A Test Of Cross-wiki Search: Helping Users Discover Content On Wikipedia's Sister Projects

Erik Bernhardson *Senior Software Engineer, Wikimedia Foundation*
Jan Drewniak *User Experience Engineer, Wikimedia Foundation*
Dan Garry *Product Manager (Search Backend), Wikimedia Foundation*
Mikhail Popov *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 9 Feb 2017 to 22 Feb 2017 to assess the effectiveness of performing cross-wiki searches and showing Catalan, Italian, Persian, and Polish Wikipedias' users results from sister projects such as Wikisource and Wikiquote. We found that while the overall engagement with the search results was higher for the two test groups compared to the control group, there was no sufficient evidence to definitively say that the additional search results increased user engagement. We suspect that a critical UX design issue – links shown in black, rather than standard blue – and the particular languages this test was deployed on (resulting in a higher zero results rate than seen across other languages) had a negative effect on the results, and recommend performing a follow-up test.

Introduction

Within the Wikimedia Foundation's Engineering group, the Discovery department's mission is to make the wealth of knowledge and content in the Wikimedia projects (such as Wikipedia) easily discover-able. The Search team is responsible for maintaining and enhancing the search features and APIs for MediaWiki, such as language detection (i.e. if a French Wikipedia visitor searches in German, then in addition to results from French Wikipedia, they would also get results from German Wikipedia).

Specifically, the Search team's current goal is to add cross-wiki searching – that is, providing search results from other (also referred to as "sister") Wikimedia projects ("*wikis*") within the same language. For example, if a work (e.g. a book or poem) on French Wikisource contained that German phrase, that user would be shown results from French Wikisource in addition to any results from French and German Wikipedias.

For the users who received the experimental user experience (UX), each additional wiki's top result was shown as a box in a sidebar with a link to view more results (see Figure 10). There were two groups of users who received the experimental UX and one control group that did not:

- **Control**: This group received the baseline user experience, which only includes the search results from the wiki they are on. To make their experience comparable to the test groups with respect to latency, we performed the search across the additional indices, but did not show the results to the end user.
- **Test (Random)**: This group received the experimental user experience, which includes search results from other wikis (if any were returned). The boxes holding the results (one box for each wiki) were ordered randomly.
- **Test (Recall)**: This group received the experimental user experience, which includes search results from other wikis (if any were returned). The boxes holding the results (one box for each wiki) were ordered according to recall – the volume of search results returned for each respective wiki.

