Swap2and3 Search Test Analysis

Erik Bernhardson *Senior Software Engineer, Wikimedia Foundation*
Trey Jones *Senior Software Engineer, Wikimedia Foundation*
Chelsy Xie *Data Analyst, Wikimedia Foundation*
Mikhail Popov *Data Analyst, Wikimedia Foundation*
Deb Tankersley *Product Manager (Analysis, Search Frontend), Wikimedia Foundation*

Wikimedia Technology's Search Platform team ran an A/B test from April 7 to April 25, 2016 to see how much our users care about the position of the result vs the actual content of the result by swapping the second and third search results. We found that position 2 had a higher clickthrough rate than position 3 in both groups, but the clickthrough rate of the 3rd result was higher in the test group than in the control group. We also found that test group users were less likely to click on the 2nd result first than the control group and were more likely to click on the 3rd result first. Based on the resultsof this analysis, we think that both position and quality matter in user behavior, but with different weights. Further experiments are needed to figure out the quantitative relationship between user behavior and these two factors.

Data

We ran this test from April 7 to April 25, 2016. Only full-text search results were affected by this test. Around half of the traffic was put into the test group randomly, where we swapped their second and third search results. The rest of the traffic was seen as control group. We collected a total of 1.2M events from 295.1K unique sessions.

As in another A/B test analysis we did before, there is an issue with the event logging that when a user goes to the next page of search results or clicks the Back button after visiting a search result, a new page ID is generated for the search results page. The page ID is how we connect click events to search result page events. For this analysis, we de-duplicated by connecting search engine results page (searchResultPage) events that have the exact same search query, and then connected click events together based on the searchResultPage connectivity.

After de-duplication, we collapsed 1.16M events into 1.04M events and 483.9K searches.

Last but not least, it is worth noting that there are some issues in our data collecting process:

- Sometimes click events were not recorded while visitPage events were. This problem was solved in June 2016 by T137262. For this analysis, we treat both click event and visitPage event as a "click"

Source code and data are available on GitHub (wikimedia-research/Discovery-Search-Test-Swap2and3)

Test group	Search sessions	Events recorded
Control	157,713	675,211
swap2and3	137,416	483,407
Total	295,129	1,158,618

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

Test group	Search sessions	Searches recorded	Events recorded
Control	157,713	279,749	599,925
swap2and3	137,416	204,167	438,227
Total	295,129	483,916	1,038,152

***Table 2**: Number of searches and events used for analysis by group after de-duplication. Each search session may have several individual searches.*

