Wikipedia Search Phrase Rescore Boosting A/B Test

Second Edition

Mikhail Popov (Analysis & Report)
Erik Bernhardson (Engineering)
David Causse (Engineering)
Dan Garry (Product Management)
Trey Jones (Review)

April 21, 2016

Executive Summary

From 15 March 2016 to 22 March 2016 the Discovery/Search team ran an A/B test to assess how changing the phrase rescore boost from its current value of 10 to a proposed value of 1 would affect our users' behavior. Phrase rescore reorders the returned results, ranking results that have the same phrase higher. It appeared to be overboosted so we hypothesized that it may yield sub-optimal results.

It is important to note right up front that the differences in metrics/distributions between the test group (users with phrase rescore boost of 1) and the control (users with phrase rescore boost of 10) were close to 0 and were not statistically significant, even after making sure we only analyzed the eligible (affected) sessions – queries with two or more words. Even in the lab it was a very small effect (PaulScore of 0.59 to 0.60).

- The test group has a slightly higher proportion of sessions with only 1 or 2 search results pages than the control group, which has a slightly higher proportion of multi-search sessions.
- The control group had a 1.3% higher probability of clickthrough and was 1.05 (1.03–1.08) times more likely to click on a result than the test group.
- Most of the users clicked on the first result they're presented with. We had some expectation that this might change, but it is unsurprising that the two groups behaved almost exactly the same. Most of the users clicked on a search result within the first 25 seconds, with 10s being the most common first clickthrough time. This did not vary by group.
- Number of results visited did not change by much between the two groups, although it looks like a slightly larger proportion of the test group visited fewer (1-2) results than the control group (larger % of sessions with 3+ clickthroughs). It also looks like more test group users have shorter sessions than the control group, with a slightly greater number of test users having sessions lasting 10-30s and a slightly greater number of control users having sessions lasting more than 10 minutes. Users in the test group remained just a little bit longer on pages they visited than the control group, but barely so.

Putting the close-to-0-differences aside, if we take a very naive look at the differences and focus on their direction, we still cannot determine whether the change is a positive or negative (however small) impact for our users. Fewer searches may mean better results, or it may (cynically) mean users figuring out faster that they're not going to find what they're looking for. Certainly that's what a lower clickthrough rate and a slightly shorter average session length imply.

Perhaps in this particular case it may be worth making the config change decision based on how the different phrase rescore boost values affect the performance and computation time, since it doesn't appear to affect the user's behavior, at least in terms of the metrics analyzed in this report.

Backgrounds

Rescoring can help to improve precision by reordering results using a secondary (usually more costly) algorithm, instead of applying the costly algorithm to all items in the index. When there are more than two words (generally a phrase) in the query ***phrase rescore*** *function* tries to rank higher documents that have the same phrase (see MediaWiki).

The phrase rescore appeared to be overboosted to us at its current value of 10, and we hypothesized that maybe it caused the first search engine results page (SERP) to be flooded by sub-optimal results. The phrase rescore is applied to the **all field** which means if a category perfectly matches all its article have good chance to be part of the first result page. In a preliminary analysis (T128071 & these notes) using Relevance Forge (*RelForge* is also on GitHub), it appeared that changing the boost to a value of 1 resulted in 20% of the queries having different pages in the top 20, which is not indicative of a positive or a negative change, just *a* change.

So we decided to pursue this as an A/B test (T129593) to see whether the change would have a positive impact on search results with regards to various user behavior metrics. Specifically, we decided to judge the impact of the configuration change by assessing: number of searches per session, clickthrough rate, position of first clicked result, time to first clickthrough, number of results visited (pages opened), session duration, and time spent on visited page(s).

We ran the test from 15 March 2016 to 22 March 2016 on 0.5% of the users searching various wiki projects. This test did not affect autocomplete (prefix/completion suggester) and did not have any effect on zero results rate, hence the use of event logging to collect user behavior metrics. After removing known spiders, the dataset we analyzed in this report contained 90,262 independent full-text search sessions with 2 or more words in the search query. (The configuration change would not have impacted single-word queries, see T132077.)