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

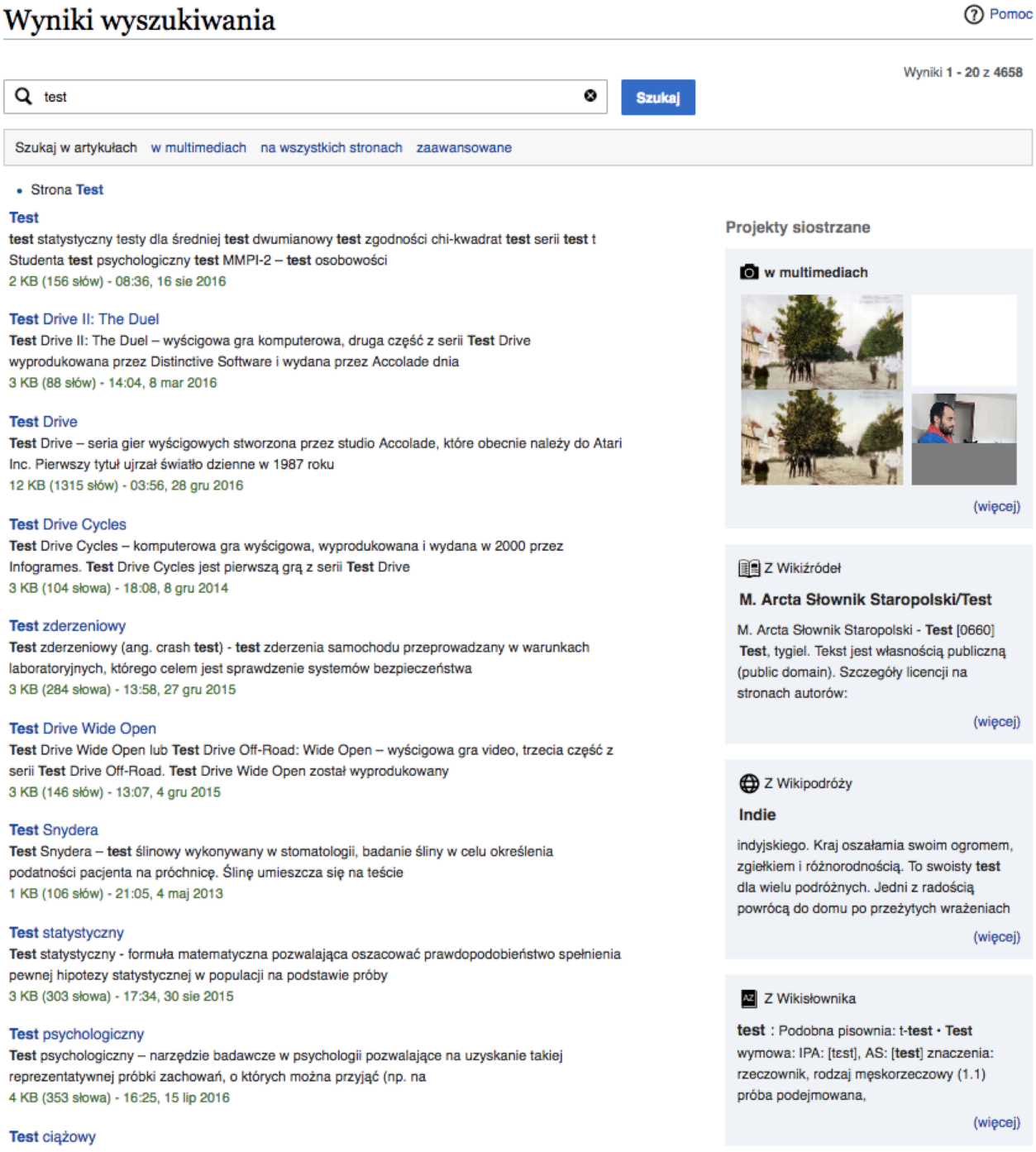


***Figure 1**: Example of cross-wiki search results on Polish Wikipedia, with sister wikis randomly ordered in the sidebar. Multimedia results (including results from Wikimedia Commons) are shown first, regardless of the sidebar ordering.*

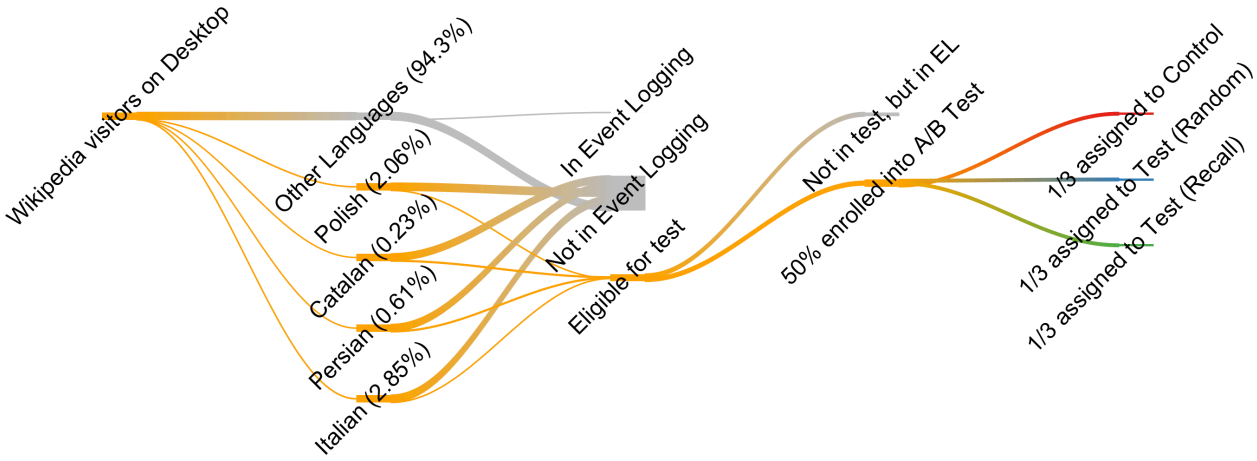


***Figure 2**: Flow of Wikipedia visitors into the A/B test.*

The primary questions we wanted to answer are:

- Did users who saw the additional cross-wiki results engage with those results?
- Was the overall engagement with search results better or worse compared to the controls?

On 9 February 2017 we deployed an A/B test on the desktop version of Catalan, Italian, Persian, and Polish Wikipedias to assess the efficacy of this feature. The test concluded on 22 February 2017, after a total 6620 search sessions have 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

The data was collected according to the TSS2 schema, revision 16270835. Figure 11 shows the flow of Wikipeda visitors on Desktop. Approximately 5.7% of the unique desktop devices that visit the 270 Wikipedias are accounted by Catalan, Italian, Persian, and Polish Wikipedias. In general, desktop users are randomly selected for anonymous tracking at a rate of 1 in 200, but for these wikis we changed the sampling rate to 1 in 200 for Catalan and Persian, and 1 in 50 for Italian and Polish. After a user was randomly selected into event logging, they had a 50% chance to be selected for the A/B test. Users who made it into the test were then randomly assigned to one of the three groups described above: Control, Test (Random), and Test (Recall).

We would like to note that our event logging does not support cross-wiki tracking, so after the user leaves the search results page, we cannot learn if they have performed subsequent searches, how and how long the user engaged with the visited result's page.

