How important are user-generated data for search result quality? Experimental evidence

Supplementary materials

Abstract

This document contains five appendices. Appendix A provides background information on the Cliqz search engine. Appendix B provides details on the experiment. Appendix C describes how we used human assessment to rate the quality of search results. Appendix D contains further details on the instructions we gave to the assessors. Appendix E contains robustness checks.

A The Cliqz search engine

Cliqz was a privacy-oriented web browser and search engine developed by Cliqz GmbH and majority-owned by Hubert Burda Media. It was available as a desktop and mobile web browser as well as an extension for Firefox itself.[1] Since 2013, Cliqz has built an own search index (a list of all URLs that exist, including tags for each URL describing its contents). In 2015, an anonymity-oriented browser was released (a fork of Mozilla Firefox) that included a search functionality. Since 2017, 1% of Firefox downloads in Germany came with a Cliqz extension in order to collect more anonymized user information.

On April 27, 2020, the experiment reported in this paper was conducted. On April 29, 2020, Cliqz announced that it will shut down its browser and search engine on May 1, 2020.[2]

[1] See https://en.wikipedia.org/wiki/Cliqz for some detail.

[2] See https://cliqz.com/announcement.html for their final annoucement.

B Details on the Cliqz experiment

Overview. We randomly drew 1,000 queries, respectively, in 5 buckets from the population of queries submitted on the Human Web (see below). Then, we conduct the experiment by obtaining search results for each query at 12 levels of available data on past searches. This leaves us with a data set consisting of 60,000 result sets. We augment this data set with 5,000 result sets from Google and Bing, respectively.

Human Web. The Human Web is a software integrated in the Cliqz browser or, alternatively, a software extension to Mozilla Firefox. It allows for the anonymous collection of user browsing activity and user-generated query logs. For example, if a user of a Cliqz browser – or a Firefox browser with installed Cliqz extension – searches for "ebay auto" using Google, Bing, Cliqz, or any other search engine, the information on the search, the results and choices made by the user were transferred in an anonymized manner to Cliqz. Hence, these search queries represent all searches on any search engine for that subpopulation of users.

Mozilla, as part of another experiment, installed the Cliqz software extension for a 1% random sample of all Firefox downloads in Germany starting in October 2017.[3] This makes the population of the Human Web users somewhat more representative of the general German population than the population of users of the Cliqz browser.

Sample of queries. In online search, very few queries are searched many times, while many others are searched only very rarely. To account for this, we ordered all queries that were submitted on the Human Web between April 20 and April 26, 2020, by their frequency, from the most popular to those that appeared only once within the week. Then, we formed five buckets using the following thresholds: 0.2%, 1%, 5%, and 25%. This means that, for example,

[3] See ZDNet article and Human Web blog.

Table B.1: Example of query logs

query	**clicked URL**
google	http://www.google.com
wnmu	http://www.wnmu.edu
ww.vibe.com	http://www.vibe985.com
www.accuweather.com	http://www.accuweather.com
weather	http://asp.usatoday.com
college savings plan	http://www.collegesavings.org
pennsylvania college savings plan	http://www.patreasury.org
pennsylvania college savings plan	http://swz.salary.com

Notes: Taken from Cliqz blog 0x65.dev and AOL query logs dataset.

the first bucket represents the top 0.2% of search queries by frequency, while the last bucket represents the last 75%. Next, we randomly drew 1,000 queries from each of the 5 bucket, leaving us with a stratified sample of 5,000 queries.

Index, query logs, and query log counts. Like other search engines, the Cliqz search engine relied on two main input components. The first one is their own index of webpages, which is generated by crawling the web to maintain the up-to-date directory of all webpages.

The second input is the data on user-generated query logs, i.e., actual user queries linked to the URLs they clicked on. Table B.1 provides an example of several query logs. These data are useful because past choices of users might be predictive of future choices. Hence, the search engine may want to put the most clicked result in the past at the top of the new search results.

Query logs are aggregated into query log counts. These say how many times a given URL was clicked by users who searched for a given query. These are the raw data.

Starting from those, Cliqz performed semantic analysis to also use information from its own index of webpages. This allows Cliqz to use the data more efficiently. For example, if someone searches for “Lady Gaga best hits”, the search engine also uses the query log counts from other

similar queries such as "Lady Gaga best songs" or "Lady Gaga hits".[4] This is held fixed in our experiment, in the sense that the algorithm is not re-trained when less data is available.

The experiment. The experiment with the Cliqz search engine was conducted in the evening of April 27, 2020. For each of the 5,000 sampled queries, Cliqz obtained search results at different fractions of the query log counts. Thereby, we simulate the counterfactual search engine results at different availability of user-generated data.

Specifically, Cliqz provided results at twelve different levels of data on past searches: 100% (or full data), 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10%, 1%, and 0.1%. To obtain respective query log counts, we multiply the query log counts by the assumed fraction of available data and take the floor of that value as the new log counts. For example, if a given query/URL pair has a count of 10 (i.e., people who searched for that query clicked on that URL ten times in the past), then the new count for that query/URL pair would be 5 under 50% of user data availability: 1 under 10%, and 0 under 1% or below. Hence, if the Cliqz search engine would only have 1% of its actual user data, it would completely miss that query/URL pair.

Table B.2 provides an example. The left part shows the query log counts for full data (top) and half of the data (bottom, obtained in the way that was described above). We can see that the query log counts for all but two URL's are lost when we move from the top to the bottom panel. Search results for a new query, query 4, are generated for the full data (top right) and half of the data (bottom right), using the respective query log counts for query 1, 2, and 3. We can see that for half of the data, URL3 and URL4 are not anymore part of the results, as the query log counts for those URL's become zero (bottom left vs. top left). Also, we can see that the order of search results is affected.

[4]More details of how Cliqz search engine works are at https://0x65.dev/blog/2019-12-06/building-a-search-engine-from-scratch.html.

Table B.2: Example of removing 50% of user data and its effect on search results

query	**URL**	**count (100%)**
query 1	URL 1	5
query 1	URL 2	1
query 1	URL 3	1
query 2	URL 2	5
query 2	URL 3	1
query 2	URL 4	1
query 3	URL 4	1

algorithm $\Rightarrow$

query	**search results**
query 4	1. URL2
	2. URL1
	3. URL3
	4. URL4

query	**URL**	**count (50%)**
query 1	URL 1	2
query 1	URL 2	0
query 1	URL 3	0
query 2	URL 2	2
query 2	URL 3	0
query 2	URL 4	0
query 3	URL 4	0

algorithm $\Rightarrow$

query	**search results**
query 4	1. URL1
	2. URL2

Notes: This table illustates how the algorithm generates search results at full data and at half the data. The search results on the right are for a new query. Removing data affects the search results because it affects the query log count.

Search results. For each query and each level of data, the Cliqz algorithm produces two sets of search results, each consisting of a set of ranked URLs: organic search results and news search results. For example, at the time of the experiment, searches for "Kim Jong-un" – the supreme leader of North Korea – were popular due to rumours of his death. Hence, the Cliqz search engine provided two sets of results: the usual organic search results that also contained some with recent news, and separate consisting only of news. Our analysis uses only the organic search results. Out of the original 5,000 queries, Cliqz provided news for 47 queries. Most of them are popular queries. For our sample of 500 queries that were assessed, only 16 had separate results related to recent news, 8 each in the two most popular categories. This suggests that the separate news category that we ignored was not very important.

Data used for analysis. Cliqz provided us with 60,000 search result sets, one for every query at every different fraction of query logs.

We also collected search result sets from Google and Bing for the same 5,000 queries. For this we used the application programming interface (API) of a for-pay service called SerpAPI (see https://serpapi.com/), for http://www.google.com and http://www.bing.com. The API allowed us to specify that we would like to obtain results for users from Germany.

C Human Assessment

Why not click-through-rate? Other papers, for instance (*1*) and (*2*) , use the click-through-rate (CTR) as a measure of quality. The CTR is the likelihood that a user clicks on one of the search result of a given set. We do not use this measure for two reasons. First, we would like to make comparisons across search engines and do not have access to data on CTR's from Google and Bing. Second, we create artificial search results in our experiment, which were never shown to actual users. For this reason, there are no data on CTR's. In principle, Cliqz could have shown the results to randomly drawn users and record the CTR, but did not want to do so, because it would lower their user experience.