Statistical analysis of impact on user behavior metrics

Number of searches per session

To compare the these two discrete distributions, we used bootstrapping with the Kullback-Leibler divergence metric, which allowed us to assess the probability of obtaining a K-L statistic equal to or greater than the K-L statistic observed. The observed K-L value was 0.0247 (close to 0, signifying similarity) with a 95% Confidence Interval of (0.015, 0.025). The p-value was 0.966 so we cannot reject the hypothesis that the two distributions are the same.

Clickthrough rate (CTR)

Table 1: Comparison of Test Group vs Control Group (Baseline).

n_B	n_A	π_B (%)	π_A (%)
39.5K	44K	24.3 (23.8, 24.7)	25.5 (25.1, 25.9)
% Diff (B vs A)	**Relative Risk (B vs A)**	**Odds Ratio (B vs A)**	
-1.3 (-1.9, -0.7)	0.95 (0.93, 0.97)	0.93 (0.90, 0.96)	

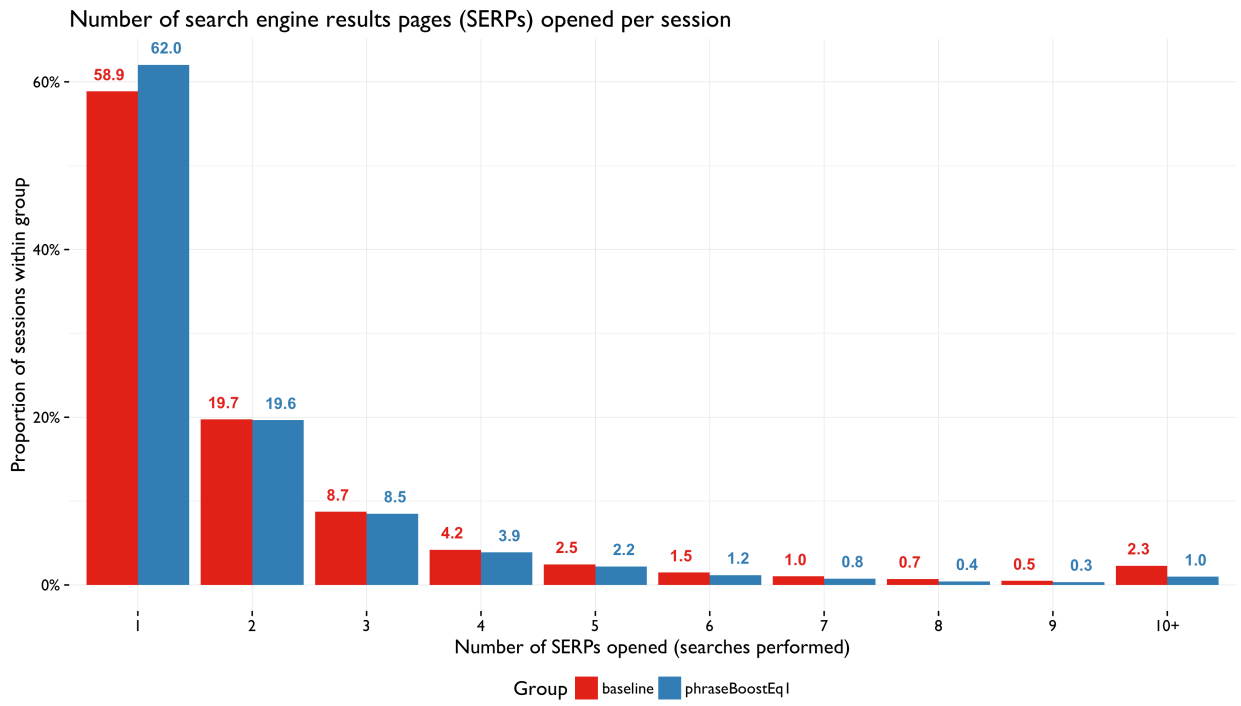

Figure 1: Number of searches made per session. The test group has a slightly higher proportion of sessions with only 1 or 2 search results pages than the control group, which has a slightly higher proportion of multi-search sessions.