See Phabricator ticket T149806 and Gerrit changes 334314, 313318, 332991, 334685, and 336896 for full details of the implementation on both back-end and front-end.

Analysis

We employed the *binom* [Dorai-Raj, 2014] and internally-developed *BCDA* [Popov] packages for Bayesian statistical analysis and confidence intervals in Figures 12, 13, 14, and 15.

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

	Control	Test (Random)	Test (Recall)	All 3 groups
Catalan Wikipedia	460	413	414	1287
Italian Wikipedia	664	664	710	2038
Persian Wikipedia	664	631	658	1953
Polish Wikipedia	454	456	432	1342
All 4 wikis	2242	2164	2214	6620

***Table 2**: Number of click events by group.*

	Same-wiki clicks	Sister-project clicks	Overall clicks
Control	1321	0	1321
Test (Random)	1260	38	1298
Test (Recall)	1274	42	1316
All 3 groups	3855	80	3935

Results

After the test has concluded on 22 February 2017, we processed the collected data and filtered out duplicated events, extraneous search engine result pages (SERPs), and kept only the searches for which we had both event logging (EL) data and logs of searches (Cirrus requests). This left us with 6620 search sessions (see Table 1) with the full breakdown by wiki and group. Table 2 breaks down the counts of clicks on same-wiki results (e.g. a Catalan Wikipedia visitor clicking on a Catalan Wikipedia article) and clicks on sister-projects results (e.g. an Italian Wikipedia visitor clicking on an Italian Wikinews article).

In Table 5, we observed the following three interactions:

- User A was a Persian Wikipedia visitor who got enrolled into the test and received their cross-wiki results ordered according to recall. The Wikiquote box was displayed first (below the Multimedia box). The user clicked on the specific result, then clicked to see more results from Wikiquote.
- User B was an Italian Wikipedia visitor who got enrolled into the test and received their cross-wiki results ordered according to recall. This user had a similar interaction, although Wikisource was the first box displayed (below the Multimedia box), and the switch to viewing more results (after the initial probe into the specific result) was quicker.
- User C was an Italian Wikipedia visitor who got enrolled into the test and received their cross-wiki results randomly ordered. The Wikiquote box was randomly displayed first (below the Multimedia box) and the user clicked to see more results. Perhaps unsatisfied with Wiktionary's search results, the user clicked Back, but Wikisource was displayed first (below the Multimedia box) on this second load of the original search results.

Zero Results Rate (ZRR)

The zero results rate (ZRR) – proportion of searches yielding zero results – is one of Discovery's Search Team's key performance indicators (KPIs), and we are always interested in lowering that number (but not at the expense of results' relevance). While we were primarily interested in searchers' engagement with the

***Table 3**: A sample of some visitors' interaction with the cross-wiki search results after they were enrolled into the test groups.*

User	Wikipedia	Group	Timestamp	Click	Position	Sister Project	Destination
A	Persian	Recall	2017-02-15 18:16:20	1st	2nd	Wikiquote	Article
A	Persian	Recall	2017-02-15 18:17:03	2nd	2nd	Wikiquote	More Results
B	Italian	Recall	2017-02-17 09:04:18	1st	2nd	Wikisource	Article
B	Italian	Recall	2017-02-17 09:04:32	2nd	2nd	Wikisource	More Results
C	Italian	Random	2017-02-19 17:43:38	1st	2nd	Wikiquote	More Results
C	Italian	Random	2017-02-19 17:45:50	2nd	2nd	Wikisource	More Results