Sample of queries. Recall that our data set consists of 60,000 result sets for Cliqz (5,000 for 12 levels of data) and 5,000 result sets each for Google and Bing (only full data, as we did not have the opportunity to conduct an experiment with them).

Since human assessment is costly, we use a random sample of the 5,000 sampled queries for evaluation. We restrict attention to queries which are either in German or in English and that are at least 3 characters long. Then, out of 3,918 queries, we sample 500 randomly: we draw 50 queries from buckets 1 and 2 each, 100 queries from buckets 3, and 150 queries from each of buckets 4 and 5. We over-sample rare queries (buckets 4 and 5) to reduce possible noise as we expect that rare queries might be more difficult to assess. After sampling, we remove 7 queries with inappropriate content, resulting in 493 queries for human assessment.

Top-5 results and mixed result sets. Previous studies (*3, 4*) show that search engine users usually look only at the results that appear at the top of the result list. In order to reduce the load on the assessors, we therefore restrict the result sets only to top-5 results.

Additionally, for each sampled query, we construct a "mixed" result set from Google, Bing,

and Cliqz (at full data) result sets, using the following algorithm:

1. **Assign order:** randomly map Google, Bing, and Cliqz result sets to set_1, set_2, set_3;
2. **Pop the first element:** add the first URL (i.e., result) of set_1 to the mixed result set, and remove that URL from set_1, and also from set_2 and set_3, if those sets also contain that URL;
3. **Rotate the order:** make set_2 to be the new set_1, set_3 to be the new set_2, and set_1 to be the new set_3;
4. **Repeat steps 2 and 3** until the mixed set has 5 elements;
5. **Shuffle the mixed set:** randomize the order of results within the mixed set.

By randomizing the order with which Google, Bing, and Cliqz result sets contribute to the combined mixed set, we ensure that all search engines get an equal chance to contribute to the mixed set for each query. For example, if all three result sets – Google, Bing, and Cliqz – are distinctly different, the union of the top-5 results will give 15 results in total. However, the mixed set is limited for 5 results only. Hence, those search engines that have been chosen to be the first two to contribute to the mixed set contribute two results each, while the last one will contribute only one. But which search engine is chosen to be the first is random, therefore, the mixed sets on average provide equal opportunities to every search engine. By randomizing the final order of the results in the mixed set, we also remove any residual correlation in the positions of results supplied by the same search engine in the mixed set.

Assessors. In order to measure the quality of these result sets, we asked human assessors to rate their satisfaction with the search results on a seven-level Likert scale for a random sample of queries. We hired two research assistants (RA's) at Tilburg University and 587 people in

Germany (37% women, median age 34) through the *clickworker.com* platform to perform the assessment. One of the research assistants received all the result sets corresponding to queries in German language: and another, to queries in English (each assessor was proficient in the relevant language). 563 clickworkers provided evaluations, on average for fifteen result sets. In total, each of the $2,848$ result sets was evaluated on average by four different people (one RA and three clickworkers). Appendix D contains details on the instructions we gave to the RA's and the clickworkers.

In general, individual assessments of the same result set might differ from person to person, which will generate noise. However, since the assessors were unaware about which search engine had generated the results, we expect this noise to be unsystematic and to vanish for the average assessment.

Assessment. For each result set, human assessors was asked to rate the quality of the result set on a scale from 1 to 7, where 7 means "completely satisfied", 4 is "neither satisfied, nor dissatisfied", and 1 means "completely dissatisfied, as if no results." See Table C.3. The assessors were explicitly asked to take the order of the results into consideration when rating the result set.

As an alternative measure of quality, we also asked human assessors to pick the best and second-best results within each result set. The assessors could choose an option "None of the above", in case they find none of the results satisfactory. Although we collected the choices of the best and second-best results for all result sets, we were interested mostly in their choices within mixed result sets. The idea is that the assessors were not aware about the fact that they were evaluating a mixed result set. We use this to conduct a robustness check in Appendix E, where we measure which search engine produces the best result by looking at the fraction of times the best rated result from the mixed result set was produced by that search engine.

Table C.3: Likert scale

value	description
7	completely satisfied
6	mostly satisfied
5	somewhat satisfied
4	neither satisfied, nor dissatisfied
3	somewhat dissatisfied
2	mostly dissatisfied
1	completely dissatisfied, as if no results

Notes: This table shows the Likert scale we used for human assessment by the RA's and the click-workers.

Presentation. The assessors were shown the result sets simulating a browser experience, where each result showed not only the URL itself, but, in most cases, also the title and the snippet of the page (Figure D.1). The titles and snippets for Google and Bing results were provided directly by the API we used to obtain them (see above). For Cliqz results, we directly copied the title and snippet from Google or Bing results if those results also contained that URL. In this way we recovered titles and snippets for 2,166 out of 3,846 unique URLs in Cliqz results. For the remaining 1,680 URL, we queried those URLs to Google API and scraped titles and snippets provided by Google to those URLs. This helped to find titles and snippets for 1,512 URLs, leaving just 168 URLs without a match. The remaining URLs were mostly web-pages which no longer existed. We kept those 168 URLs in the result list, asking human assessors not to penalize the result simply for the absence of the title and snippet. We discuss the potential influence of the missing titles and snippets on the ratings by human assessors in the robustness checks in Appendix E.

In total, there were 3,944 result sets to be evaluated: i.e., 493 queries times eight result sets per query (mixed, Google, Bing, and Cliqz at five different fractions). However, 400 out

Table C.4: Empty result sets out of 493 queries used for human assessment.

search engine	fraction of user data	empty result sets in all queries		empty result sets in rare and rarest queries	
		number	share	number	share
Cliqz	1%	250	0.51	195	0.40
Cliqz	10%	64	0.13	54	0.11
Cliqz	20%	42	0.09	36	0.07
Cliqz	50%	28	0.06	27	0.05
Cliqz	100%	16	0.03	15	0.03
Google	100%	0	0.00	0	0.00
Bing	100%	0	0.00	0	0.00

Notes: This table shows the number and fraction of empty result sets in all queries (third and fourth column) and the number and fraction of empty result sets in rare and rarest queries (fifth and sixth column). Rare and rarest queries are from bucket 4 and 5, respectively.

of those results sets were empty: i.e., a search engine did not provide any result to the query. Unsurprisingly, empty result sets mostly occurred for rarer queries and at lower fractions of user data (See Table C.4). Moreover, out of 3,544 non-empty result sets, 696 result sets were duplicates, so we did not need to evaluate them again. The duplicate result sets are those that have the same set of URLs in exactly the same order as an already evaluated result set. Overall, there were $2,848$ result sets to be evaluated by the human assessors, net of duplicates and empty sets.

We decided not to remove empty result sets from our analysis, as it would bias severely our results. We believe that the fact that the Cliqz search engine struggled to provide results at lower fractions of user data or for rarer queries is in itself a sign of deteriorating quality. Hence, even if the actual empty result sets were not evaluated by the assessors to save costs, we restored empty sets for our analysis by imputing the lowest rating of 1 for them. We used

Table C.5: Number of assessments

evaluator	unique	with duplicates	with duplicates and zero sets
click workers	8,544	10,632	11,832
research assistant 1 (DE)	1,544	1,901	2,301
research assistant 2 (EN)	1,304	1,643	2,043
total	11,392	14,176	16,176

Notes: This table shows the number of assessments by type of evaluator. Duplicate result sets have the exact same search results in the same order. Zero result sets are empty.

the "as if no results" wording for the lowest rating, in order to anchor all the other ratings by the assessors with respect to empty result sets. In robustness checks, we show that our results remain qualitatively similar even if we restrict attention to non-empty result sets only.

In total, we received 11,392 assessments of result sets (without duplicates and empty sets). Then we restored evaluations for duplicate result sets and imputed evaluations for zero sets per each worker, resulting in 16,176 evaluations ready for the analysis. See Table C.5 for more details about the sample size per each type of the evaluator.

