereas users in the "underestimate" condition tended to click the sidebar the most. This goes counter to what we would have expected – that is, users in the condition that underestimates quality clicking the least and users in the condition that overestimates quality clicking the most. This is not a particularly large difference, but it could point towards a more general trend.

We also looked at the number of clicks in the main result ranking, as shown in Figure 13. We see here that the result is flipped from that in the sidebar between the "underestimate" and "overestimate" conditions; whereas the sidebar gets more clicks for users in the condition that underestimates quality than in the condition that overestimates quality, it gets *fewer* clicks in the main results. This relationship between the rankings follows the expected inverse relationship if we assume that attention spent on one means less attention spent on the other, but we would expect the main ranking to receive more attention when users are shown an underestimate of sidebar effectiveness. This could perhaps mean that the visualisation is having the opposite effect rather than the one intended.

Oddly enough, Figure 13 also shows users in the "accurate" condition making one click per task, which tended to be the first result.

Query Behaviour. We compared the number of queries issued per task by each user, as seen in Figure 14. We see a similar phenomenon here as with the number of main ranking clicks per task (Figure 13) – users in the "accurate" condition have the lowest degree of interaction, while the "overestimate" condition shows the highest degree of interaction. Here, users in the "accurate" condition issued approximately 3.9 queries per task, those in the overestimate condition issued approximately 6.6, and those in the underestimate condition issued approximately 4.7.

Gaze Behaviour. We show heatmaps for the gaze data we collected on the results pages by condition in Figure 15. We limited the gaze data to the area within the sidebar for the plot to more easily compare the level of attention the sidebar received between conditions. The "accurate" condition shows the strongest degree of attention paid to the sidebar in terms of visual fixations, while the "overestimate" condition shows the weakest. Performing a set of pairwise chi-square tests with Holm correction, we find that these proportions are not equally likely as we would expect by chance: "accurate" differs from "overestimate" and "underestimate" at the $p < 0.05$ level, and "overestimate" is also significantly different from "underestimate" at the $p < 0.05$ level.

If we compare the number of sidebar clicks (Figure 12) to the visual attention the sidebar received, we see that the "accurate" and "underestimate" conditions do show more clicks in the sidebar than the "overestimate" condition, which lines up proportionally to the amount of visual attention for "accurate" and "underestimate" compared to "overestimate". This may hint at less attention being

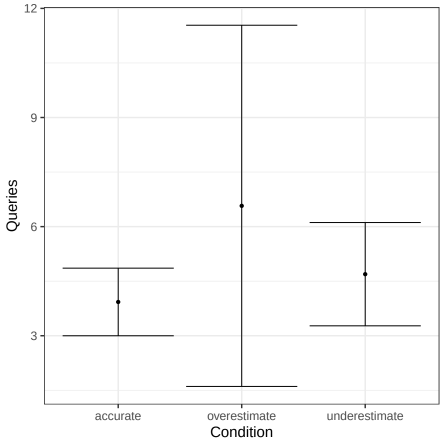


Figure 14: Number of Queries Issued Per Task. Error bars show the standard deviation. Users in the condition that accurately estimates quality issued the fewest queries (approximately 3.9 per task) whereas users in the condition that overestimates quality issued the most (approximately 6.6 per task) with the largest standard deviation.

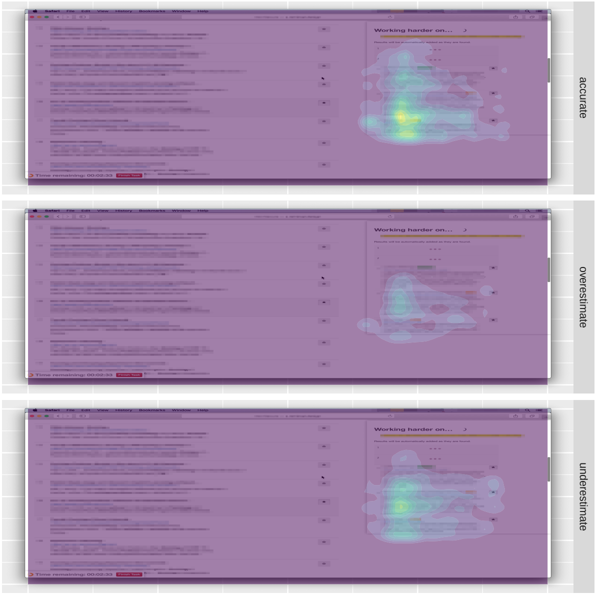


Figure 15: Eye gaze heatmaps limited to the area within the sidebar. The sidebar receives more fixations in the "accurate" and "underestimate" conditions than the "overestimate" condition, with the "accurate" condition receiving the most.

paid generally to the sidebar in the "overestimate quality" condition compared to the others. The number of clicks remains sparse however, and the exact nature of the differences may be better left to a larger-scale full experiment.