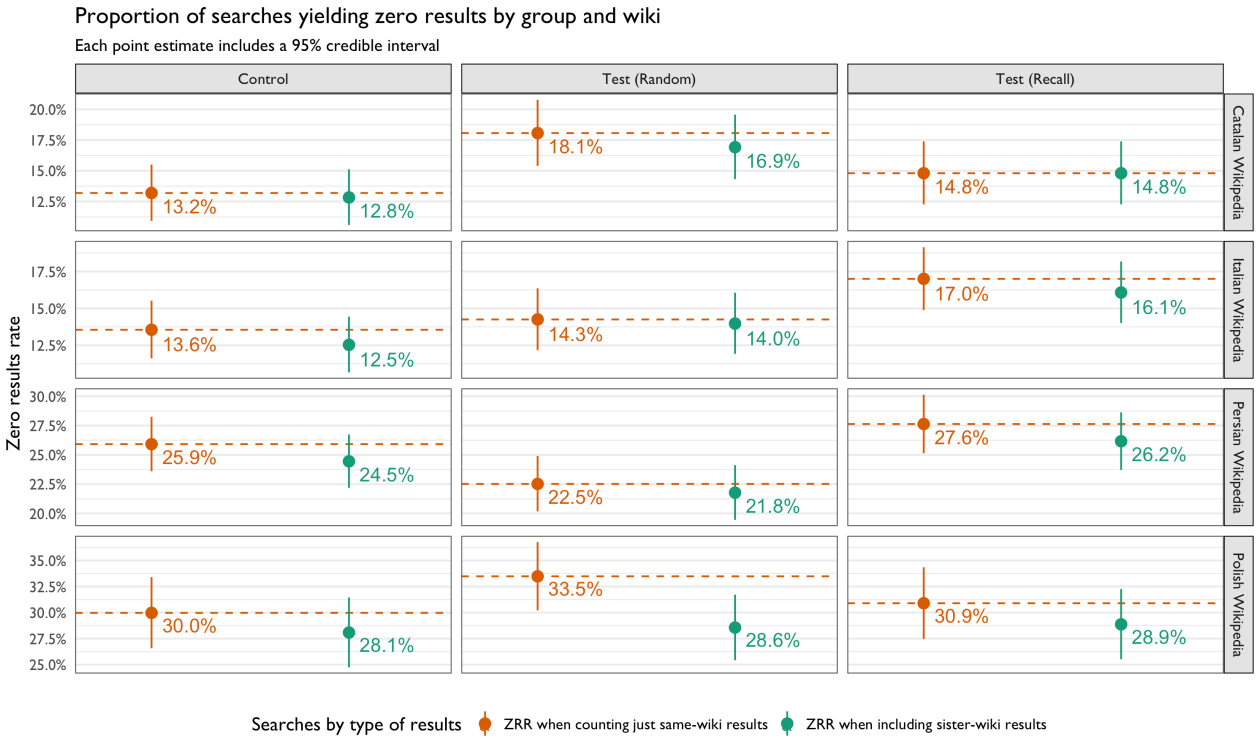


***Figure 3**: Proportion of searches yielding zero results broken up by group, wiki, and type of results (same-wiki only vs. including cross-wiki results).*

search result for this test, we included this section as a consistency check – that the zero results rate is lower when a cross-wiki search is performed (see Figure 12).

In Figure 4, we broke the ZRR from Figure 3 down by language and project and included a reference marker for each project's overall ZRR (aggregated across all the languages the project is available in). Almost all of the projects are available (at the time of the test and at the time of writing this report) in Catalan, Italian, Persian, and Polish.

Namely, the overall ZRR of projects like Wikinews and Wiktionary (both exist in Catalan, Italian, Persian, and Polish) appears to be much lower than the ZRR observed in this test.

In fact, the ZRR in these four languages is much higher than the overall ZRR for every project. We suspect this is responsible for the low sister-project click counts seen in Table 2.