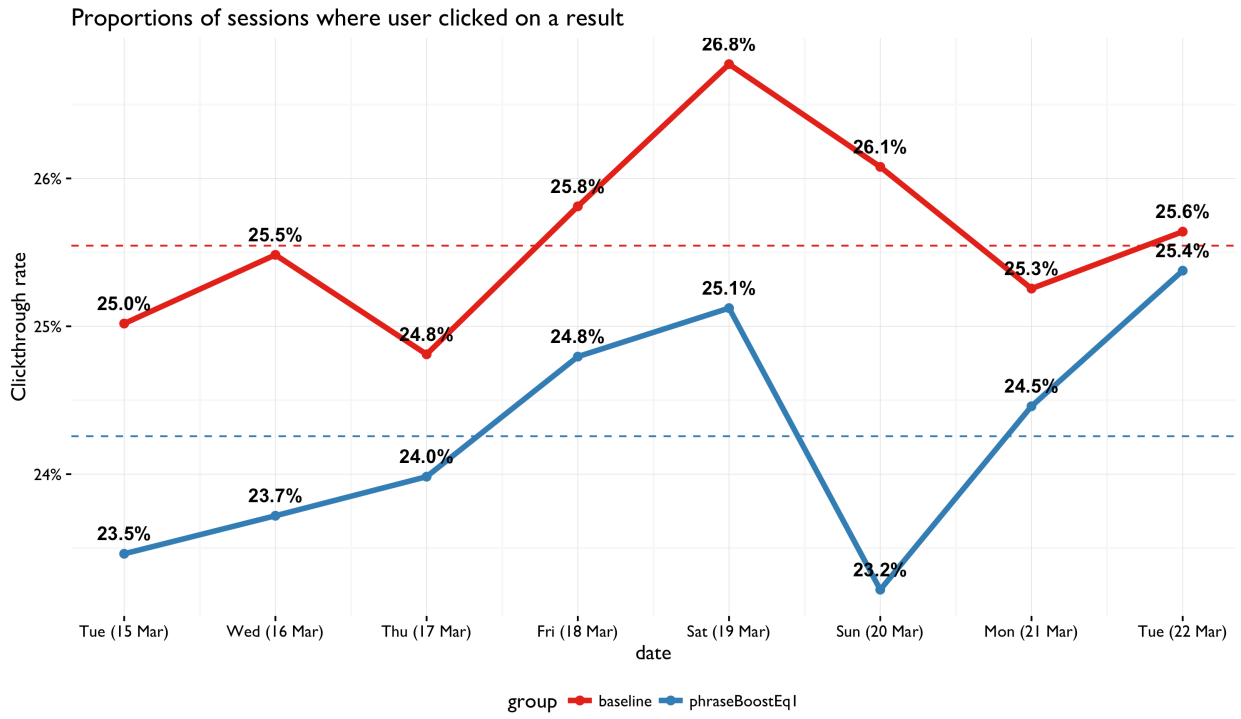

Figure 2: Daily clickthrough rate (CTR) for users who received some results. The control group had a 1.3% higher probability of clickthrough and was 1.05 (1.03-1.08) times more likely to click on a result than the test group.

