Test of the Explore Similar Widget: Providing users with related pages, categories and suggested languages in search results

Erik Bernhardson *Senior Software Engineer, Wikimedia Foundation*
Jan Drewniak *User Experience Engineer, Wikimedia Foundation*
Chelsy Xie *Data Analyst, Wikimedia Foundation*
Deb Tankersley *Product Manager (Analysis, Search Frontend), Wikimedia Foundation*

Wikimedia Engineering's Discovery's Search team ran an A/B test from 30 Jun 2017 to 24 Jul 2017 to assess the effectiveness of adding 'explore similar' widget to provide randomly selected users on English Wikipedia with related pages, categories and suggested languages in addition to each search result. We found that about 10% of users hover over the widget but almost no one clicked on the similar result links. Users hover over the 'Related' section most often, followed by 'Categories' and 'Languages' sections. More than 80% of hovers lasted less than 2 seconds, which means that the user was either not interested in the related content or hovered over the links by accident. Our analysis also showed that this widget did not improve user engagement, but it did not hurt user experience neither. Given the fact that very few users saw the similar results and we got positive feedback from usertesting.com, we suggest making the feature to be more visually distinct and then run another test.

Introduction

In the ongoing effort to improve the discoverability of all the world's knowledge, the Discovery team is proposing an enhancement to the search result snippets on the search results page on Wikipedia (Special:Search). The goal of this feature is to surface related content for each individual search result, so that even when the search result itself isn't entirely useful, perhaps its related content could be.

For the users who received the experimental user experience (UX), three new links would be added beside the green metadata on each search result: related pages, categories and languages. When hovered over, these link would expand the search result into a 'card' with related content (see Figure 1). These links would reveal the following related content:

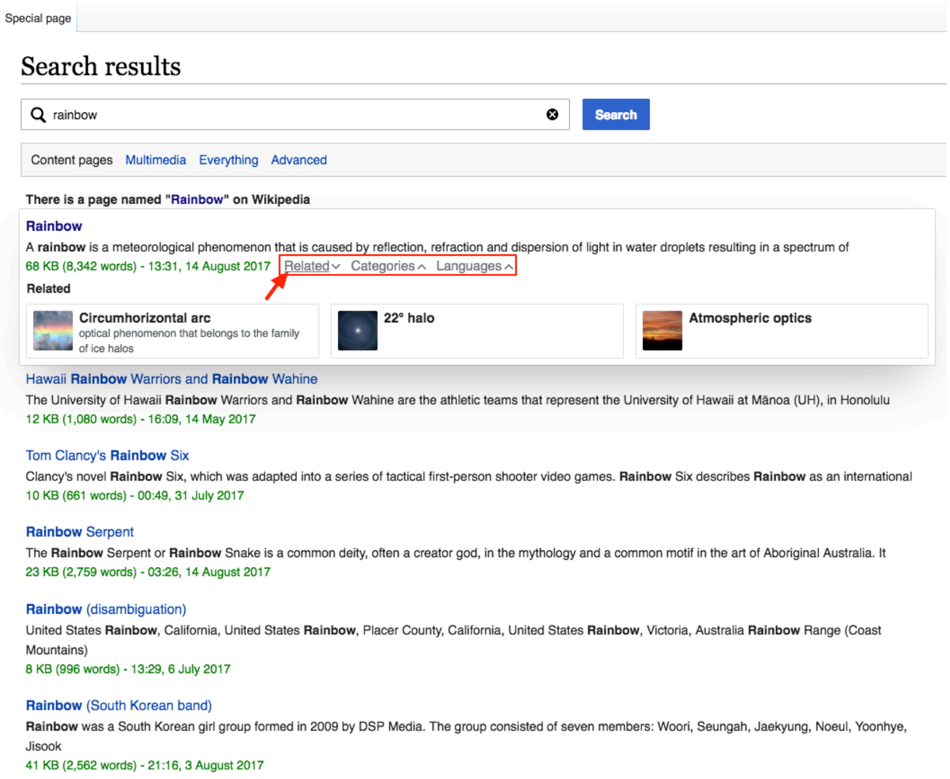
- Up to 3 related page links containing metadata (as is available): thumbnail image, name of related page and metadata description
- Up to 9 links to categories that the article was tagged with
- Any other languages the page is available in

This extended content would be activated by hovering over one of the three links. Also, in order to avoid overcrowding the UI, these links would only be visible when hovering over the search result. There was one test group of users who received the experimental UX, and one control group that did not and received the currently existing search results page.

The primary questions we wanted to answer are:

- Did users who saw the additional related content engage with those results?