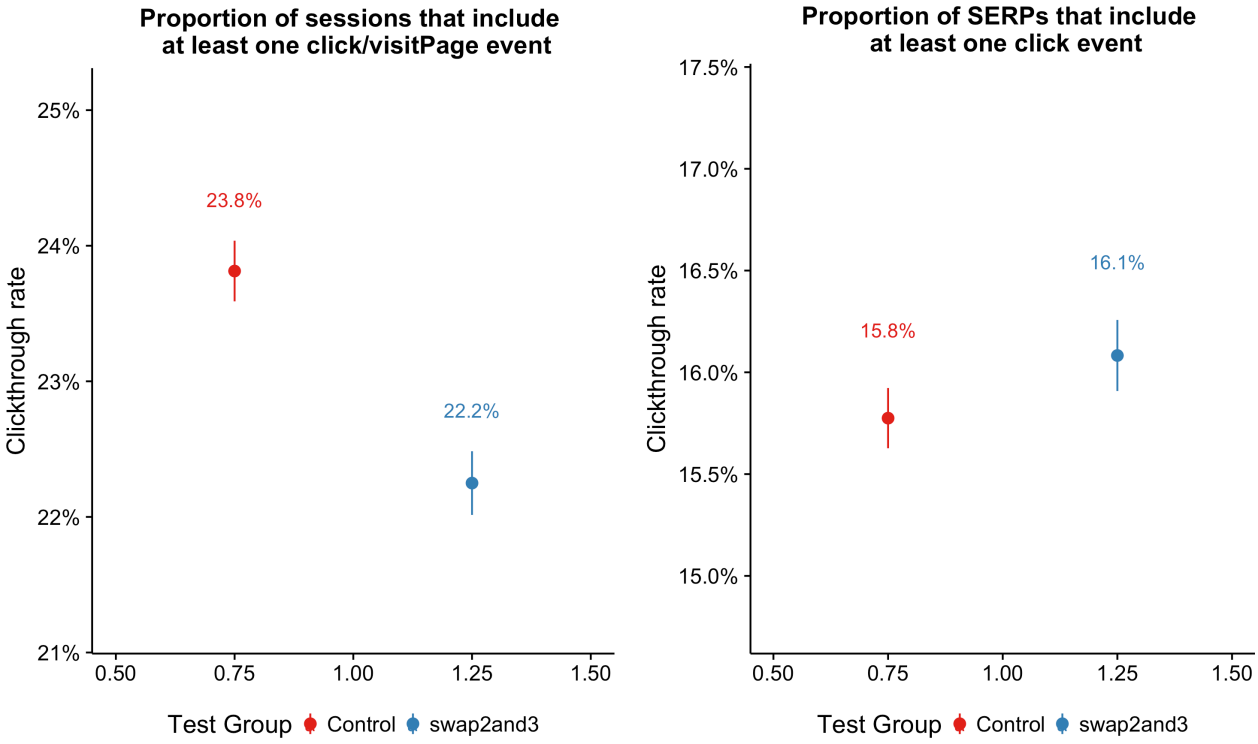

***Figure 1**: Proportion of sessions or searches that include at least one click/visitPage event by group.*

when computing clickthrough rate, i.e. if there is either a click or a visitPage event in a session, we will say there is a clickthrough in that session.

- There were 6474 out of 301,603 search sessions falling into both control and test buckets. We deleted those sessions in the data cleansing step.

Results

Engagement

Firstly, we compared the overall clickthrough rate between control and test group. We've checked that there is no significant difference in zero results rate between these two groups. The plot below shows that for each session, the control group has a significantly higher engagement rate; for each search results page, the test group has higher engagement, but the difference is not significant.

Next, we compared the clickthrough rates in second and third position by test groups. All the differences in the graph below were significant. In both control and test groups, the clickthrough rates in position 2 were higher than position 3. When comparing the same position between groups, the engagement of the 3rd result