Position of first clicked result

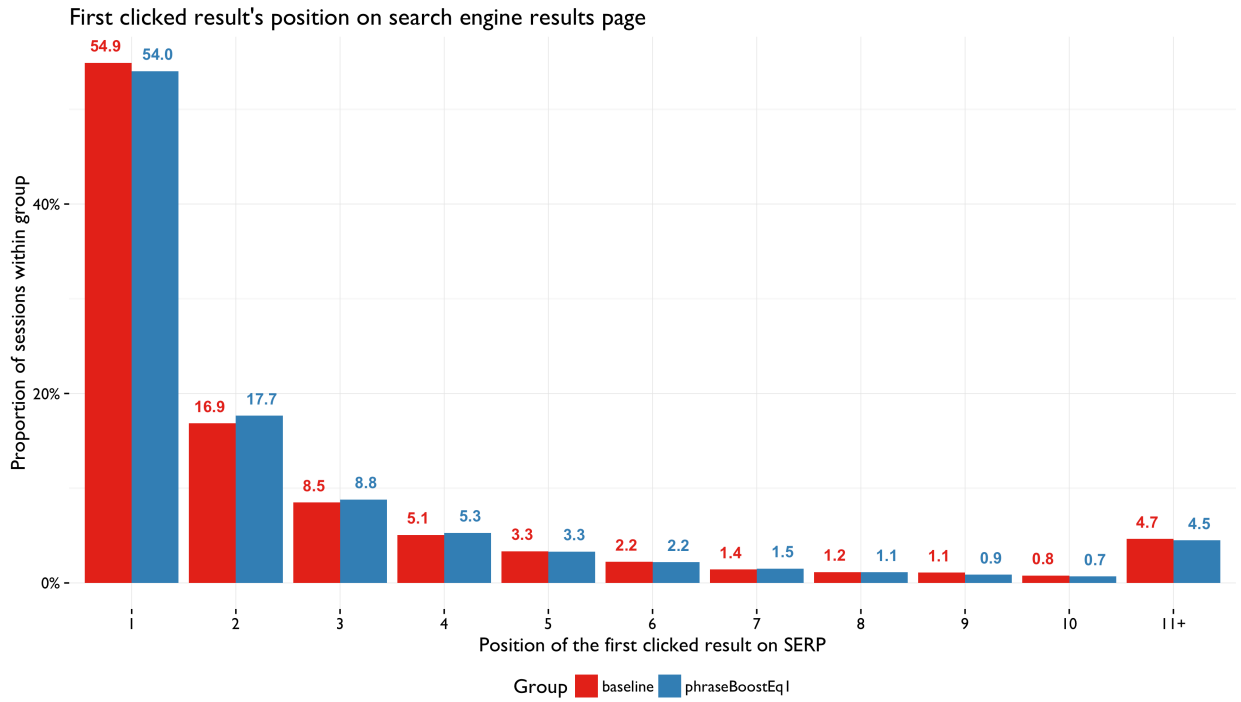

Figure 3: Position of the first clicked result on the search results page. Most of the users click on the first result they're presented with.

To compare the these two discrete distributions, we used bootstrapping with the Kullback-Leibler divergence metric, which allowed us to assess the probability of obtaining a K-L statistic equal to or greater than the K-L statistic observed. The observed K-L value was 0.018 (close to 0, signifying similarity) with a 95% Confidence Interval of (-0.002, 0.016). The p-value was 0.987 so we cannot reject the hypothesis that the two distributions are the same.

Time to first clickthrough

We performed the Kolmogorov-Smirnov Test to check if the two continuous distributions of time to first clickthrough were different between the two groups and found that we **cannot reject** (p-value $= 0.6469$, $D = 0.011$) that the times were drawn from the same distribution – in other words, that the two distributions are the same.

Number of results visited

To compare the groups' number of results visited distributions, we bootstrapped the Kullback-Leibler divergence metric. The observed K-L value was 0.0096 (close to 0, signifying similarity) with a 95% Confidence Interval of (0.003, 0.010). The p-value was 0.894 so we cannot reject the hypothesis that the two distributions are the same.