Source code and data are available on GitHub (wikimedia-research/Discovery-Search-Test-ExploreSimilar)



Special page

Search results

rainbow

Search

Content pages Multimedia Everything Advanced

There is a page named "Rainbow" on Wikipedia

Rainbow

A **rainbow** is a meteorological phenomenon that is caused by reflection, refraction and dispersion of light in water droplets resulting in a spectrum of

68 KB (8,342 words) - 13:31, 14 August 2017 Related Categories Languages

Related

Circumhorizontal arc
optical phenomenon that belongs to the family of ice halos

22° halo

Atmospheric optics

Hawaii **Rainbow** Warriors and **Rainbow** Wahine

The University of Hawaii **Rainbow** Warriors and **Rainbow** Wahine are the athletic teams that represent the University of Hawaii at Mānoa (UH), in Honolulu

12 KB (1,080 words) - 16:09, 14 May 2017

Tom Clancy's **Rainbow** Six

Clancy's novel **Rainbow** Six, which was adapted into a series of tactical first-person shooter video games. **Rainbow** Six describes **Rainbow** as an international

10 KB (661 words) - 00:49, 31 July 2017

Rainbow Serpent

The **Rainbow** Serpent or **Rainbow** Snake is a common deity, often a creator god, in the mythology and a common motif in the art of Aboriginal Australia. It

23 KB (2,759 words) - 03:26, 14 August 2017

Rainbow (disambiguation)

United States **Rainbow**, California, United States **Rainbow**, Placer County, California, United States **Rainbow**, Victoria, Australia **Rainbow** Range (Coast Mountains)

8 KB (996 words) - 13:29, 6 July 2017

Rainbow (South Korean band)

Rainbow was a South Korean girl group formed in 2009 by DSP Media. The group consisted of seven members: Woori, Seungah, Jaekyung, Noeul, Yoonhye, Jisook

41 KB (2,562 words) - 21:16, 3 August 2017

***Figure 1**: Example of explore similar widget on English Wikipedia when a user move the mouse on the 'Related' section of the first search result.*

- Was the overall engagement with search results better or worse compared to the controls?

On 30 June 2017 we deployed an A/B test on the desktop version of English Wikipedia to assess the efficacy of this feature. The test concluded on 24 July 2017, after a total of 5180 search sessions had been anonymously tracked.

Methods

This test's event logging (EL) was implemented in JavaScript according to the TestSearchSatisfaction2 (TSS2) schema, which is the one used by the Search team for its metrics on desktop, data was stored in a MySQL database, and analyzed and reported using R [R Core Team, 2016].

Data

1 in 1000 users are included in Event-logging (on English Wikipedia that's about 2000 full-text searchers, according to T163273). Of those 1 in 1000 users, 1 in 2 are included in the test (that's about 1000 sessions daily). Of those 1 in 2 users, 50% go in the test group (as described above), and the other 50% of users will go in the control group (that's about 500 sessions for each bucket). However, we failed to notice that the sampling didn't work as we expected – we got only about 100 sessions per day for each bucket. But we got enough data since we ran this test longer than planned – for 25 days.

After the test has concluded on 24 July 2017, we processed the collected data and filtered out duplicated events, extraneous search engine result pages (SERPs), events without an associated SERPs, and remove

Test group	Search sessions	Searches recorded
Control	2,570	4,618
Test	2,566	4,719
Total	5,136	9,337

***Table 1**: Number of search sessions used for analysis by group. Each search session may have several individual searches.*

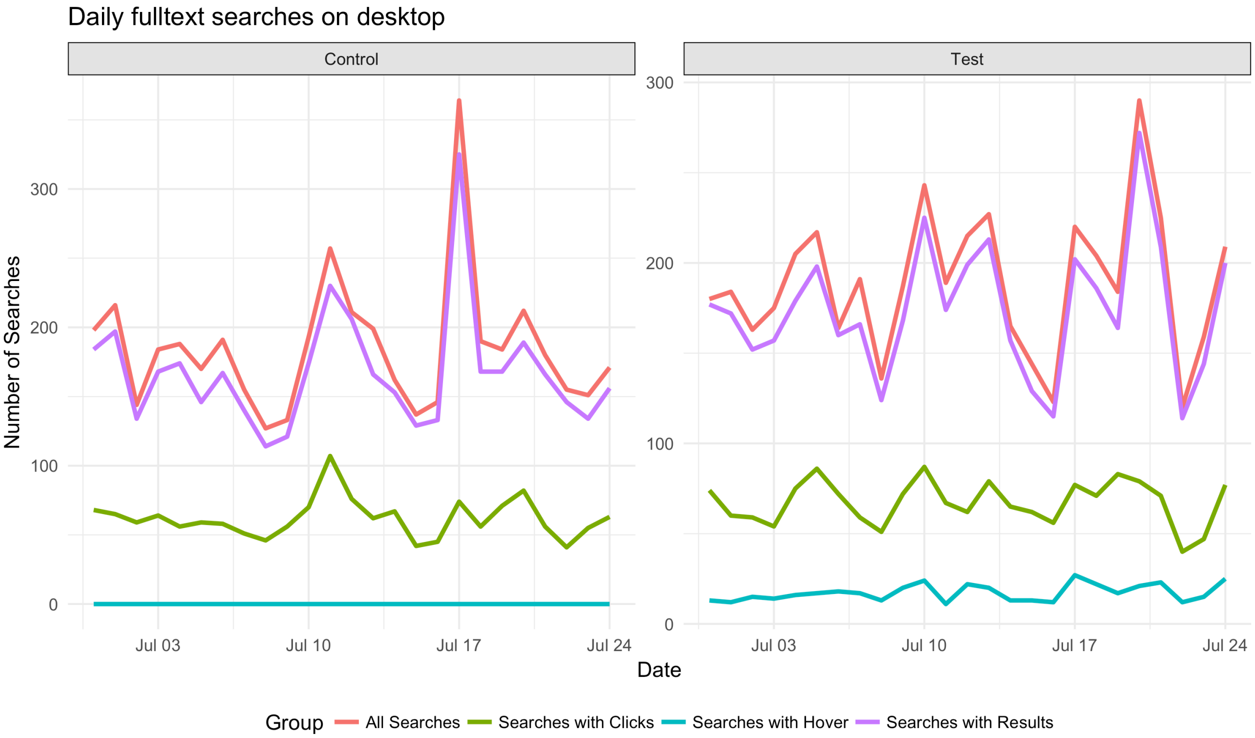


***Figure 2**: Number of all searches, searches with results, searches with clickthrough and searches with hover-over, by day and group.*

sessions with more than 50 searches but no interaction with the result. This left us with a total of 5136 search sessions (see Table 1 for the full breakdown by group).