Table C.6 provides summary statistics for the evaluations using only unique result sets (without duplicates or empty result sets). We split the answers by the type of the assessor and also by language of the query to facilitate comparison. Overall, the distribution of ratings seem to be broadly in line with each other by different assessors, although clearly there are certain idiosyncrasies. In robustness checks, we discuss relative merits of answers by research assistants relative to answers by clickworkers and show that our result remain qualitatively the same independent of which type of assessors we use.

Table C.6: Summary statistics for ratings

evaluator	lang	median rating	mean rating	shr of 7	shr of 1	shr of no best URL	n obs
clickworker	de	6	5.42	0.33	0.04	0.06	4,632
DE RA	de	6	5.21	0.40	0.11	0.11	1,544
clickworker	en	6	5.35	0.29	0.04	0.05	3,912
EN RA	en	7	6.02	0.53	0.03	0.06	1,304

Notes: This table shows summary statistics on ratings by type of assessor and language. The column headers use the following abbreviations: lang: language of the query, shr: share, shr of 7: share of "completely satisfied" ratings, shr of 1: share of "completely dissatisfied, as if no results at all" ratings, no best URL: the evaluator decided that no result in the result list is satisfactory, n obs: total number of evaluations.

D Instructions for human assessors

D.1 Instructions to research assistants

The text below contains the instructions that were given to the research assistants (university students):

Here are the detailed instructions for your RA-task:

1. You are asked to evaluate 1,544 result sets provided by a search engine.

2. Please, click on the following link: https://madinak.shinyapps.io/assessment_app_mag/

3. You will see a field which asks you to put the result list with which you want to start your evaluation. Choose result list #1 and proceed in chronological order. You would see a webpage like in an example below: [Figure D.1 was shown here]

4. Each result set consists of a search query term (highlighted with red rectangle in the picture above) and up to five results (blue rectangle), where each result usually includes a URL link to a website and a short description of that website. For example, the picture above represents results provided by a search engine to someone who was searching for "ptgui". The search engine provided five results. The first result, for example, is a company page https://www.ptgui.com/. The fifth and last result is a webpage which allows to download ptgui software within https://www.giga.de.

5. You are asked to do three things for each result set:

   (a) Evaluate how satisfied you are with the results for a given query overall on a scale from 1 to 7 from a drop down menu (see light green rectangle), where 1 would be equivalent to a situation when the search engine does not give you any results and 7 means that you are extremely satisfied with the result.

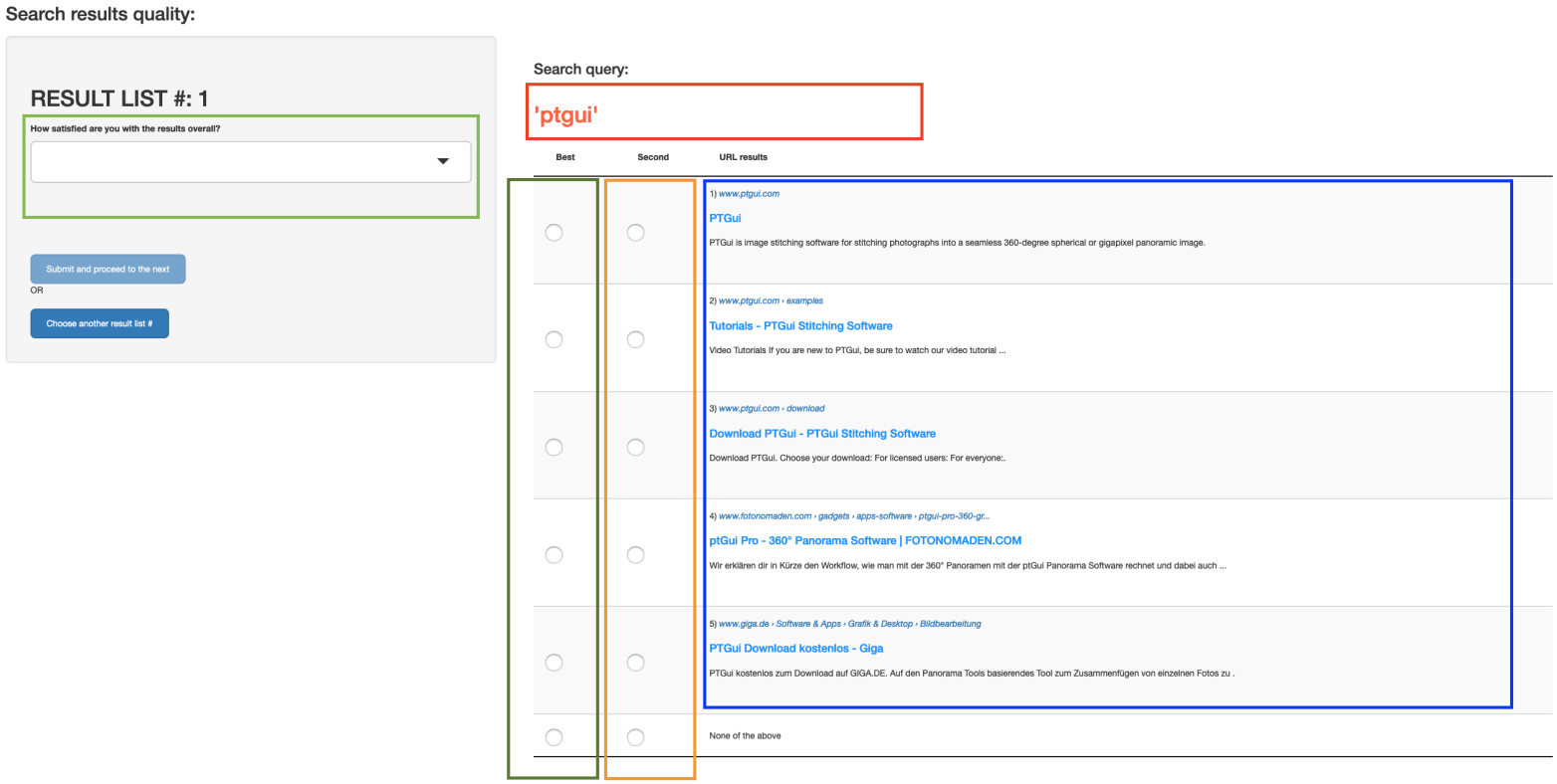


Figure D.1: Example of a web page with search results used for human assessment

Please, evaluate the quality of the result set as if you are really searching for the answer. For example, as a search engine user you want the relevant information to appear first in the search results, and less relevant — later. So, please take the order of the results into account when evaluating overall quality.

(b) Among the results provided by the search engine, please choose the result that answers the query the best. You need to click on the radio button in the first column of radio buttons (see the dark green rectangle) in the row that corresponds to the result you have chosen as the best. If you think that none of the results provided by the search engine answer the query well, you can always choose "None of the above" by clicking the radio button in the last row.

Please, click on the URL links, if brief descriptions are not enough to give you an idea about each website. Of course, sometimes it is clear without clicking, but sometimes it is not.

(c) You also need to choose the second-best result, by clicking on the radio button in the second column of radio buttons (see the orange rectangle) in the row that corresponds to the result you have chosen as second-best. You can also choose "None of the above".

6. Note that you cannot simultaneously choose the same result as best and second-best (except for, of course, the "None of the above" option).

7. After you have selected the overall rating of the results, the best, and the second-best result, you should push the submit button which will automatically load the next result list in chronological order. If you made a mistake and you want to return back to some of the result sets you have already evaluated, you can always manually choose the result set number by clicking "Choose another result list #".

8. Note that you may see that some queries may be repeating, but result sets are different. This is on purpose.

9. Also, sometimes there will be fewer than five results in a result set.

10. If there is only one result in the result set, please, choose “None of the above” as best and second-best result.

11. I expect that on average you will spend around 1 minute on evaluating one result set. Of course, there will be queries which will be harder to understand, for which you will have to click on every link and explore the results better. As for example, the example result set’s query on “ptgui”. If you have never heard about such software, you would need more time to click on the URL links in the result set, to get acquainted with the concept. However, there will be queries which are straightforward for you (some common knowledge popular queries). So those will not take much time to evaluate. Moreover, as many queries will repeat from time to time, the process should go faster than at the beginning.

12. Also, sometimes some results will not have a brief description under the URL. It will say ”(Description not available)”. This may happen at random. Or this may happen because the web-page no longer exists (the result lists have been collected several months ago). Please, do not penalise such results, this is not the search engine’s fault. Rather try to infer whether it was a valid result or not. You are encouraged to click on those URLs.