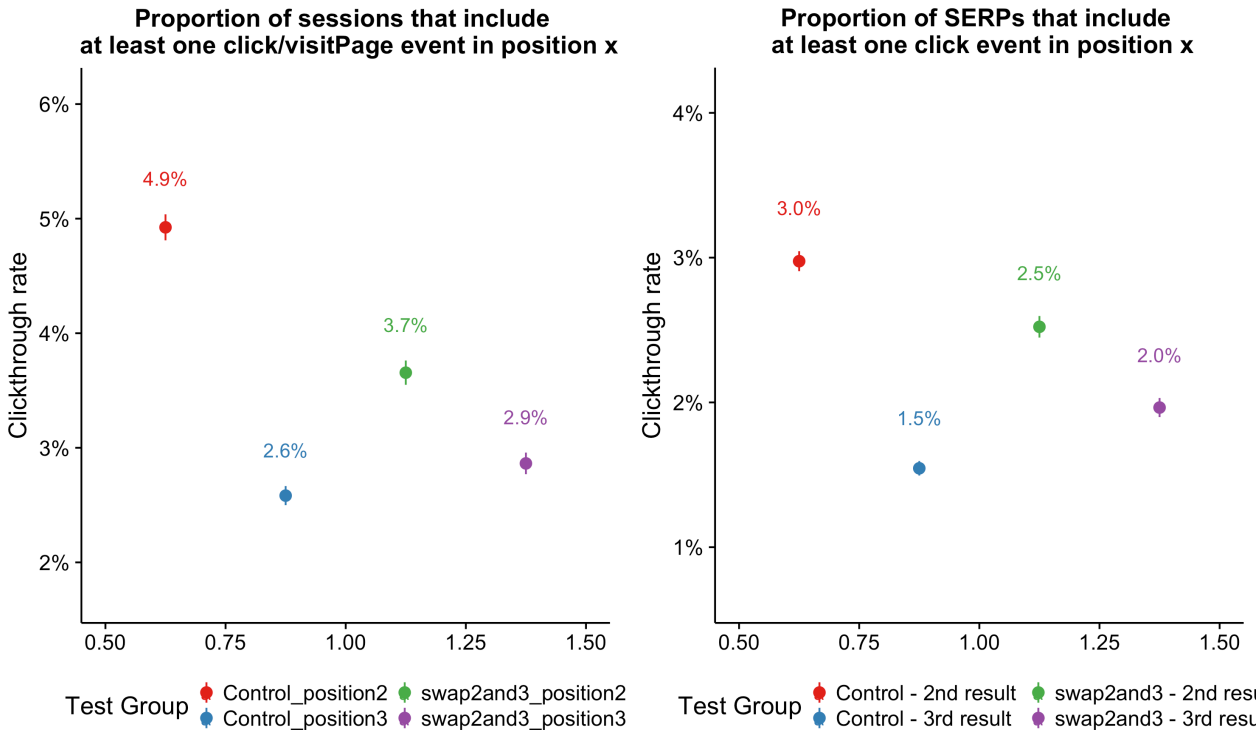

***Figure 2**: Proportion of sessions or searches that include at least one click/visitPage event by position and group.*

in the test group was higher than that in the control group, but not as high as its counterpart, the 2nd result in the control group.

First Clicked Result's Position

We can see that test group users were less likely to click on the second result first than the control group, while they were more likely to click on the third result first. There is no significant difference in other positions.

Dwell Time per Visited Page

When we compared the overall survival curves between the two groups, we found that the test group users were significantly more likely to stay longer on visited pages.

When we compared the survival curves for users who clicked on the second or third result, we found that users in the test group who clicked on the 3rd result have a longer dwell time than others, but the difference is not significant.

Scroll

We found that users in the test group were significantly more likely to scroll on the visited page.

Users in the test group who clicked on the 3rd result were significantly more likely to scroll on the visited pages than those who clicked on the 3rd result in the control group, but the differences were not significant in the 2nd result comparison and in the within-group comparisons.