Results

There are around 185 searches everyday in each group, around 170 of them return at least one same-wiki result, and about 65 searches have clickthroughs. In the test group, about 20 searches have hover-over actions (see Figure 2 for daily counts and Figure 3 for daily zero result rates and clickthrough rates). The hover-over rate is 9.8% (see Figure 4).

Engagement with the results

Figure 5 compares same-wiki clickthrough rates between control and test group, and also shows the clickthrough rates among searches with at least one hover-over. The three clickthrough rates are not significantly different, which means showing related information didn't lead to a higher clickthrough rate.

After we show the search result page to a user, if the user doesn't do anything – no click, scroll, hover-over, not reload the search result page or check the next page of results, we will say this search is abandoned.

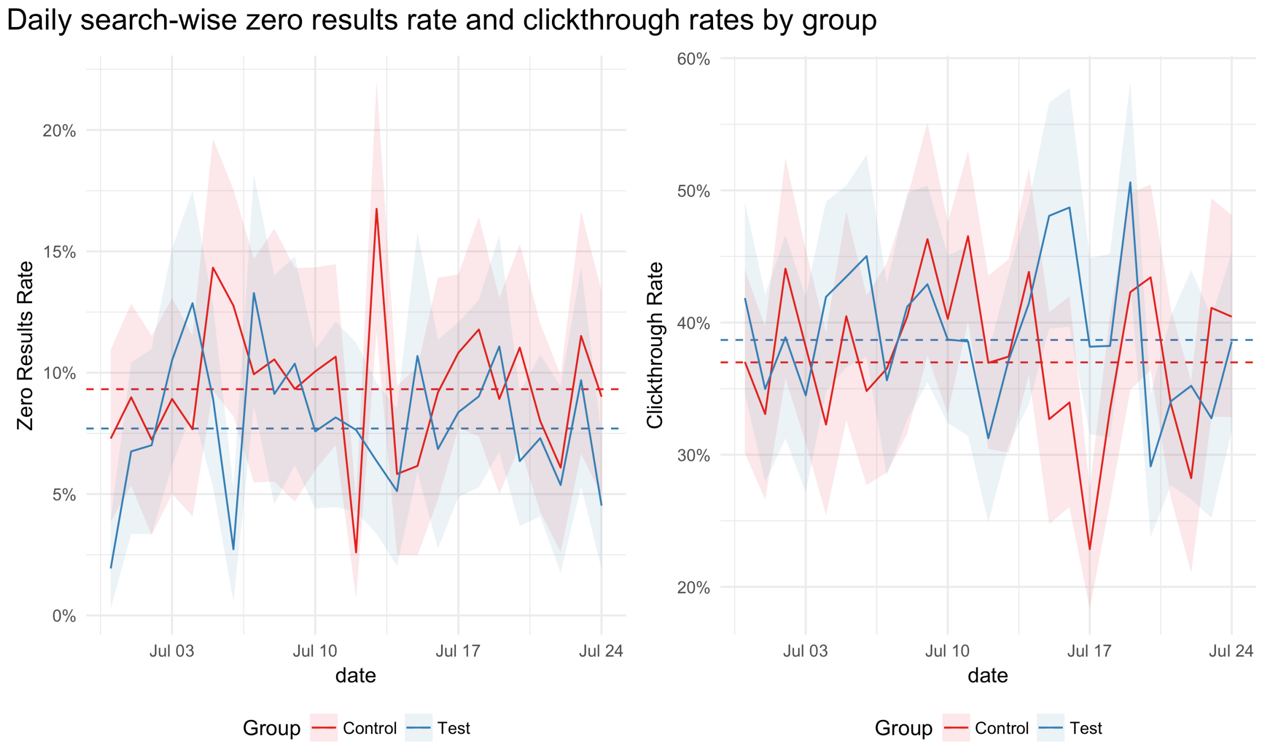


Figure 3*: Daily search-wise zero results rate and clickthrough rates of experimental groups. Dashed lines mark the overall zero results rate and clickthrough rate.*