13. In general, do not hesitate to click on the links if you want to understand more the context of the query and the results.

14. Note that the result sets are real results by a search engine for a random sample of queries people search on the internet. I filtered out inappropriate content, however, should you

still find any inappropriate queries and/or URL results, please skip that result list and let me know the number of the problematic result set, so I would know the reason you skipped.

15. When you want to make a break in your work, please write down the number of the last result list for which you completed the evaluation, and continue with the next one after the break.

16. You have 3 weeks to finish the evaluations, i.e., by July 15. Please let me know when you start evaluations, so we can cross-check for the first few evaluations that the app works as intended.

D.2 Instructions to clickworkers

The text below contains the instructions that were given to people hired through the *clickworker.com* platform to perform the assessment.

Please decide **how good search results match a search term**.

We will show you up to 5 results.

Important:

- If you are not sure how good a result matches the query please follow the link.
- Please keep in mind that the order of results is also relevant for the quality of results.
- If titles or snippets are missing do not evaluate the results. Only judge the results that are visible.

E Robustness

Human assessment of search results shows that if we reduce the amount of user data available to the search engine algorithm, users find that the quality of search results becomes worse, especially for rare queries. In this Appendix, we present additional robustness checks. First, we keep our preferred measure of quality – the average rating on the Likert scale – and show that the main result holds even if we remove empty result sets from the analysis. We also show that the result holds independent of the identity of the assessor. Finally, we show that the result holds if we use alternative measures of quality, whether coming from human assessment or through automated comparison of the overlap with Google results.

E.1 Empty result sets

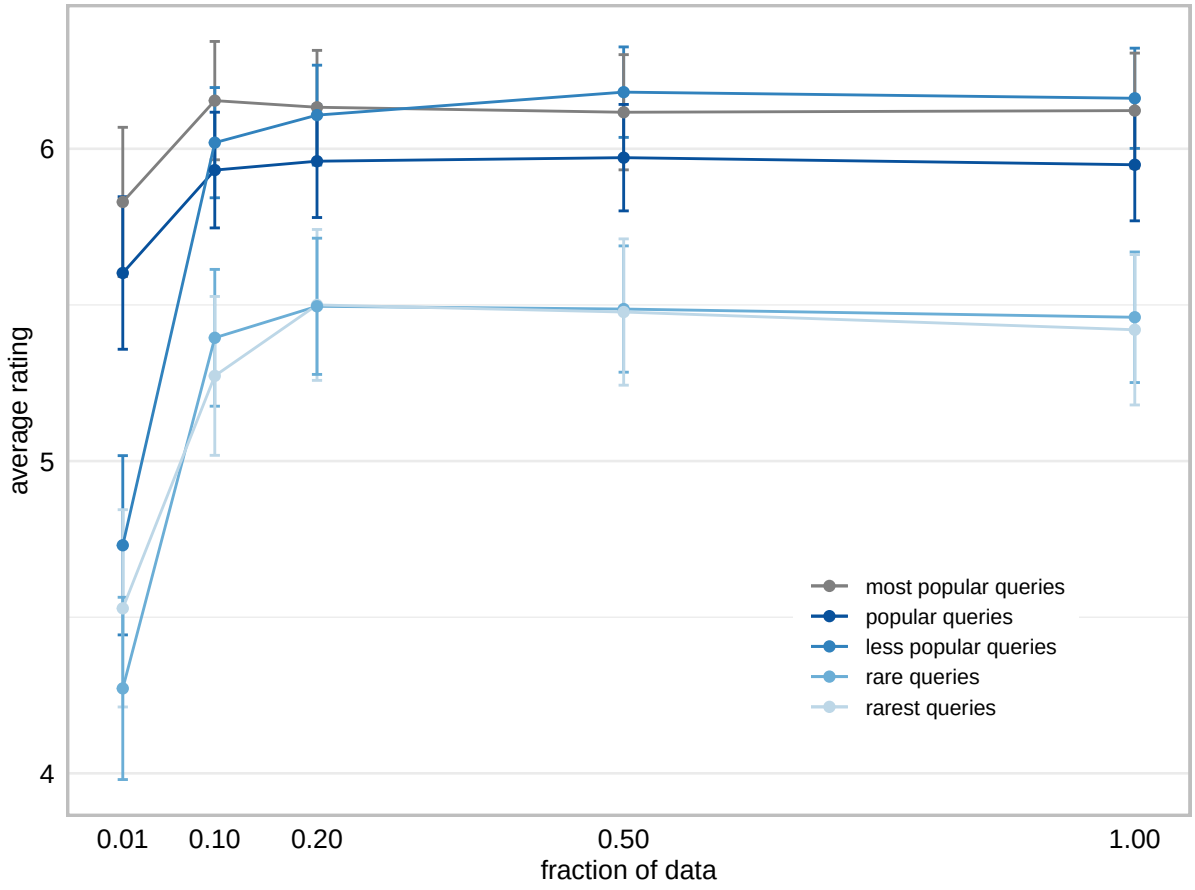
In Table C.4 of Appendix C, we showed that the incidence of empty result sets was increasing for rarer queries and for lower fractions of user data used to generate the results. In other words, as data available to Cliqz search engine became scarcer, the search engine found it harder to provide search results. At full data, the Cliqz search engine failed to return any results for 3% of queries, while at 1% of user data, it failed for half of the queries.

In our analysis, we assumed that such empty result sets should receive the lowest quality rating of 1 (and we anchored the rating scale by explicitly stating that a rating of one is "as if no results at all"). Here, provide additional results for the case in which we only use the queries that generated non-empty result sets at *all* five levels of user-data availability.

As we now drop queries without results for any level of data and as it was more likely that results were missing for low levels of data, we expect that the average rating for the remaining ratings should be higher and that this effect is particularly pronounced at low levels of data.

Figure E.2 shows the result. As compared to Figure 1 in the main text, the patterns are qualitatively similar. For each type of query, the human assessors still give lower ratings to

Figure E.2: Average ratings as function of query popularity and user-data availability: queries with no empty sets at any fraction of data



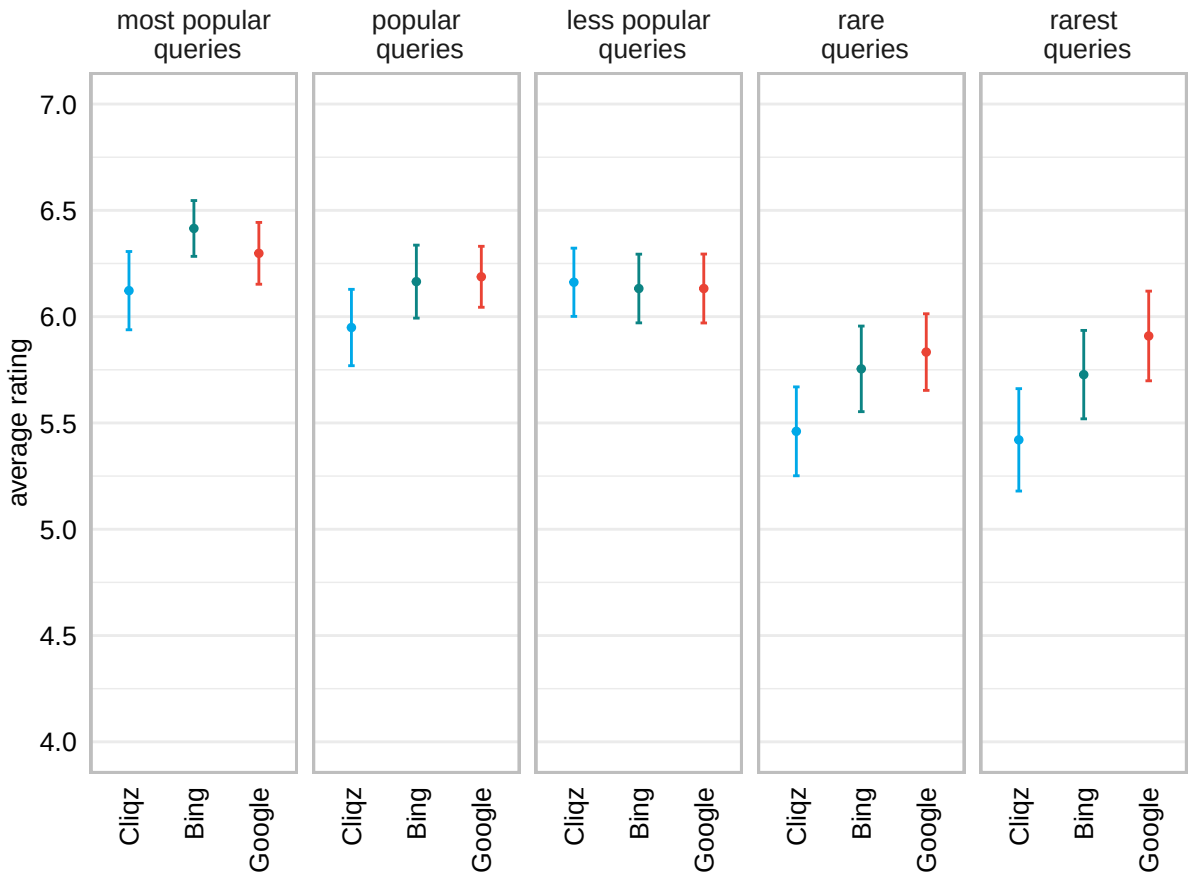
Notes: This figure is based on the sample of queries for which Cliqz was able to generate a non-empty result set at all five levels of data availability. In total, there are 4,860 assessments for 243 queries (most popular: 47, popular: 44, less popular: 51, rare: 57, rarest: 44).