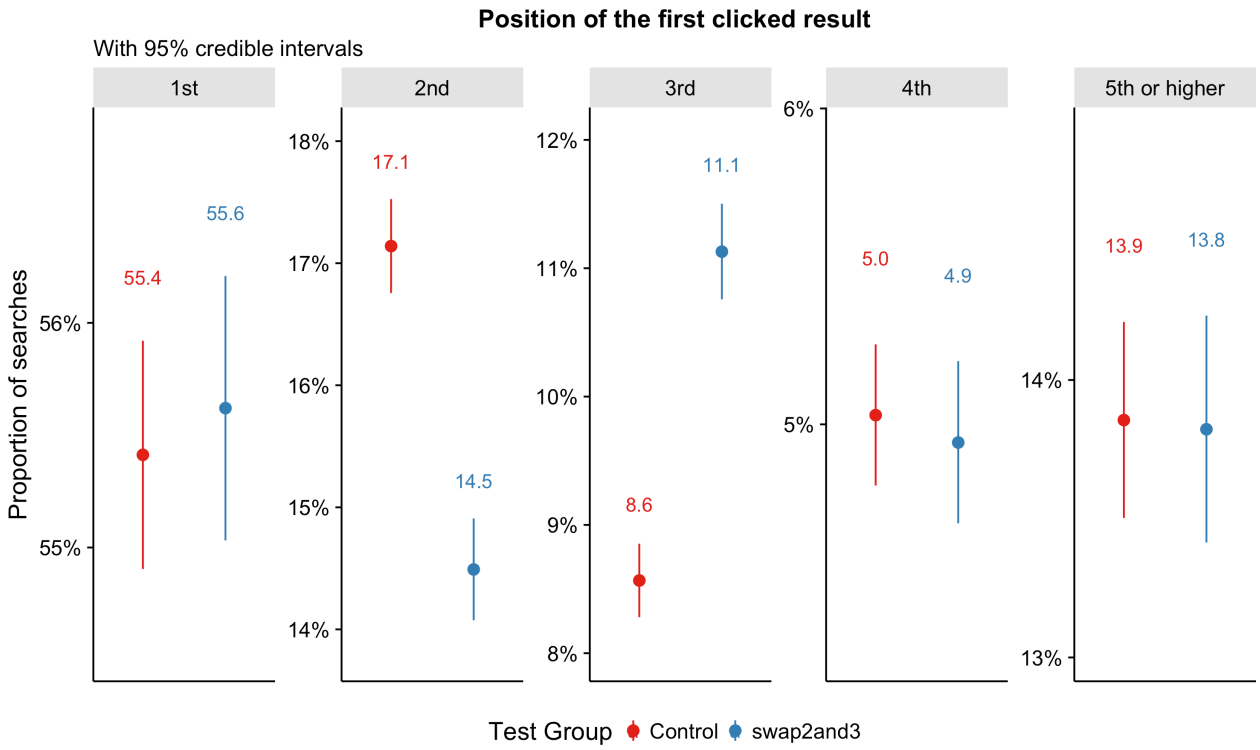

***Figure 3**: Proportion of searches that clicked a result first by position and group.*

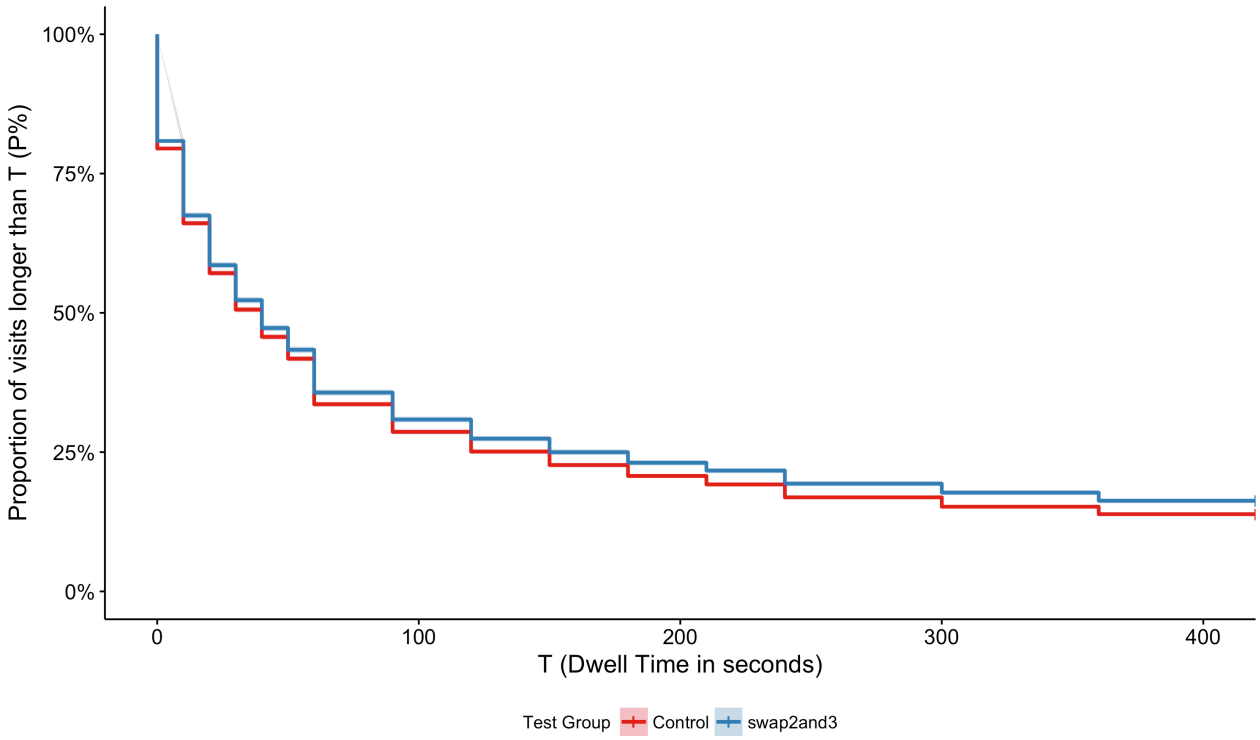

***Figure 4**: Survival curves with 95% confidence interval on visited pages after users click through, broken up by group. This shows the length of time