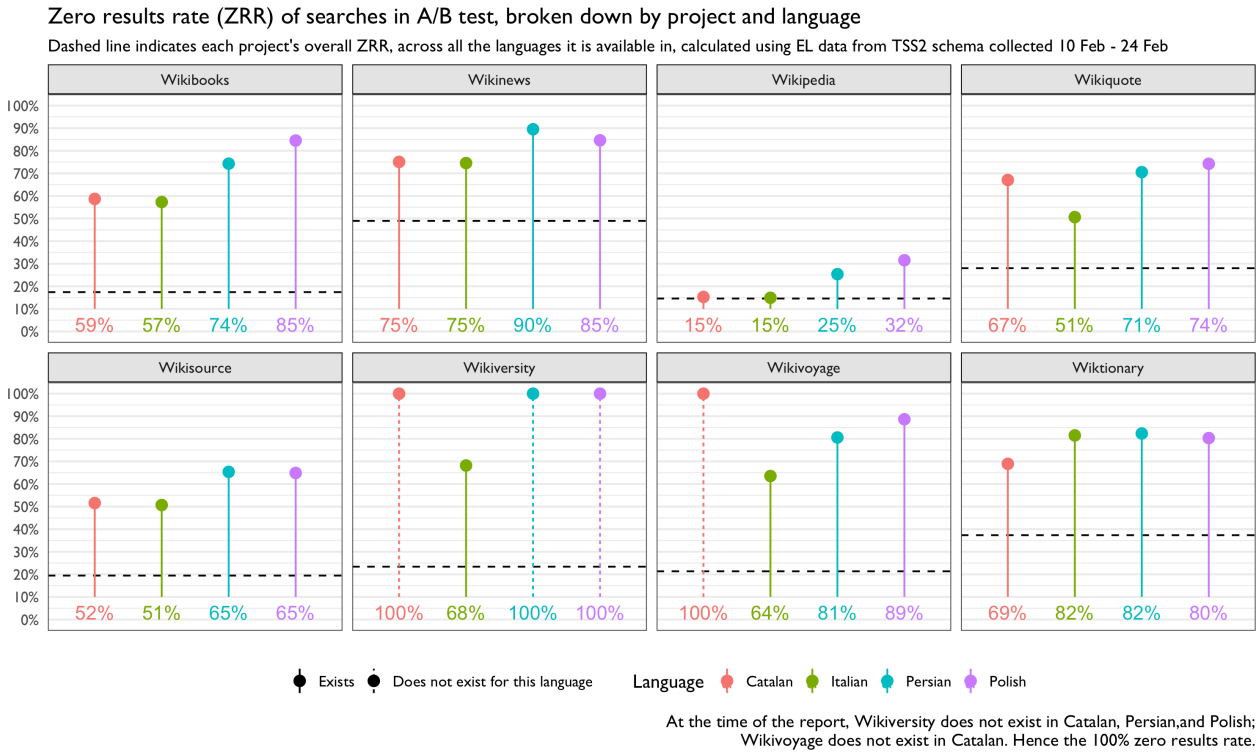


***Figure 4**: The proportion of searches that yielded zero results was the lowest for Wikipedia and Wikisource, with the other projects having very high zero result rates.*

Engagement

We used the clickthrough rate as an indicator of users' engagement with search results and as a measure of the results' relevance. That is, if we present users with more relevant results (such as those from Wikipedia's sister projects), then we expect the clickthrough rate to be higher in the two test groups compared to that of controls. Figure 5 shows that various search activity measures did not vary too much from one group to another.

In Figures 6, 7, and 8, we saw that the clickthrough rate was higher in Test (Random) and Test (Recall) than in Control on almost all of the four wikis. The only exception being the clickthrough rate of users in the Test (Random) and Test (Recall) group on the Italian and Polish Wikipedias, respectively.

Table 6 shows the relative risk – how more likely each respective test group is to engage with the search

***Table 6**:*

Wiki	Comparison	Relative Risk	95% CI
Catalan Wikipedia	Test (Random) vs Control	1.103	(0.920, 1.302)
Catalan Wikipedia	Test (Recall) vs Control	1.096	(0.903, 1.289)
Italian Wikipedia	Test (Random) vs Control	1.047	(0.941, 1.160)
Italian Wikipedia	Test (Recall) vs Control	0.976	(0.881, 1.085)
Persian Wikipedia	Test (Random) vs Control	1.055	(0.904, 1.199)
Persian Wikipedia	Test (Recall) vs Control	1.131	(0.977, 1.283)
Polish Wikipedia	Test (Random) vs Control	0.999	(0.850, 1.161)
Polish Wikipedia	Test (Recall) vs Control	1.122	(0.955, 1.304)