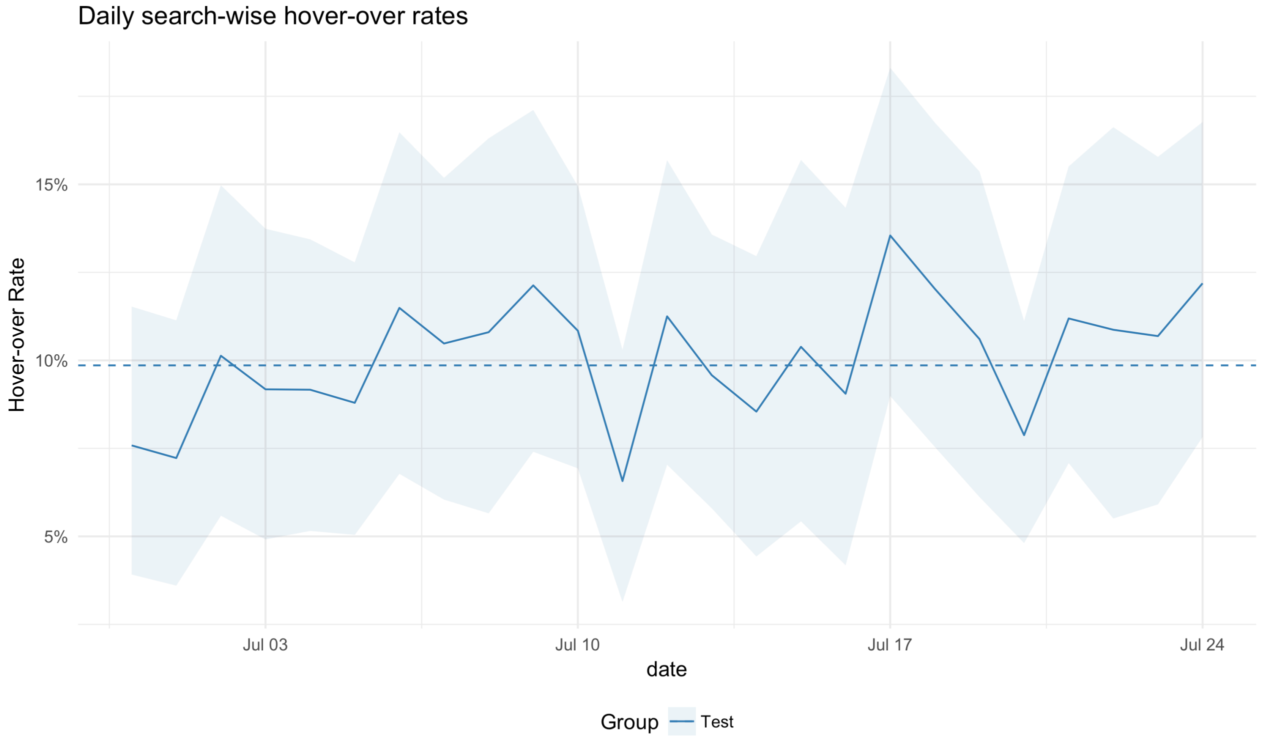


Figure 4*: Daily search-wise hover-over rate of the test group. Dashed lines mark the overall hover-over rate.*

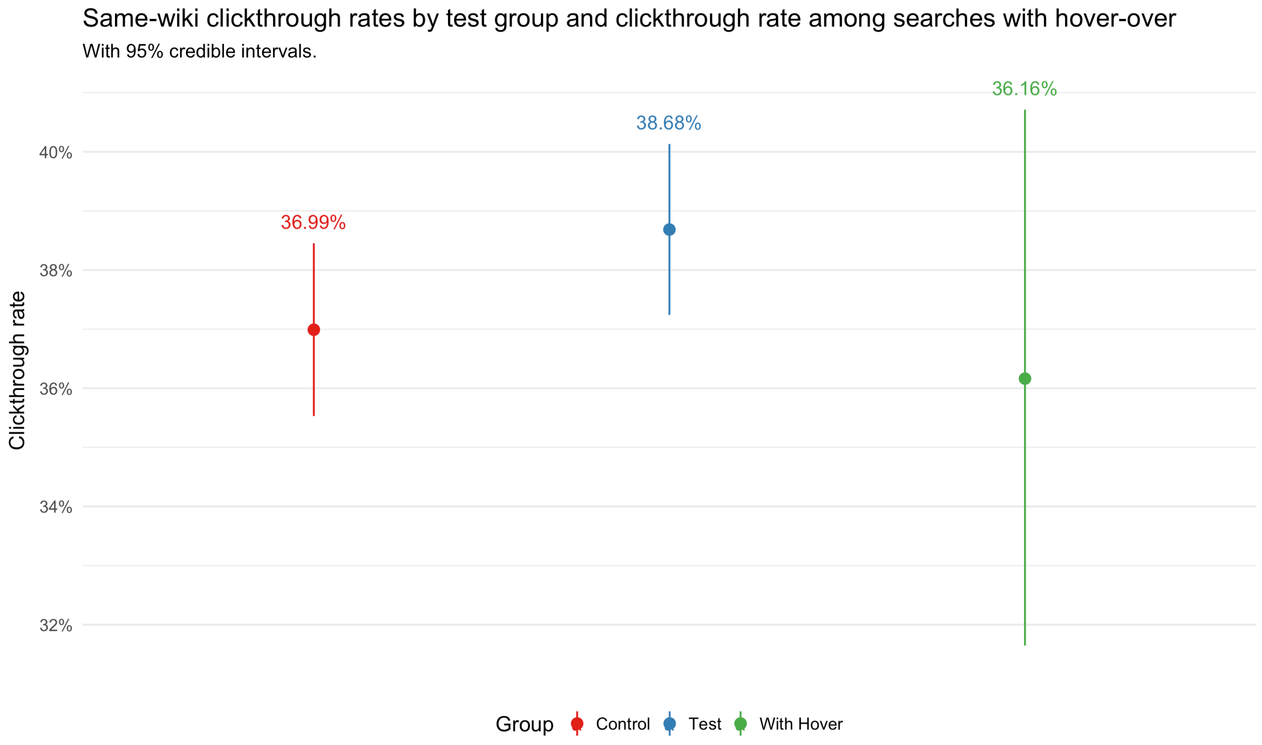


***Figure 5**: Clickthrough rates of experimental groups and clickthrough rate among searches with hover-over.*

In Figure 6, the left graph shows the proportion of searches without any actions on the result page, and the abandon rate of test group is significantly lower. However, the left graph is not comparing apples to apples, because:

- The control group didn't track hover-over and explore similar clicks and we cannot prove that the hover-over in test group didn't happen by accident. In other words, the control group users might move their mouse on the pages, but they didn't get recorded;
- According to our EL schema, the more actions are recorded, the more likely a scroll can be recorded, thus a scroll in the test group is more likely to be tracked (Although Figure 7 shows that the proportion of search result pages with scroll in the test group is not significantly higher.).

Therefore we exclude hover-over, explore similar clicks and scrolls, and generate the right graph which shows the proportion of searches that have no click, reload or page-turn. There is no significant difference between the two groups, which means the test group did not have more search result page reload, page-turn and clicks of any kind (same-wiki, cross-wiki or other).

Figure 8 shows that the time users spent on search result pages are not significantly different.

Interactions with explore similar widget

There are 721 hover-on actions and Table 2 shows the breakdown by section and number of results shown. 12.8% (92) hovers sees 0 result, most of them are in the "Languages" section. 44.1% (318) hover on the "Related" section, 34% (245) hover on the "Categories" section and 21.9% (158) hover on the "Language" section.

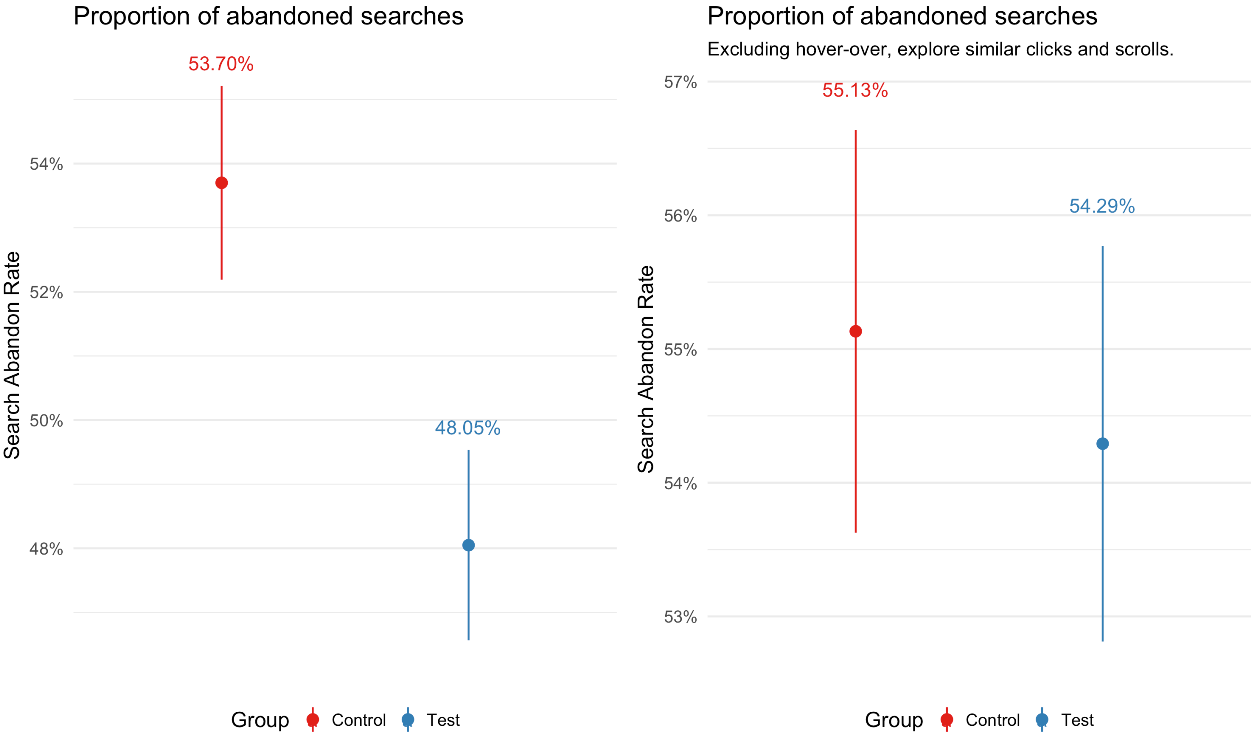


Figure 6*: Search abandon rate by test group. The graph on the left shows the proportion of searches without any actions on the result page. The right one shows the proportion of searches that have no click, reload or page-turn.*