Post-Task Questionnaires. We now turn towards the results of our post-task questionnaires, which were completed by participants after every task, in order to glean users' perceptions of the system in addition to their behaviours.

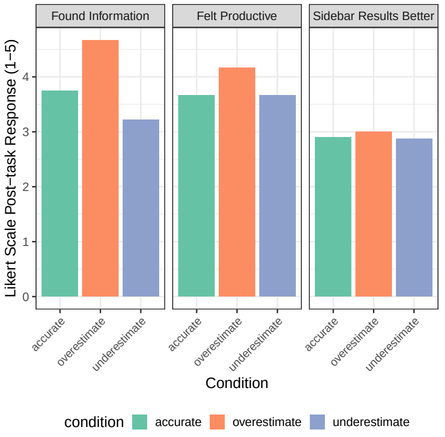


Figure 16: Average Responses to a subset of post-task questions on a Likert scale from 1 to 5. Users in the "overestimate" condition tended to agree more than other conditions that they found the information they were looking for, that they felt productive, and that the sidebar gave better results than they could find on their own.

Figure 16 shows the responses to the following questions: "I was able to satisfactorily find the information I needed" ("Found Information" in the figure), "I felt that my time was spent productively, and that I made substantial progress while doing this task" ("Felt Productive"), and "I think the sidebar in fact gave me better results than I could find on my own" ("Sidebar Better"). We see that users in the "overestimate" condition seemed to agree with these statements more than other conditions, though this effect is relatively small.

In Figure 17, we plot the responses to two questions we asked in the post-task questionnaire about their expectations of the sidebar: "How likely do you think the sidebar was to give you a good result based on what you saw?" ("Estimated Probability of Success" in the figure), and "When you saw a result in the sidebar that you thought was good, how good do you think that result was on average?" ("Estimated Value"). We offered options ranging from "very likely" to "very unlikely" for the estimated probability of success, and "very good" to "very bad" for the estimated value. These were converted to a scale of 1–5 for the figure. Although the questions were phrased to elicit impressions of the sidebar based on their experience using it, there does seem to be an effect related to the experimental condition: users in the "overestimate" condition indeed thought that the sidebar would have given better results than those in the other conditions, and users in the "underestimate" condition had the lowest expectations.

4 DISCUSSION & CONCLUSIONS

We began our work by looking at the effects of different types of visualisations on the willingness of crowdsourced workers to wait on better results, if the visualisations presented the probability seeing better results in the sidebar as well as the expected value of the better results. Although there were no significant differences

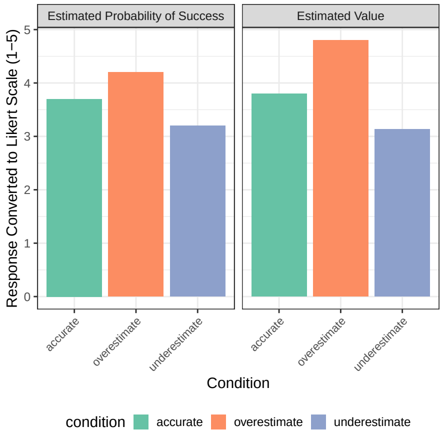


Figure 17: Average Responses to estimates of expected sidebar effectiveness during post-task questionnaire. Responses were converted to a scale from 1–5. Users in the "overestimate" condition had higher expectations for the sidebar than the other conditions. Users in the "underestimate" condition had the lowest.

between visualisation types, we maintained the direction of frequency framing our pilot study with a quantile dotplot that showed how likely the sidebar was to surface results of varying degrees of quality at a given point in the future. One aspect that we did not investigate was the degree of granularity of the plot. Although users could count the number of dots, feedback from the pilot suggested that there may have been too many to count. We aim to revise this in future work. In our pilot user study, we incorporated operational transparency through the combination of a visualisation to convey the expected future performance of an optional sidebar that gives higher-quality results relevant to a search task, as well as feedback in the sidebar that additional results will be shown over time. The preliminary results from the pilot indicate that giving accurate information about the expected performance of the sidebar may have a positive effect on the attention paid to the sidebar – as seen through gaze tracking data – and fewer queries issued. However, subjectively, pilot participants exposed to an overestimate in the future performance gave post-task questionnaire responses that indicated higher productivity and a higher estimated value of the sidebar. These warrant further study with more participants, but point to a subjective effect of deception applied to operational transparency.