that must pass before we lose 1-P% of the population. For example, it appears we have 80% of control group users stayed on visited pages by 10s.*

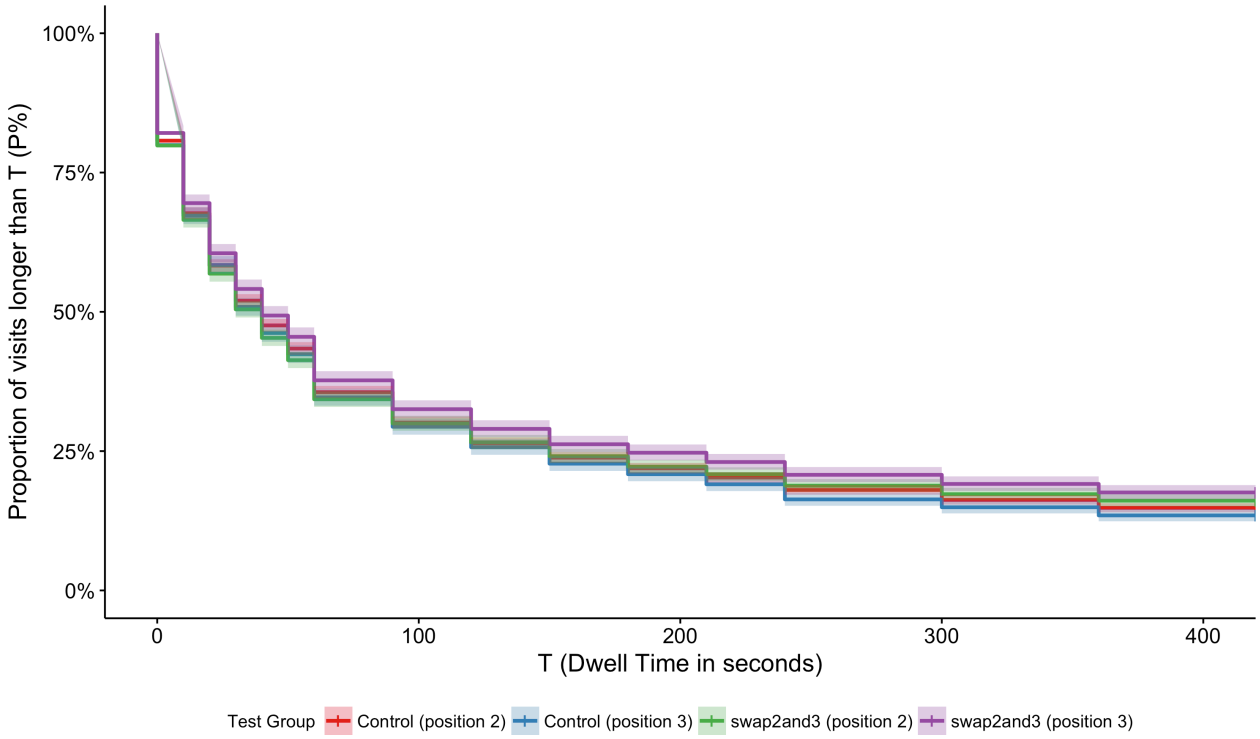

***Figure 5**: Survival curves with 95% confidence interval on visited pages after users click through, broken up by position and group.*