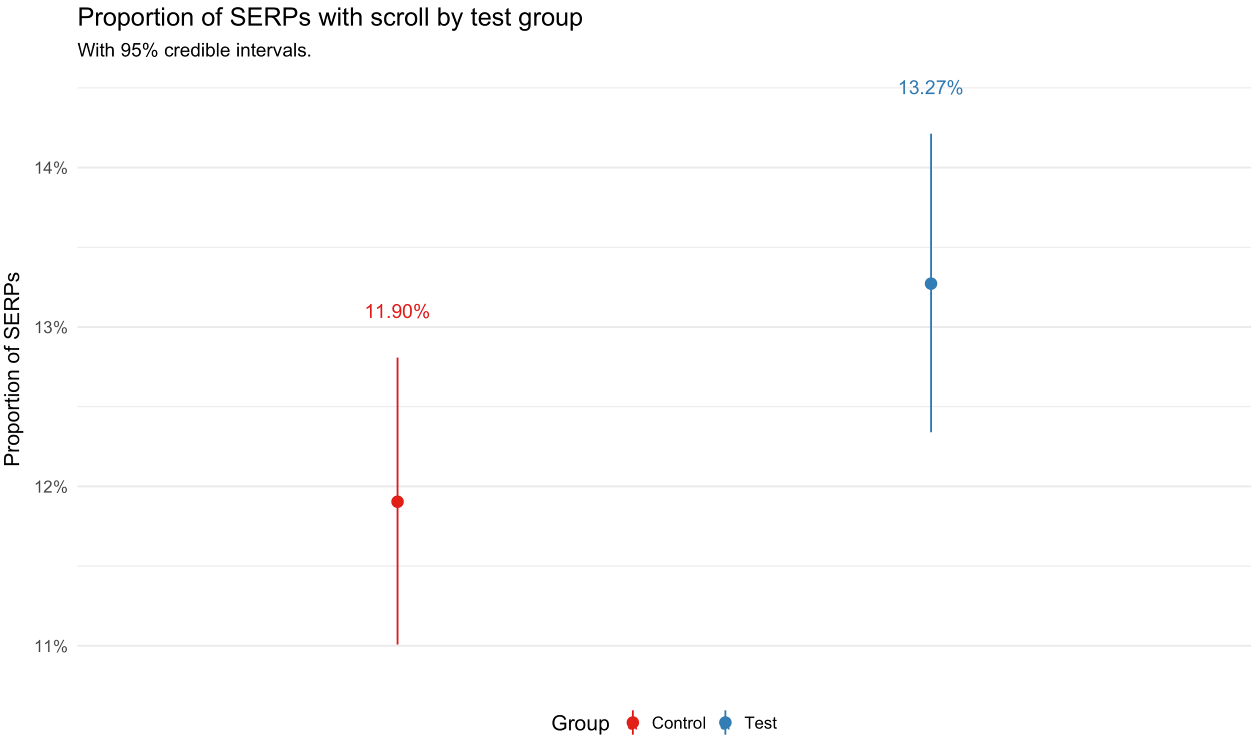


Figure 7*: Proportion of search results pages with scroll by test group.*

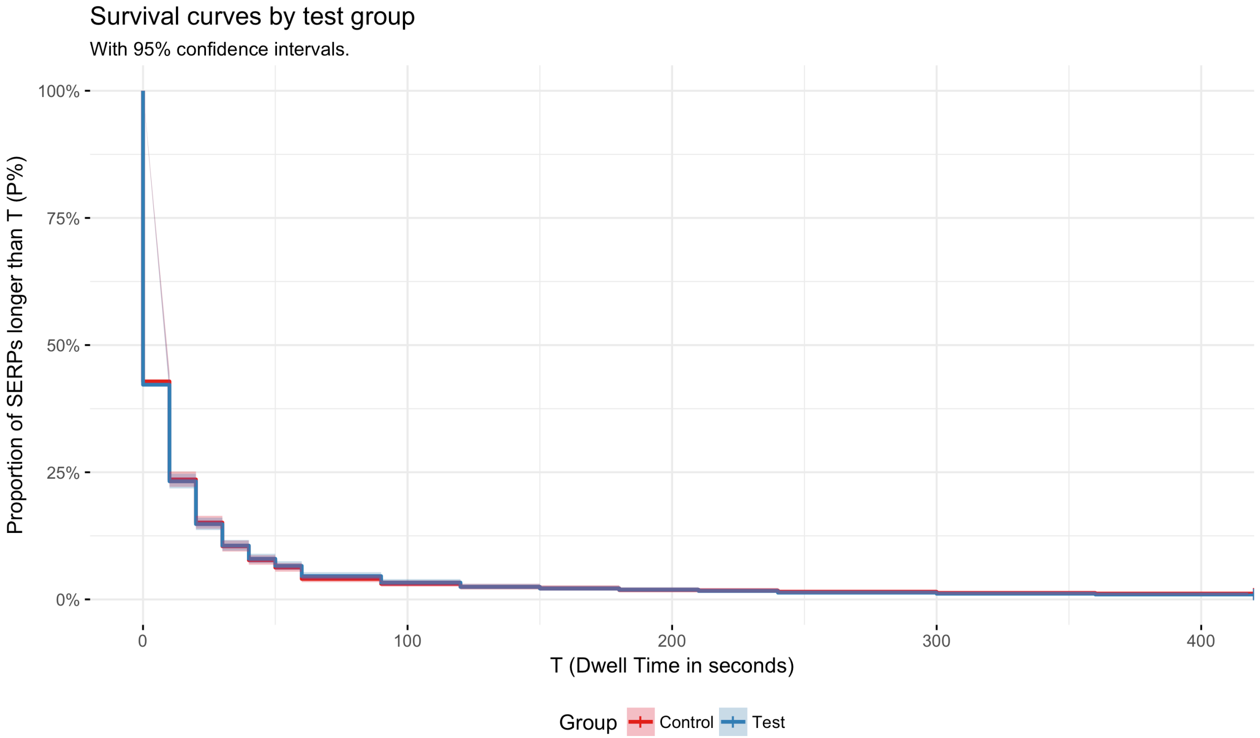


***Figure 8**: Proportion of search results pages from autocomplete last longer than T. Currently we are only able to track the dwell time of search result pages from autocomplete search.*