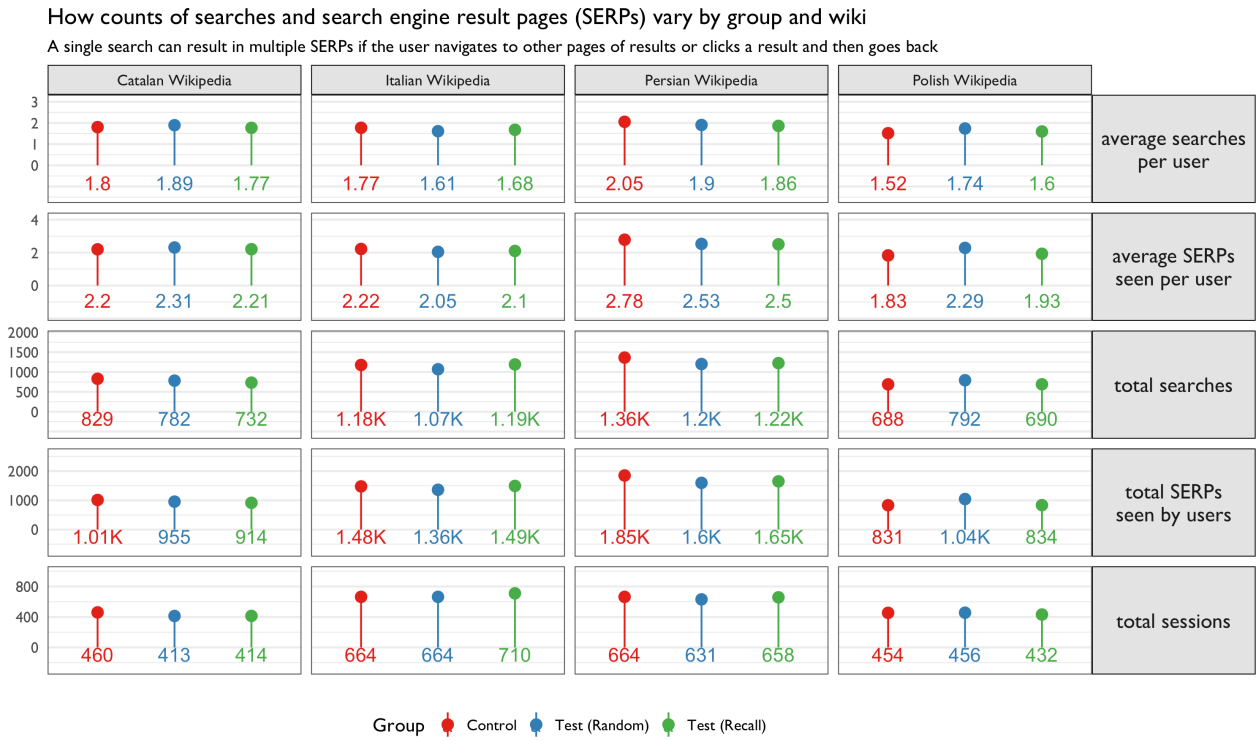
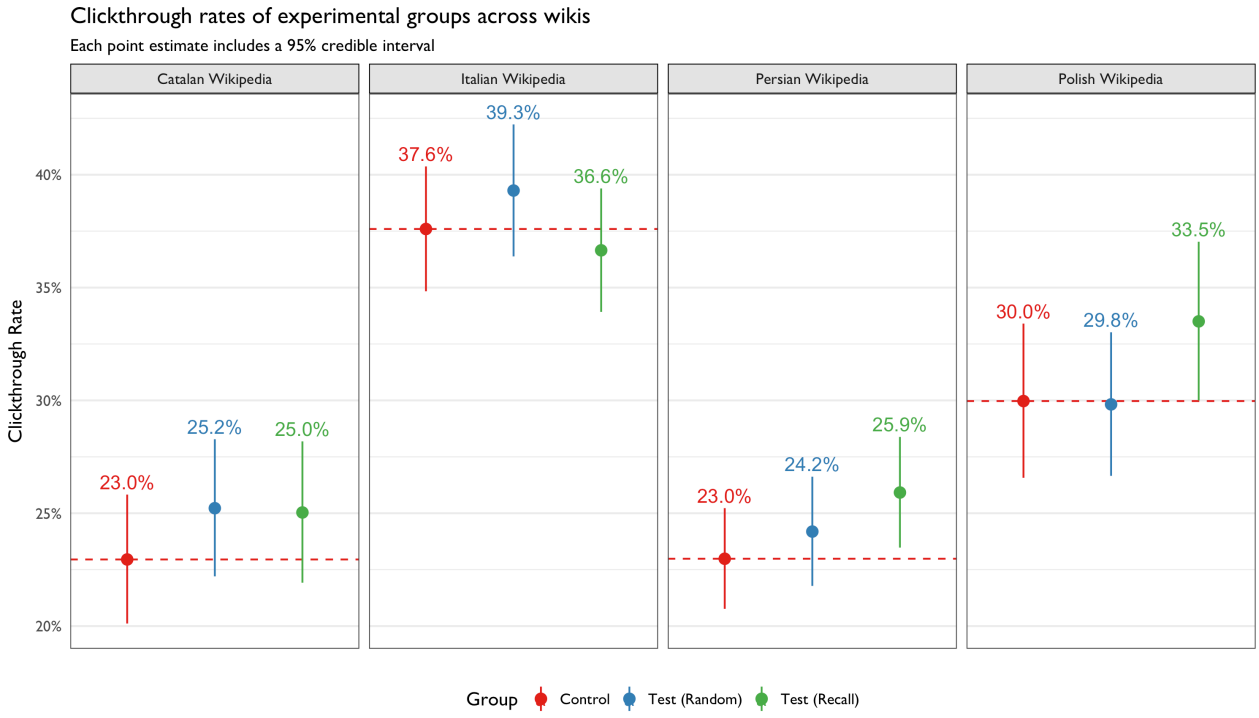


***Figure 5**: Average number of searches, average number of search engine result pages (SERPs), total searches, total SERPs, and total sessions by group and wiki. The groups did not appear to behave too differently. For example, the three groups had very similar average searches per user.*



* For this engagement analysis we focused on: (1) controls' searches with same-wiki results, and (2) test groups' searches that included sister-wiki results.

***Figure 6**: Clickthrough rates of experimental groups, split by wiki. In the Control group, only searches that yielded some same-wiki results were considered. In the two test groups, only searches that yielded some sister-wiki results were considered.*