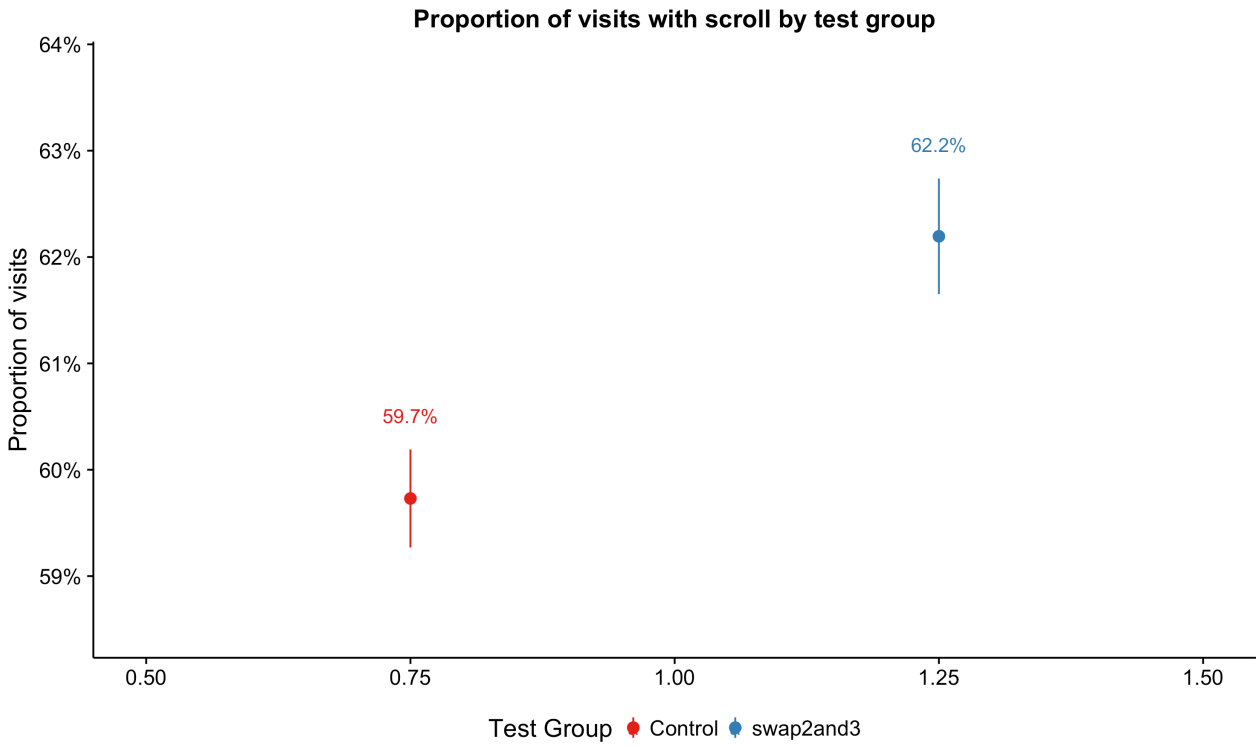

***Figure 6**: Proportion of visits with scroll by group.*

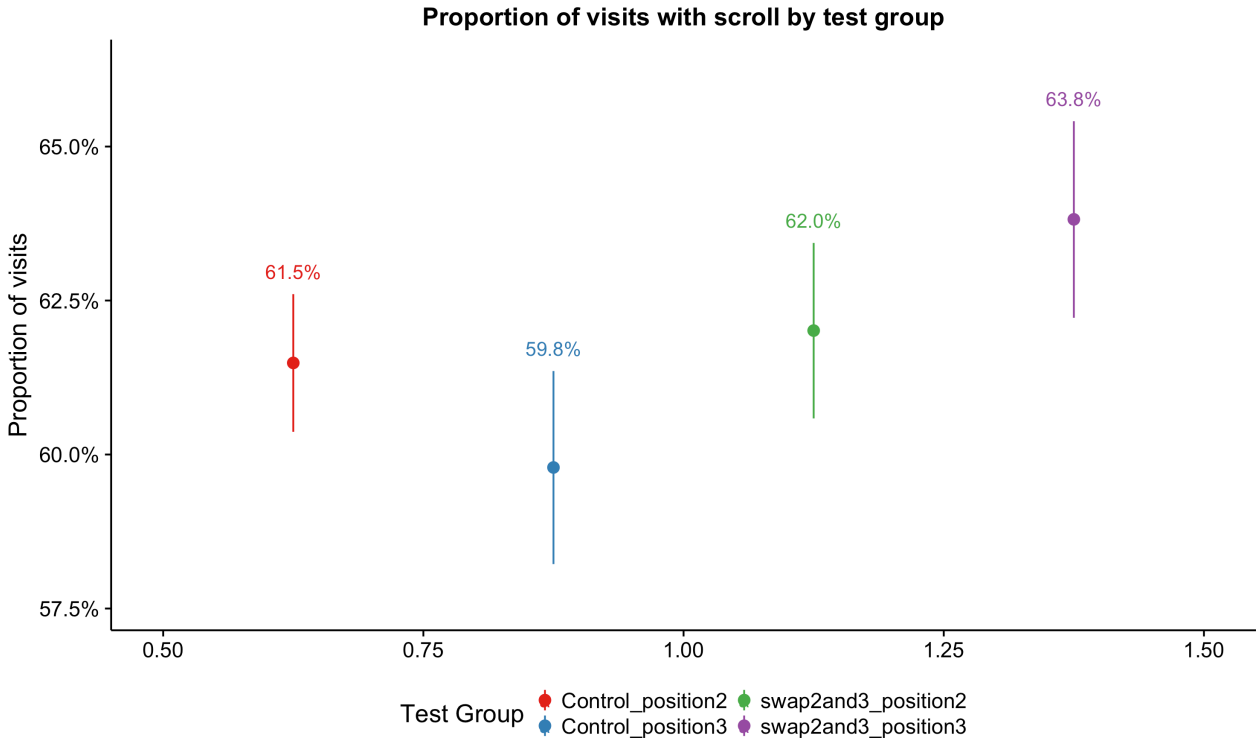

***Figure 7**: Proportion of visits with scroll by position and group.*

Discussion

Overall, we found that user engagement decreased by swapping the second and the third search results. Position 2 had a higher clickthrough rate than position 3 in both groups. However, when comparing the same position between groups the clickthrough rate of third results in the test group was higher and the clickthrough rate of second results in the test group was lower. We also found that test group users were less likely to click on the 2nd result first than the control group, while they are more likely to click on the 3rd result first. Based on these analysis results, we believe that it is not an either/or situation with position (order displayed) and "quality" (as determined by Cirrus). We suspect that *both* position *and* quality matter in user behavior, but with different weights. Further experiments are needed to figure out the quantitative relationship between user behavior and these two factors.

It seems a little bit counter intuitive that the group with a lower engagement rate tended to stay longer and was more likely to scroll on the visited page, but we saw a similar case in another test before. This may be a case of the self-selection bias that users who click through in a probably less relevant group – the test group – may be different from users who click through in the control group in motivation, experience or other factors. Further research is needed.

Additionally, instead of making explicit control buckets, we simply treated those users who didn't get assigned to the test group as control group users. We suspect that this behavior resulted in us putting all users whose browsers had a cached version of event logging JavaScript into the control group automatically, so some metrics may be biased.

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/.

Terry M Therneau. *A Package for Survival Analysis in S*, 2015. URL https://CRAN.R-project.org/package=survival. version 2.38.

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.