result sets generated at lower fractions of data. One difference (discussed below) is that the lines are now less steep for higher fractions of data.

We also calculated the average rating for results produced by Google and Bing for the exact same selected set of queries. Figure E.3 reproduces Figure 3 from the main text for the set of queries for which Cliqz was able to produce search results for any level of data we consider.

It is interesting to directly compare results with and without nonempty results by comparing Figure E.3 to Figure 3 in the main text. Table E.7 provides numbers corresponding to the dots in Figure 3 in the main text and Figure E.7. We can see that for each search engine, average ratings are very similar (comparing entried between the last two columns) for less popular, popular, and the most popular queries. At the same time, results are different for the rarest and rare queries. Our preferred explanation for this is that the queries for which Cliqz is able to produce results at all levels of data are "easier queries" on average. This could partly be because they are more popular, which means that more data is available to produce search results (see also Table C.4.) Interestingly, we see that also for the rarest queries, Google's results are better rated when we restrict attention to queries with non-missing results (5.91 vs. 5.70). That difference is bigger for Bing (5.73 vs. 5.46) and even bigger for Cliqz (5.42 vs. 4.70). This means that for this selected set of "easier queries" the gap between Cliqz and Google becomes smaller, but is still substantial. This, in turn, can also explain why the lines in Figure E.2 are less steep than in Figure 1 in the main text.

Figure E.3: Average ratings by query popularity: no empty results



Notes: This figure is based on the sample of queries for which Cliqz was able to generate a non-empty result set at all five levels of data availability: in total, 4,860 assessments for 243 queries (by popularity: 47 queries in most popular, 44 in popular, 51 in less popular, 57 in rare, and 44 in the rarest bucket).

Table E.7: Average ratings with and without empty result sets

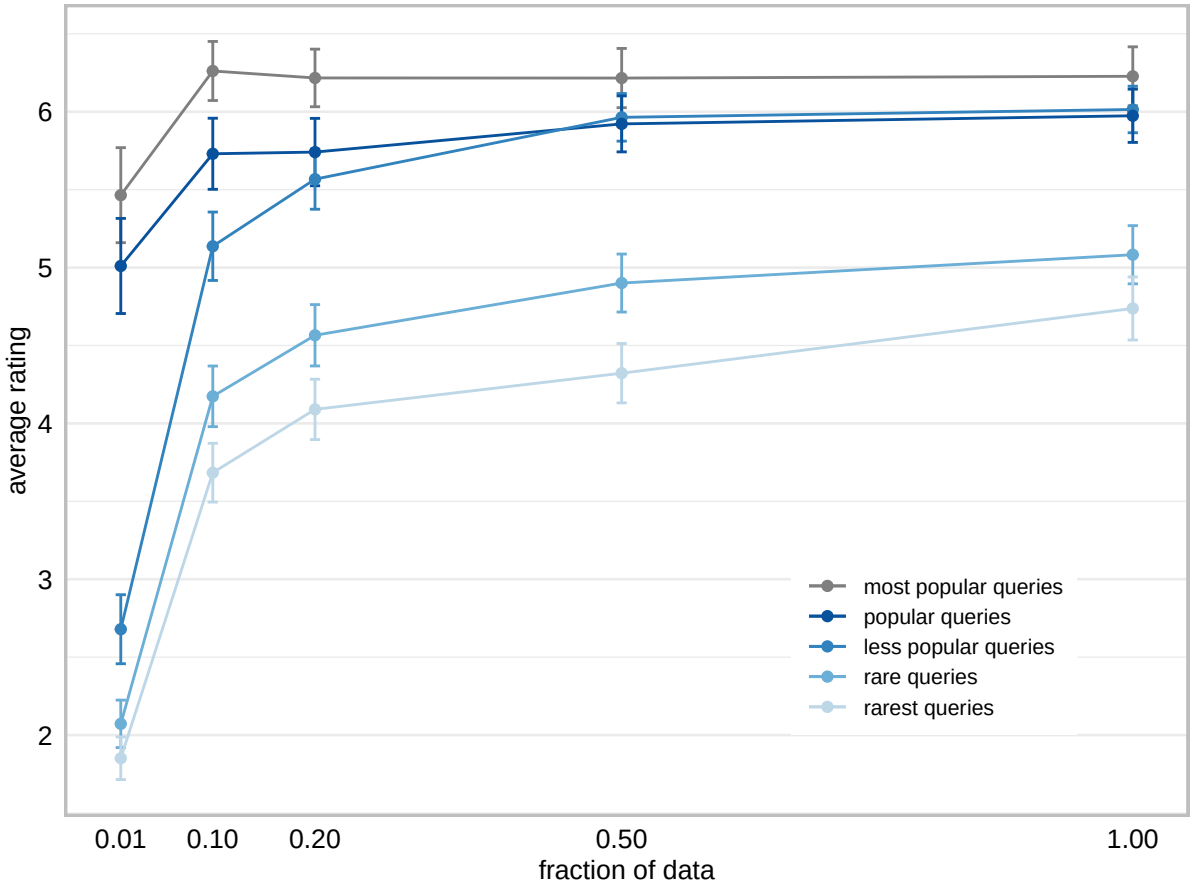
search engine	popularity	average rating (all)	average rating (nonmissing)
Cliqz	Rarest queries	4.70	5.42
Cliqz	Rare queries	4.99	5.46
Cliqz	Less popular queries	5.94	6.16
Cliqz	Popular queries	5.97	5.95
Cliqz	Most popular queries	6.13	6.12
Bing	Rarest queries	5.46	5.73
Bing	Rare queries	5.53	5.75
Bing	Less popular queries	6.08	6.13
Bing	Popular queries	6.17	6.16
Bing	Most popular queries	6.40	6.41
Google	Rarest queries	5.70	5.91
Google	Rare queries	5.77	5.83
Google	Less popular queries	6.16	6.13
Google	Popular queries	6.17	6.19
Google	Most popular queries	6.32	6.30

Notes: This table shows average ratings for all queries (third column, corresponding to the dots in Figure 3 in the main text) and average ratings for all queries for which there were nonempty results for Cliqz for all levels of data (fourth column, corresponding to the dots in Figure E.2).)

E.2 Missing snippets

By trying to construct as natural a web-browser experience for the assessors as possible, we framed each URL result with a corresponding title and a snippet as it is usually represented on the web pages of search engines. However, as we noted earlier, 168 out of 3,846 unique URLs in Cliqz results did not have a matching snippet. It meant that 1,020 result sets of Cliqz (out of 10,260) were visually distinct since some results had incomplete snippets. Figure E.4 shows that the main result remains unchanged even if we remove result sets that contained missing snippets, suggesting that our results are not driven by slightly different representation of search results across search engine sources.