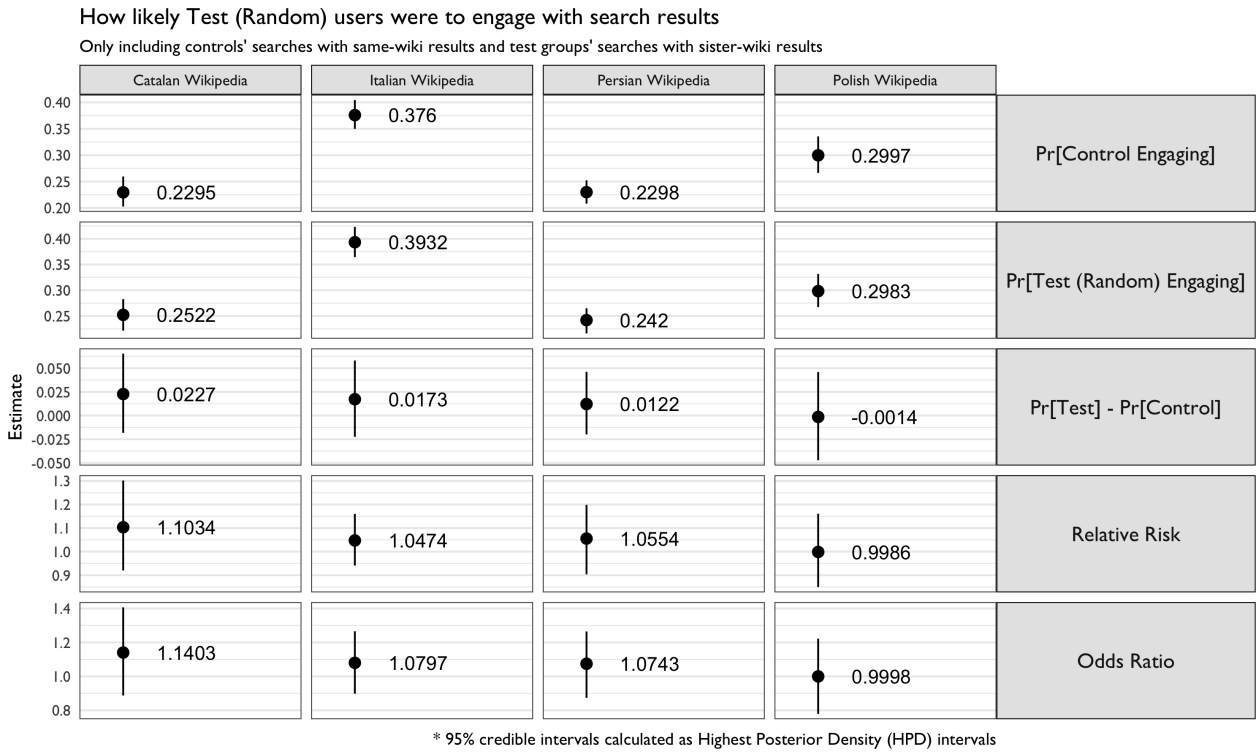


***Figure 7**: Comparison of the Control group with the Test (Random) group.*

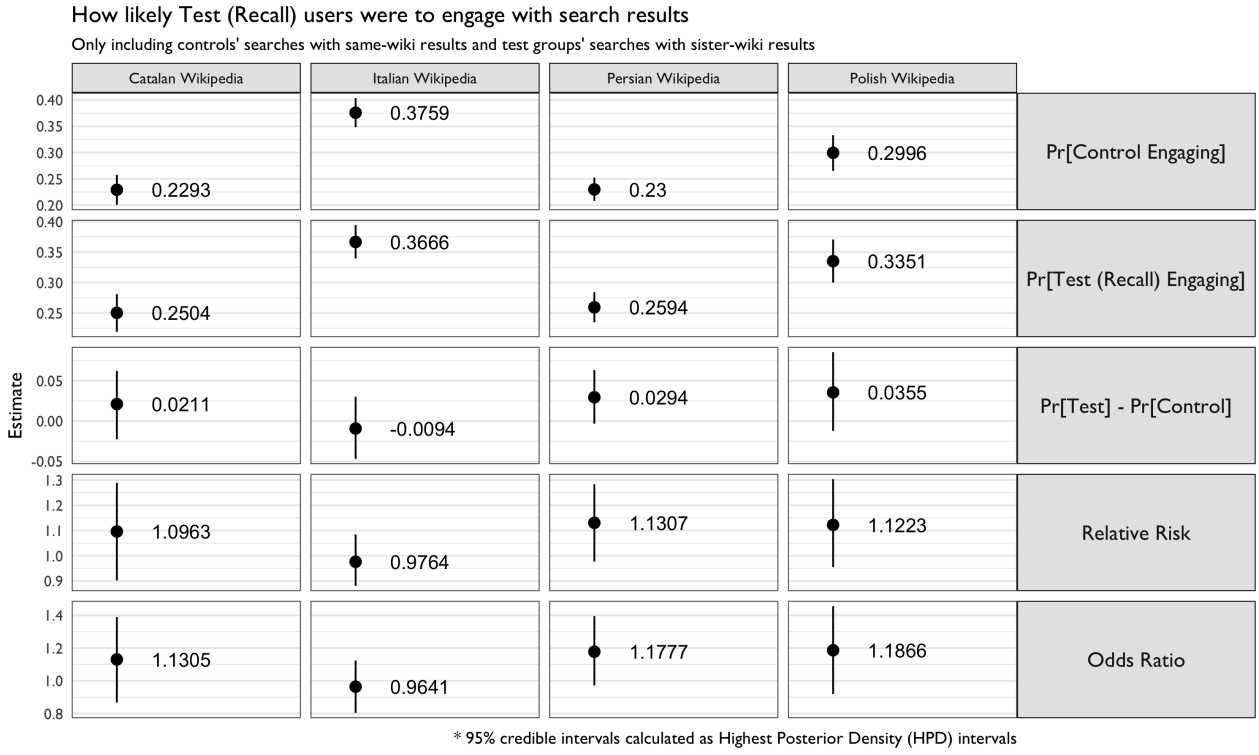


***Figure 8**: Comparison of the Control group with the Test (Recall) group.*

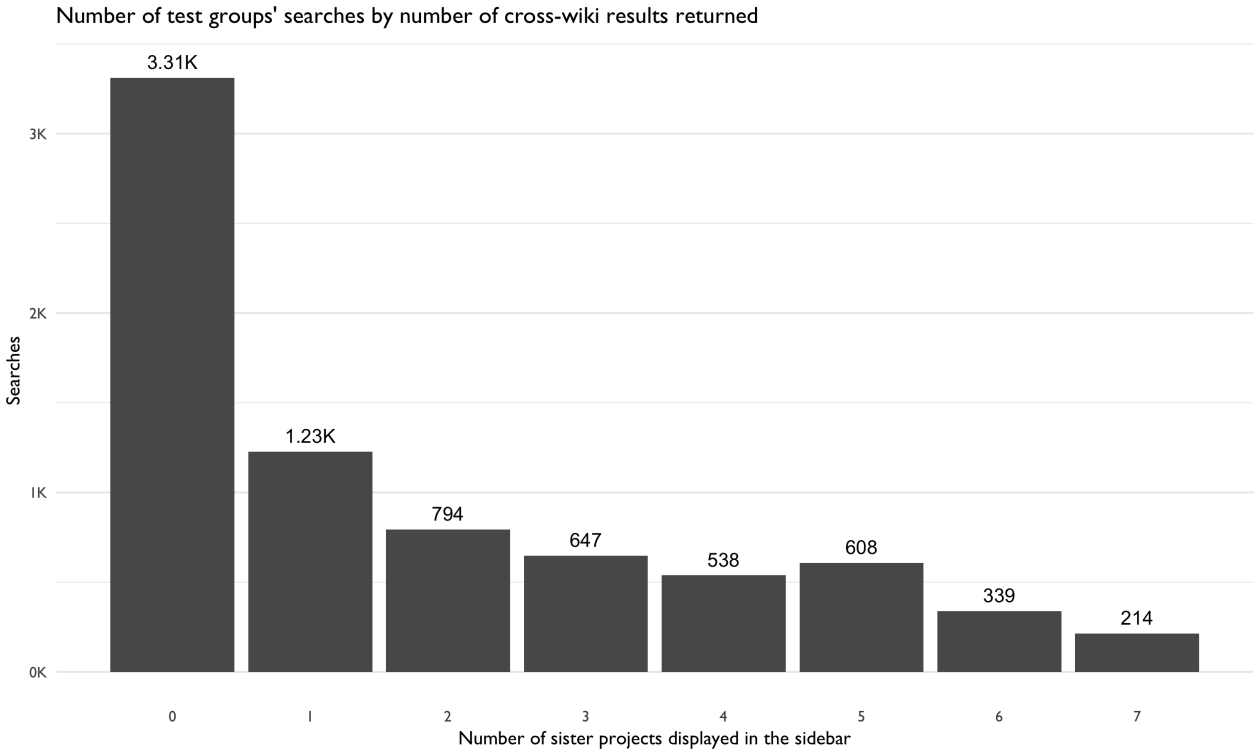


***Figure 9**: How many of the two test groups' searches returned cross-wiki results from 0 (none) - 7 (all) sister projects.*

results (same-wiki or cross-wiki) than the Control group. For example, on Catalan Wikipedia, users in the Test (Random) are 1.103 times more likely to click on a result than users in the Control group. While most of the estimates are greater than 1 (suggesting more relevant results), the 95% credible intervals contain 1, meaning we do not have sufficient evidence to draw definitive conclusions.

Discussion

As can be seen in Figure 1, the cross-wiki results were displayed in black rather than the standard blue. This issue is tracked in T158935. We cannot estimate the effect this may have had on the results of the test, but we suspect this may have had a considerable negative effect because the results did not look like click-able links.

We also suspect that the high zero results rate for each of the sister projects for these four languages may have been responsible for the few sister-project clicks. In Figure 9, relatively few searches had more than 3 cross-wiki search results.

Furthermore, since the users did not see more than the top result from each sister project, it is possible they did not event want to bother with viewing more results since the one they were shown was not relevant. Ideally, the first result would always be the most relevant one, but that is not the case, and sometimes results further down the list are more relevant to the user's actual task.

Additionally, Multimedia results were shown in a reverse order (see T158937), but we suspect this is a minor bug that did not have an effect on the test.

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.

Mikhail Popov. *BCDA: Tools for Bayesian Categorical Data Analysis*. URL https://github.com/bearloga/BCDA. R package version 0.2.3.

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.