	0 results	1 result	2 results	3 results	4 results	5+ results	Sum
categories	2	23	28	44	37	111	245
languages	83	24	14	12	8	17	158
related	7	0	0	311	0	0	318
Sum	92	47	42	367	45	128	721

***Table 2**: Number of hover-on actions by section and number of results shown. For example, there are 2 hovers on the 'Categories' section and see 0 result.*

Dwell Time	Number of Hover-over	Proportion
0 secs	196	50.39%
1 secs	128	32.9%
2 secs	34	8.74%
3 secs	11	2.83%
4 secs	4	1.03%
5 secs	4	1.03%
6+ secs	12	3.08%

***Table 3**: Number of hover-over that stay on the widget for n seconds. For example, there are 196 hover-over actions that stay on the widget for less than 1 seconds. The dwell time is the difference between the timestamp of hover-off and hover-on.*

Table 3 shows the approximate time users stay on the widget after they hover-on. More than 80% of hovers lasted less than 2 seconds, which means that the user is either not interested in the related content or hovered over the links by accident.

There are only 3 valid clicks on the explore similar results from 3 different sessions. There are 1 click on the first "Related" result and 2 clicks on the second "Related" result.

Unfortunately, we didn't record the position of the same-wiki search result when users interact with the explore similar widget.

Return Rate

Users may click back to the search result page directly after they clickthrough to an article (within 10 mins). We computed two kinds of return rate:

- Among users with at least a click in their search, the proportion of searches that return to the same search page
- Among users with at least a click in their search session, the proportion of sessions that return to search for different things (different search result page but in the same session)

From Figure 9, we can see the return rates are not significantly different between experimental groups.

Load time of search results pages

Figure 10 shows that the distributions of search result pages' load time of the two experimental groups almost overlap.

Discussion

We found that about 10% of users hover over the widget but almost no one clicked on the similar result links. Users hover over the 'Related' section most often, followed by 'Categories' and 'Languages' sections. More than 80% of hovers lasted less than 2 seconds, which means that the user is either not interested in the related content or hovered over the links by accident. Our analysis also showed that this widget did not improve user engagement, but it did not hurt user experience neither. Given the fact that very few users saw the similar

We computed the dwell time for 389 hover-over events. There are 330 hover-on events without a paired hover-off event because the users moved their mouses over the links so fast that javascript did not fire the hover-off event for them. And there are 2 pairs with duplicated hover-on events but only 1 hover-off event. Therefore, we cannot computed the dwell time for these 332 hover-over event.