Figure E.4: Average ratings as function of query popularity and user-data availability: no missing snippets

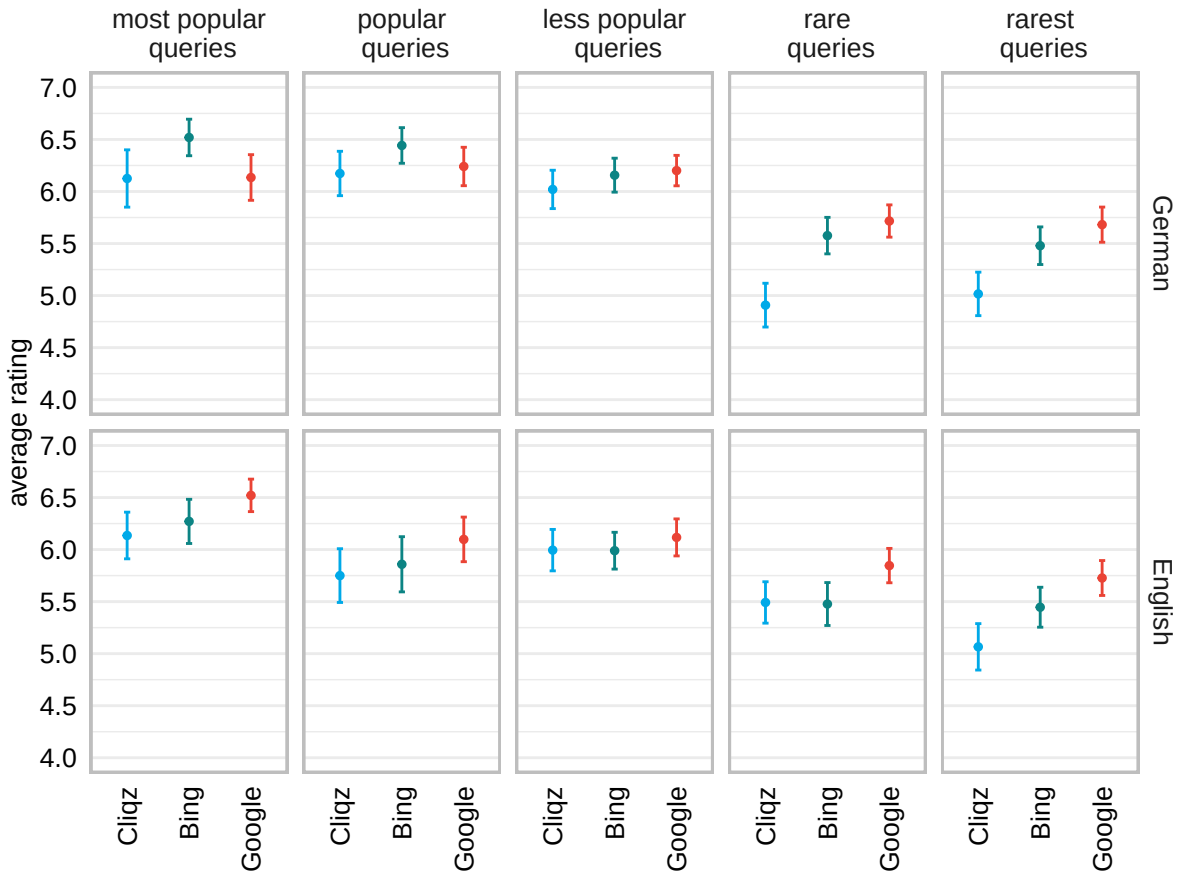


Notes: This figure is based on 9,240 assessments for Cliqz result sets for 485 queries at 5 different levels of data availability (see Appendix C for details). Result sets with no missing snippets.

E.3 Language of queries

In our analysis, we pool results across language. Figure E.5 reproduces Figure 3 from the main text by language of the query. The patterns are generally similar. If anything, it seems to be the case that the gap between Google's and Cliqz' results is wider for the rare and rarest queries in German, as compared to English. So, Cliqz does not seem to have a home advantage in the sense that it is able to produce better results in German, as compared to English.

Figure E.5: Average ratings for queries in German versus queries in English



Notes: This figure reproduces Figure 3 from the main text by language of the query. This figure is based on 10,260 assessments for Cliqz result sets for 493 queries at 5 different levels of data availability (see Appendix C for details).

E.4 Differences across types of assessors

The main analysis pools answers from research assistants and clickworkers. Here, we discuss the implications of this and show that this does not affect our conclusions.

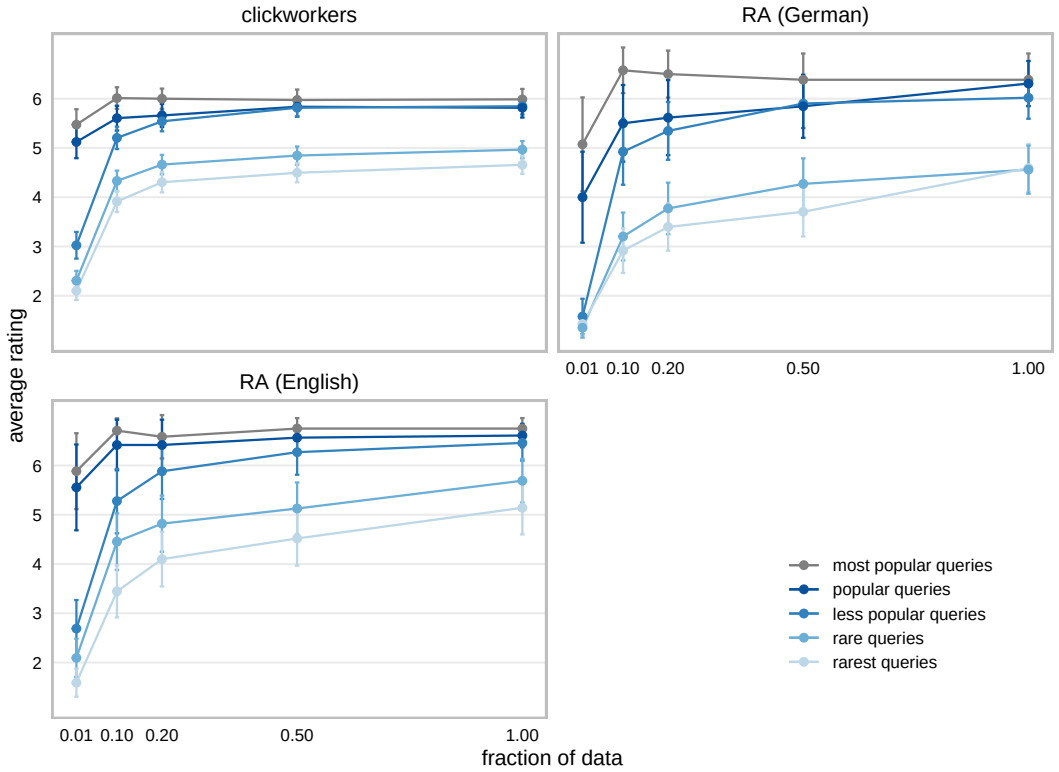
For two main reasons, we expect the assessments by the two research assistants to be more consistent. First, they evaluated more than 1,000 result sets, which provided them with the opportunity to learn what good result sets look like. Second, they first evaluated all the mixed sets (in random order), and only afterwards they were given all the original result sets (also in random order); mixed result sets were more likely to be of good quality.

A clickworker, on the other hand, did not see enough result sets to develop more experience in the given task, so the scope for their learning was limited. Thus, the ratings by each individual clickworker is expected to be noisier than the ratings by the research assistants. We believe that it is valuable to use the ratings that were provided by the clickworkers, as they represent a broader population, with the age ranging from 18 to up to 90 years old and a median age of 34 years. Moreover, the sheer number of evaluations is helpful to reduce noise. Therefore, the ratings by clickworkers might be more representative of a general German population than the ratings by the research assistants.

Finally, since the order of result sets were completely random, in expectation, learning or the absence of it should have impacted all buckets and all search engines results equally. The noise should make it difficult to find any difference at all. If despite all the noise, we still observe that human assessors rate certain types of result sets systematically higher than the other, it must be due to the fact that they are of higher quality.