Concluding Remarks. We presented an analysis of preference when users are given the option of using a sidebar with higher quality results that comes from waiting an additional amount of time, when we also present an indicator of the quality of results that can be expected over time. We began with a crowdsourced study looking at different types of visualisation that offered either point estimates or a time series of outcomes, and found that although there was not an appreciable difference between the types of visualisations, there was value in showing an indication of the quality that users could expect. We then showed the results for an interactive pilot study of a system with a visualisation that provided a potential distribution of outcomes over time that is either accurate, an overestimate of quality or an underestimate of quality. These results of indicate potential effects of the perception of quality on query behaviour, click behaviour, and attention to the sidebar when measured via eye tracking, as well as subjective opinions through post-task surveys. This work shows the value of user perceptions of quality in use, and also points towards larger full-scale studies on perceptions of the expected quality one might get from the future use of an optional component of a search system.

REFERENCES

[1] Eytan Adar, Desney S Tan, and Jaime Teevan. 2013. Benevolent deception in human computer interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1863–1872.

[2] Earl Bailey. 2017. *Measuring Online Search Expertise*. Ph. D. Dissertation. The University of North Carolina at Chapel Hill.

[3] Kathy Brennan, Diane Kelly, and Yinglong Zhang. 2016. Factor analysis of a search self-efficacy scale. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, 241–244.

[4] Bambi R Brewer, Matthew Fagan, Roberta L Klatzky, and Yoky Matsuoka. 2005. Perceptual limits for a robotic rehabilitation environment using visual feedback distortion. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 13, 1 (2005), 1–11.

[5] Ryan W Buell and Michael I Norton. 2011. The labor illusion: How operational transparency increases perceived value. *Management Science* 57, 9 (2011), 1564–1579.

[6] Ryan Burton and Kevyn Collins-Thompson. 2016. User Behavior in Asynchronous Slow Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. ACM, New York, NY, USA, 345–354. https://doi.org/10.1145/2911451.2911541

[7] Mary Czerwinski, Eric Horvitz, and Edward Cutrell. 2001. Subjective duration assessment: An implicit probe for software usability. In *Proceedings of IHM-HCI 2001 conference*, Vol. 2. 167–170.

[8] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (Philadelphia, Pennsylvania, USA) *(CSCW '00)*. ACM, New York, NY, USA, 241–250. https://doi.org/10.1145/358916.358995

[9] Matthew Kay, Tara Kola, Jessica Hullman, and Sean Munson. 2016. When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16)*.

[10] Matthew Kay, Dan Morris, mc schraefel, and Julie Kientz. 2013. There's no such thing as gaining a pound: reconsidering the bathroom scale user interface. In *Ubicomp '13*.

[11] Alexandre Pastor-Bernier, Charles R Plott, and Wolfram Schultz. 2017. Monkeys choose as if maximizing utility compatible with basic principles of revealed preference theory. *Proceedings of the National Academy of Sciences* 114, 10 (2017), E1766–E1775.

[12] Claudia Plant and Christian Böhm. 2011. INCONCO: Interpretable Clustering of Numerical and Categorical Objects. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Diego, California, USA) *(KDD '11)*. ACM, New York, NY, USA, 1127–1135. https://doi.org/10.1145/2020408.2020584

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[14] Kirsten Swearingen and Rashmi Sinha. 2002. Interaction design for recommender systems. In *Designing Interactive Systems*, Vol. 6. 312–334.

[15] Jaime Teevan, Kevyn Collins-Thompson, Ryen W White, Susan T Dumais, and Yubin Kim. 2013. Slow Search: Information Retrieval without Time Constraints. In *Proceedings of HCIR 2013*. ACM.

[16] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*. IEEE, 801–810.

[17] Dimitrios Tsekouras, Ting Li, and Izak Benbasat. 2018. Scratch My Back and I'll Scratch Yours: The Impact of the Interaction Between User Effort and Recommendation Agent Effort on Perceived Recommendation Agent Quality. *Available at SSRN 3258053* (2018).

[18] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2017. Making sense of recommendations. *Journal of Behavioral Decision Making* (2017).