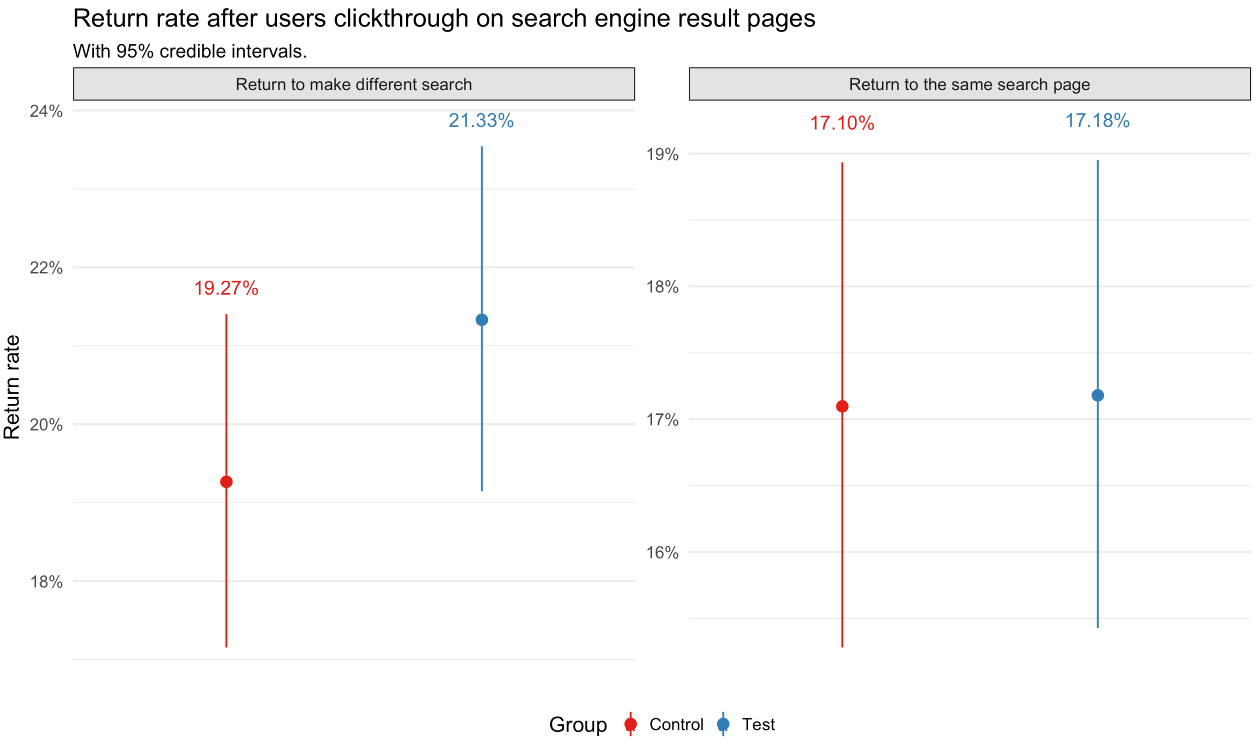


***Figure 9**: Return rate after users clickthrough on search engine result pages.*

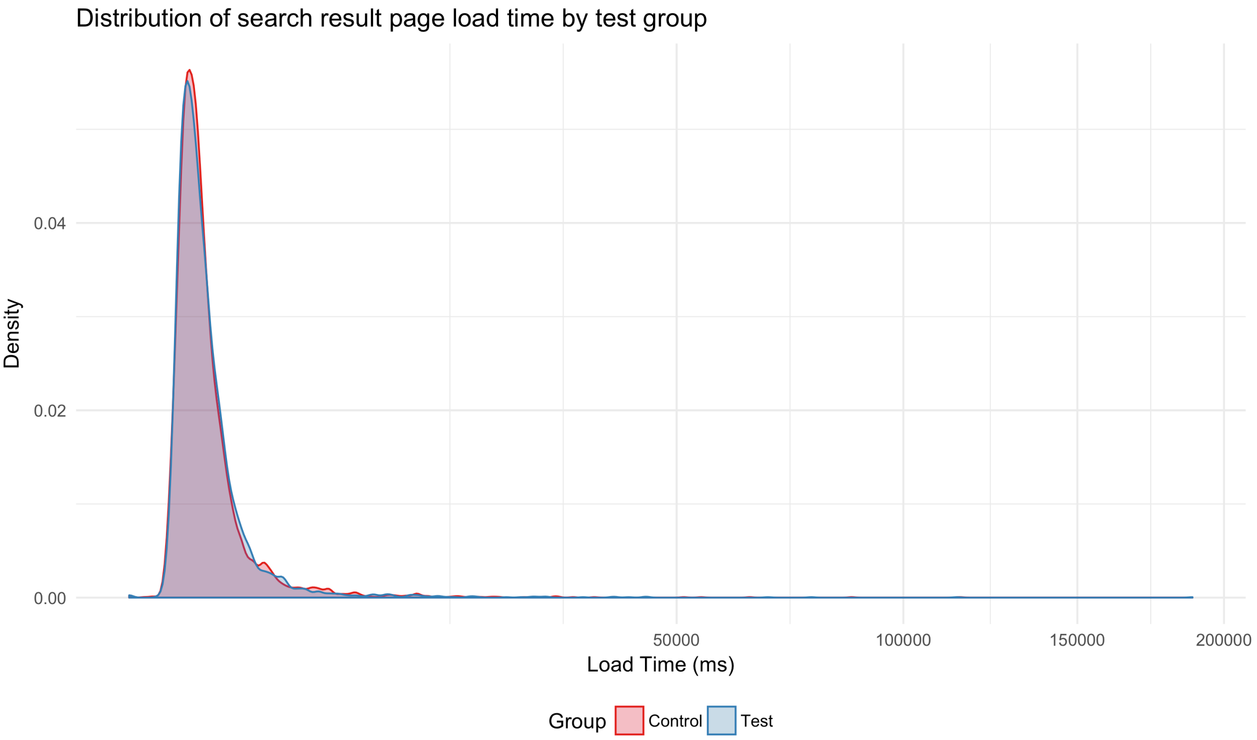


***Figure 10**: Distribution of search result pages' load time by test group.*

results and we got positive feedback from usertesting.com, we suggest making the feature to be more visually distinct but also not too visually obtrusive. Additionally, it would be helpful to record the position of the same-wiki result on the page when users hover-over or click on the related content under it.

Acknowledgements

Finally, we would like to thank our colleague Mikhail Popov (Data Analyst, Wikimedia Foundation) for his review and feedback on this report.

References

JJ Allaire, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, Rob Hyndman, and Ruben Arslan. *rmarkdown: Dynamic Documents for R*, 2016. URL http://rmarkdown.rstudio.com. R package version 1.3.9002.

Stefan Milton Bache and Hadley Wickham. *magrittr: A Forward-Pipe Operator for R*, 2014. URL https://CRAN.R-project.org/package=magrittr. R package version 1.5.

Sundar Dorai-Raj. *binom: Binomial Confidence Intervals For Several Parameterizations*, 2014. URL https://CRAN.R-project.org/package=binom. R package version 1.1-1.

Oliver Keyes and Mikhail Popov. *wmf: R Code for Wikimedia Foundation Internal Usage*, 2017. URL https://phabricator.wikimedia.org/diffusion/1821/. R package version 0.2.6.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

Hadley Wickham. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*, 2017. URL https://CRAN.R-project.org/package=tidyr. R package version 0.6.1.

Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*, 2016. URL https://CRAN.R-project.org/package=dplyr. R package version 0.5.0.