To assess this, we show the results by type of assessor. Figure E.6 shows that qualitatively the results do not change if we use answers of one group of assessors or the other. In other words, the main result holds independent of the type of the assessor.

Figure E.6: Average ratings as function of query popularity and user-data availability: by type of assessor



Notes: This figure shows the average ratings of Cliqz result sets separately for each type of human assessors. This figure is based on 10,260 assessments for Cliqz result sets for 493 queries at 5 different levels of data availability (see Appendix C for details).

We also conducted a regression analysis in order to control for assessor fixed effects and thus take into account only variation of ratings within the answers of any given assessor. We also account for query fixed effects.

We use the answers on the Likert scale as the dependent variable and estimate the changes in user satisfaction for 24 groups of result sets (5 buckets at 5 different fractions minus one baseline group, which is the group of most popular queries at full data). We fit the linear model

$$y_{iqf} = \alpha_i + \sum_{b=1}^{5} \sum_{f=1}^{5} \beta_{b,f} I\{q \in b, f\} + \delta_q + \varepsilon_{iqf}, \tag{E.1}$$

where y_{iqf} is the rating assessor i ($i \in \{1, \ldots, 565\}$, i.e., 563 clickworkers plus two research assistants) provided for query q ($q \in \{1, \ldots, 493\}$) when fraction f of the data was used. α_i is an assessor fixed effect. $\beta_{b,f}$ are bucket-specific effects of the fractions of data used. Technically, each query q is in bucket b; we use this to construct indicators $I\{q \in b, f\}$ for bucket-fraction combinations that we use as regressors. δ_q is a query fixed effect and ε_{iqf} is the error term. We normalize $\beta_{b,f}$ to be zero for the most popular queries at full data. Given this, the parameters $\beta_{b,f}$ are the difference in the ratings between the group of result sets in bucket b at fraction f and the baseline group of result sets (most popular queries at full data). Reported standard errors are clustered at the assessor level.

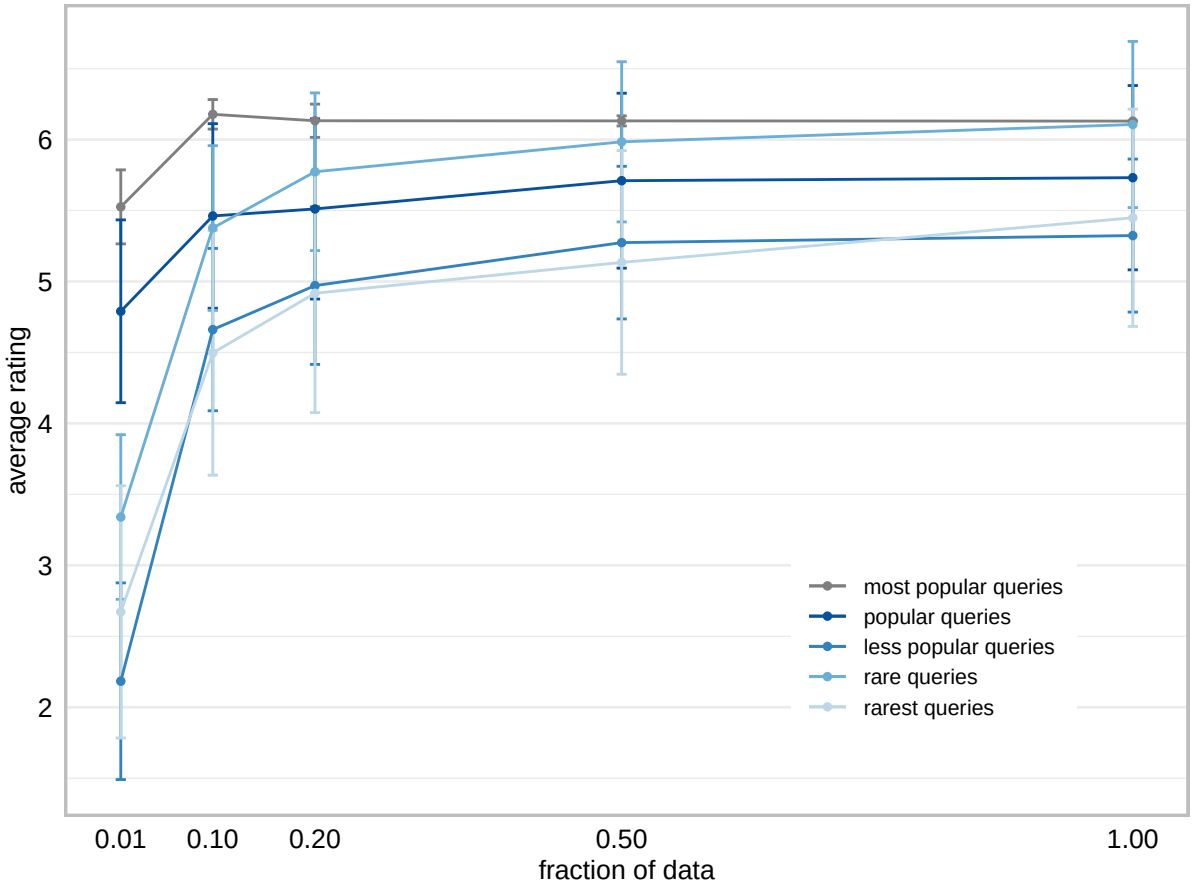
Table E.8 provides the results. For instance, the coefficient estimate -0.52 for "popular; fraction = 0.2" means that the average rating was 0.52 less for popular queries with 20 percent of the data, as compared to the average rating for the most popular queries under full data. Based on the regression results, Figure E.7 again plots the predicted average ratings and shows that the lines have similar shape to the ones in Figure 1 in the main text.

Table E.8: Average ratings as function of query popularity and user-data availability

bucket × fraction	est.	s.e.	t-stat	p-val
most popular; fraction = 0.5	-0.00	0.02	-0.01	0.99
most popular; fraction = 0.2	-0.00	0.06	-0.08	0.94
most popular; fraction = 0.1	0.05	0.05	0.97	0.33
most popular; fraction = 0.01	-0.61	0.14	-4.22	0.00
popular; fraction = 1.0	-0.31	0.25	-1.23	0.22
popular; fraction = 0.5	-0.34	0.23	-1.44	0.15
popular; fraction = 0.2	-0.52	0.24	-2.14	0.03
popular; fraction = 0.1	-0.58	0.24	-2.38	0.02
popular; fraction = 0.01	-1.28	0.28	-4.59	0.00
less popular; fraction = 1.0	-0.93	0.30	-3.09	0.00
less popular; fraction = 0.5	-0.98	0.29	-3.33	0.00
less popular; fraction = 0.2	-1.26	0.30	-4.16	0.00
less popular; fraction = 0.1	-1.59	0.30	-5.32	0.00
less popular; fraction = 0.01	-4.07	0.39	-10.52	0.00
rare; fraction = 1.0	-0.02	0.26	-0.08	0.93
rare; fraction = 0.5	-0.15	0.25	-0.62	0.53
rare; fraction = 0.2	-0.37	0.25	-1.49	0.14
rare; fraction = 0.1	-0.75	0.26	-2.87	0.00
rare; fraction = 0.01	-2.80	0.25	-11.01	0.00
rarest; fraction = 1.0	-0.83	0.44	-1.87	0.06
rarest; fraction = 0.5	-1.14	0.45	-2.53	0.01
rarest; fraction = 0.2	-1.35	0.48	-2.83	0.00
rarest; fraction = 0.1	-1.76	0.51	-3.47	0.00
rarest; fraction = 0.01	-3.62	0.52	-6.96	0.00

Notes: This table reports results from a regression of ratings on bucket times fraction of available data indicators. Based on 10,260 assessments for Cliqz result sets for 493 queries at different levels of data availability (see Appendix C for details). Standard errors are clustered at the assessor level.

Figure E.7: Regression results: Average ratings as function of query popularity and user-data availability



Notes: This figure shows the predicted ratings for Cliqz result sets at different fractions of data. Based on estimating model (E.1), which controls for across-assessor and across-query variation in ratings using fixed effects.