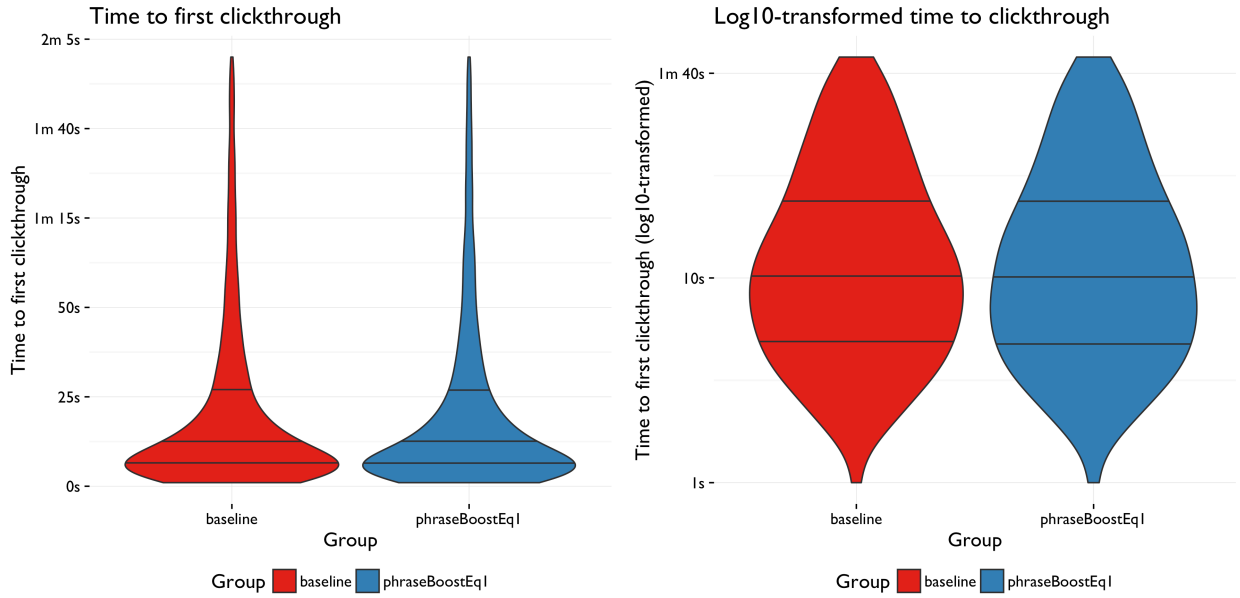

Figure 4: Most of the users click on a search result within the first 25 seconds, with 10s being the most common first clickthrough time.

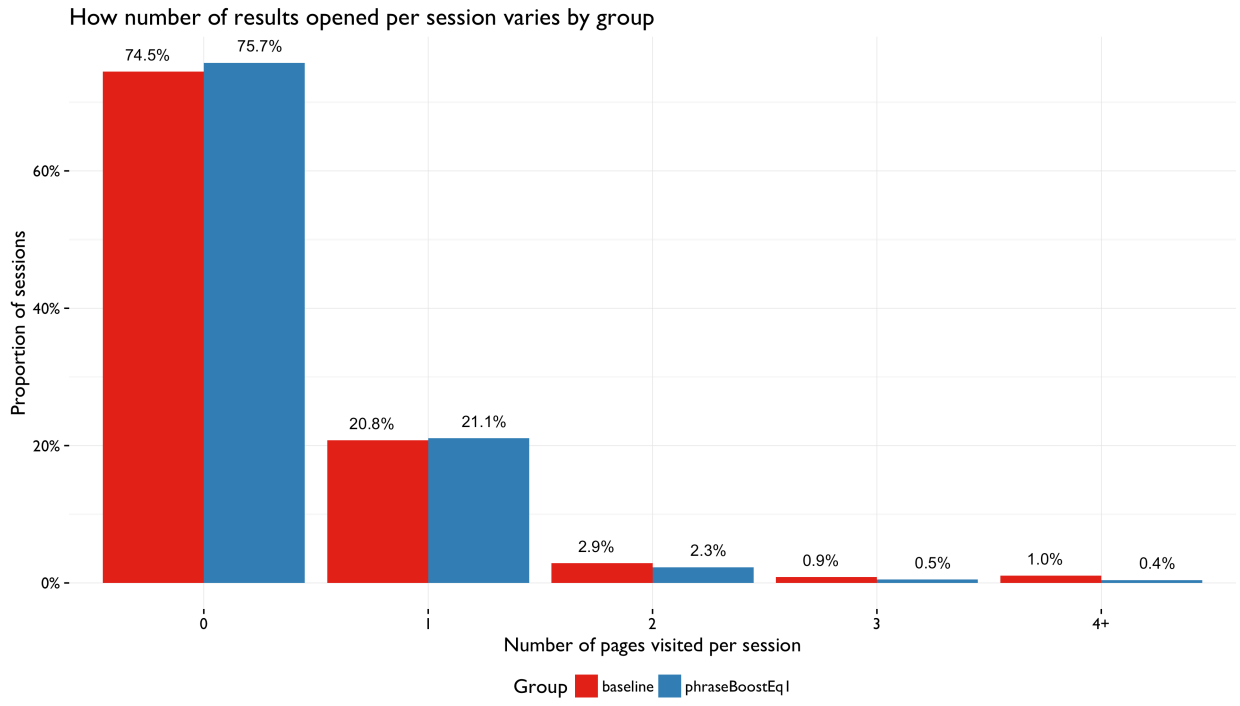

Figure 5: Number of results opened per session. 0 pages visited corresponds to session abandonment rate – aka the inverse of clickthrough rate.

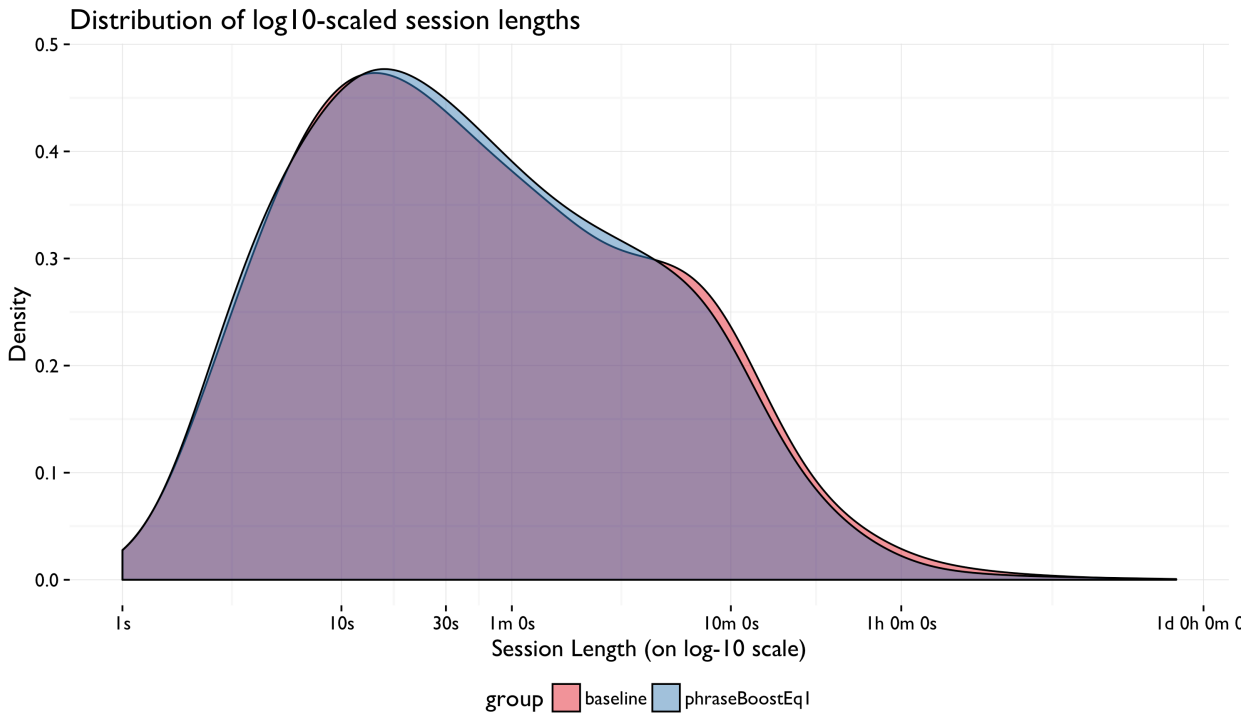

Figure 6: Session length peaks between 10s and 30s, with most of the sessions wrapping up around the 10 minute mark.

Search session duration

Due to the size of the dataset, we could not perform a straight Kolmogorov-Smirnov Test to compare the two distributions (larger sample sizes yield smaller p-values, yielding statistically significant difference where there is no practical difference). As before, we used *bootstrapping* but instead applied it to the Kolmogorov-Smirnov test statistic. The observed K-S statistic was 0.017 – with a 95% C.I. (0.010, 0.022) – and the bootstrapped p-value was 0.615, which means we do not have sufficient evidence of the groups' session length distributions being significantly different.

Time spent on pages

Figure 7 and Tables 2–4 in the Appendix show the % of each group remaining on the first visited page / randomly chosen visited page / last visited page. The two groups behave very similarly, with the test group staying just little a bit longer on visited pages than the control group.

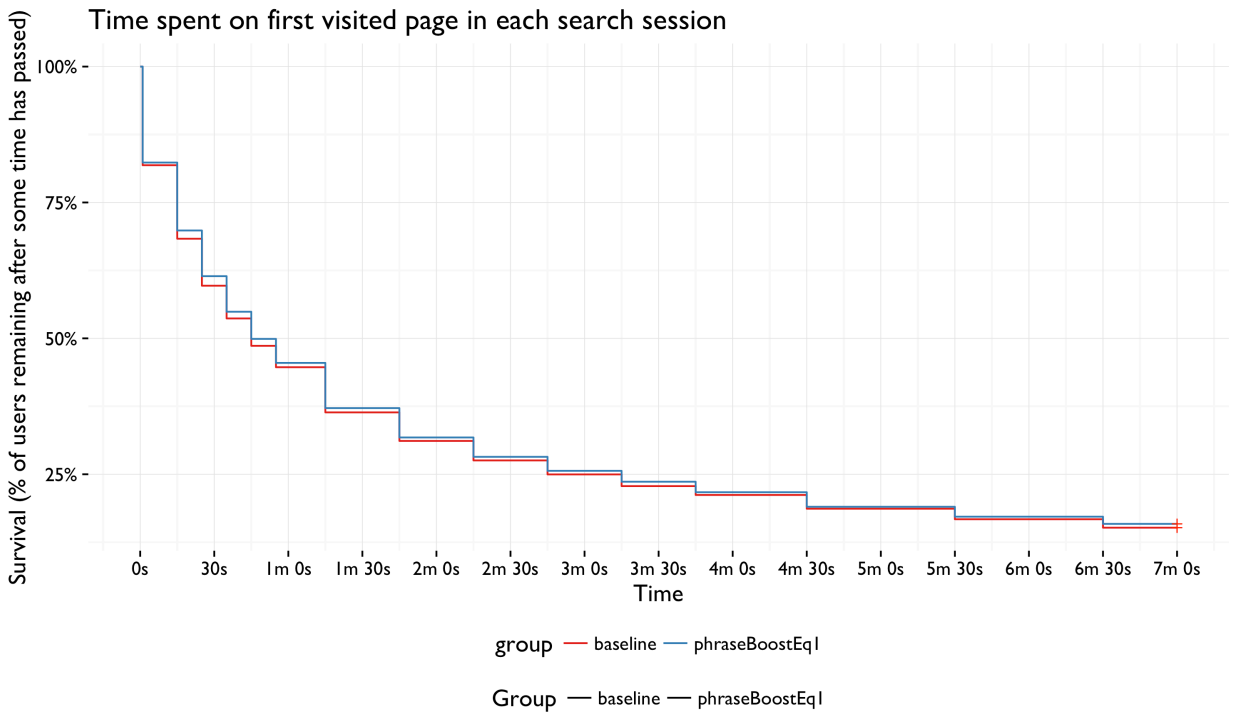

Figure 7: The test group stayed a tiny bit longer on the first result they clicked on than the control group, on average.

Discussion

Across the board, the users who received the new boost configuration behaved the same as the control group. Whether you look at number of searches per session, the clickthrough rate, position of first clicked result, time to first clickthrough, number of results visited (clickthroughs per session), search session duration, and time spent on visited pages, there are tiny, insignificant differences – both from statistical and practical perspectives.

Putting the close-to-0-differences aside, if we take a very naive look at the differences and focus on their direction, we still cannot determine whether the change is a positive **or** negative (however small) impact for our users. It is unclear how to interpret the differences. For example, the test group had a slightly higher proportion of sessions that only had 1 or 2 SERPs than the control group, which had a slightly higher proportion of multi-SERP sessions. Fewer searches may mean better results, or it may (cynically) mean users figuring out faster that they're not going to find what they're looking for. Certainly that's what a lower clickthrough rate and a slightly shorter average session length imply.

Appendix

Table 2: Time spent on first visited page.

Time (s)	% of baseline	% of phraseBoostEq1	Difference
1	81.86%	82.34%	barely higher % of test group remained on page
15	68.33%	69.85%	barely higher % of test group remained on page
25	59.69%	61.44%	barely higher % of test group remained on page
35	53.67%	54.90%	barely higher % of test group remained on page
45	48.63%	49.90%	barely higher % of test group remained on page
55	44.69%	45.49%	barely higher % of test group remained on page
75	36.38%	37.17%	barely higher % of test group remained on page
105	31.13%	31.77%	barely higher % of test group remained on page
135	27.55%	28.21%	barely higher % of test group remained on page
165	24.97%	25.63%	barely higher % of test group remained on page
195	22.82%	23.63%	barely higher % of test group remained on page
225	21.19%	21.70%	barely higher % of test group remained on page
270	18.66%	19.02%	barely higher % of test group remained on page
330	16.73%	17.20%	barely higher % of test group remained on page
390	15.18%	15.87%	barely higher % of test group remained on page

Table 3: Time spent on randomly selected visited page (useful when there were several pages opened).

Time (s)	% of baseline	% of phraseBoostEq1	Difference
1	82.16%	82.44%	barely higher % of test group remained on page
15	68.85%	70.06%	barely higher % of test group remained on page
25	60.34%	61.62%	barely higher % of test group remained on page
35	54.14%	55.09%	barely higher % of test group remained on page
45	49.27%	50.11%	barely higher % of test group remained on page
55	45.40%	45.92%	barely higher % of test group remained on page
75	37.22%	37.62%	barely higher % of test group remained on page
105	32.08%	32.27%	barely higher % of test group remained on page
135	28.45%	28.74%	barely higher % of test group remained on page
165	25.90%	26.29%	barely higher % of test group remained on page
195	23.76%	24.21%	barely higher % of test group remained on page
225	22.14%	22.45%	barely higher % of test group remained on page
270	19.53%	19.80%	barely higher % of test group remained on page
330	17.62%	17.96%	barely higher % of test group remained on page
390	16.06%	16.63%	barely higher % of test group remained on page

Table 4: Time spent on last visited page.

Time (s)	% of baseline	% of phraseBoostEq1	Difference
1	82.88%	82.94%	barely higher % of test group remained on page
15	69.75%	70.69%	barely higher % of test group remained on page
25	61.17%	62.30%	barely higher % of test group remained on page
35	55.23%	55.97%	barely higher % of test group remained on page
45	50.31%	50.97%	barely higher % of test group remained on page
55	46.51%	46.75%	barely higher % of test group remained on page
75	38.42%	38.42%	barely higher % of test group remained on page
105	33.29%	33.31%	barely higher % of test group remained on page
135	29.62%	29.67%	barely higher % of test group remained on page
165	27.12%	27.21%	barely higher % of test group remained on page
195	24.93%	25.20%	barely higher % of test group remained on page
225	23.25%	23.45%	barely higher % of test group remained on page
270	20.57%	20.72%	barely higher % of test group remained on page
330	18.60%	18.88%	barely higher % of test group remained on page
390	16.98%	17.53%	barely higher % of test group remained on page