E.5 Alternative measures of quality

Our main result, as depicted in Figure 1 in the main text, is based on the average ratings for Cliqz result sets grouped by the query's popularity (i.e., the search frequency buckets) at different levels of user-data availability (i.e., at different fractions of query logs). One may be concerned that a Likert scale is a categorical variable and not a cardinal one and that our results are solely based on human assessment. Here, we show that our results are robust to using three alternative measures of quality, including one that is not based on human assessment.

The first measure is the *share of mostly or completely satisfied ratings*, i.e., the share of results sets rated 6 at least on the Likert scale. The advantage of using this measure is that we do not have to impose cardinality. Figure E.8 shows the result. They are qualitatively the same.

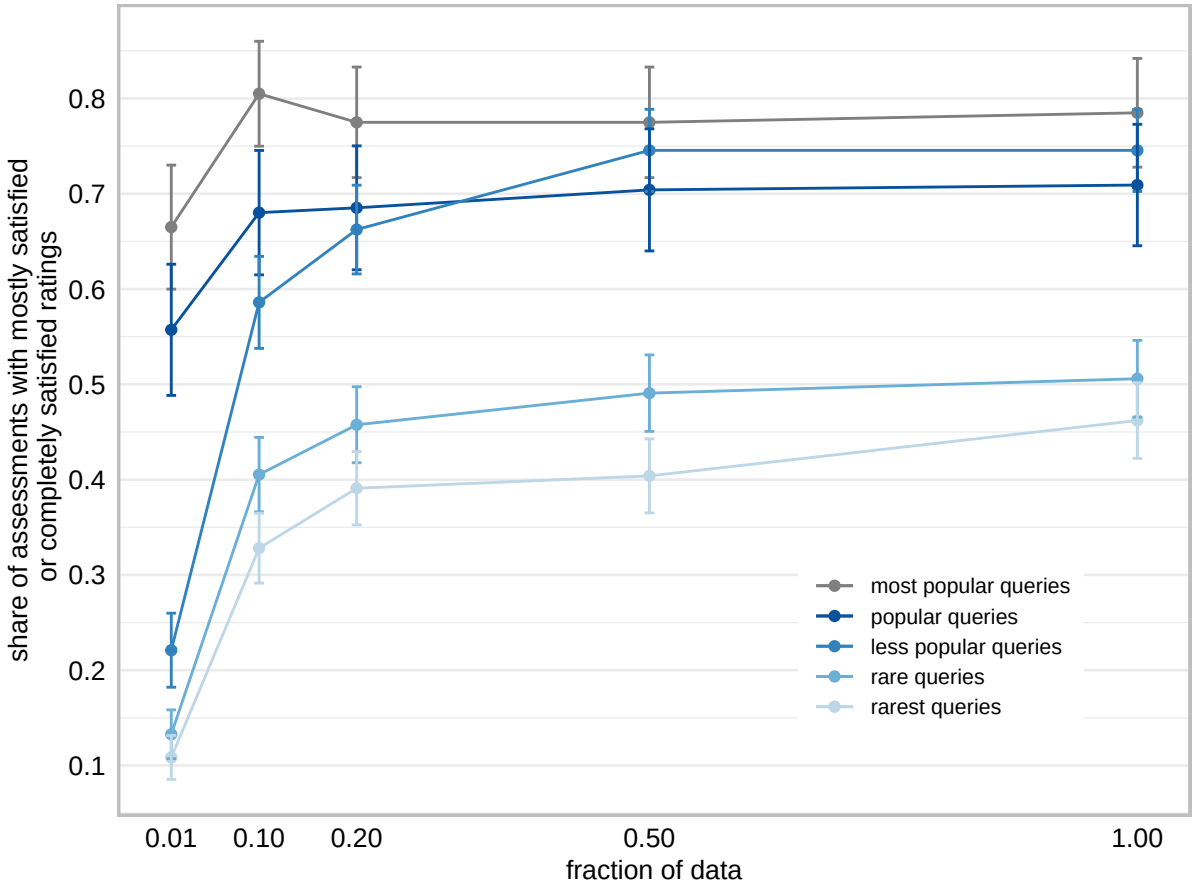
The second measure is the *position of the best rated result from the mixed result sets*. For all three search engines and all levels of available data (for Cliqz) we determine whether the best rated result for each of the 493 queries is presented as the top result, or in position 2 or 3, 4 to 10, 11 to 18, or not in the top 18. Again, for this, we do not treat the ratings as cardinal.

Figure E.9 shows the result. It confirms that the quality of search results depends on the amount of data that is used to obtain them (for Cliqz). By this measure, overall Google produces the best search results, closely followed by Bing, ahead of Cliqz.

Taken together, these two robustness checks suggests that assuming cardinality and looking at average ratings is appropriate for our purposes.

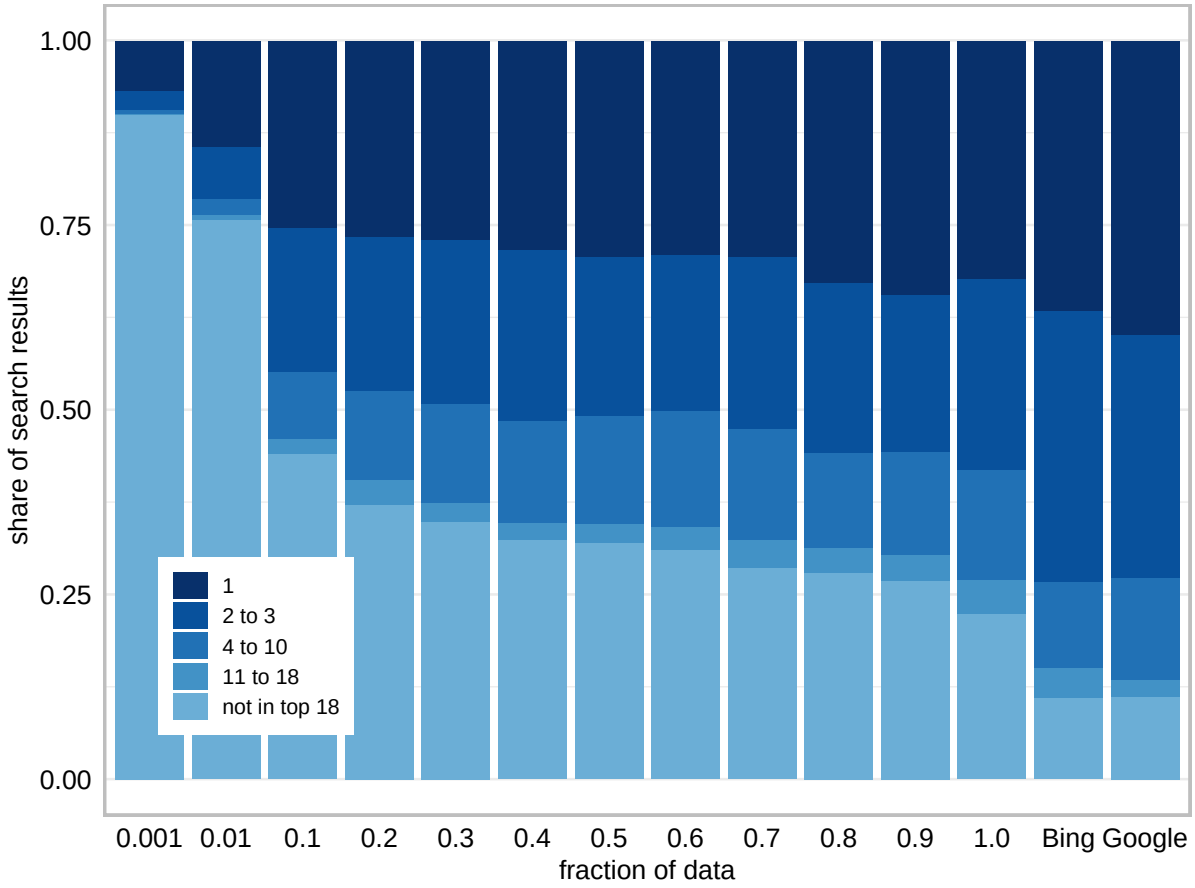
As a third alternative measure, we step away from human assessment and use the *Google results as the yardstick*. Specifically, our third alternative measure is to what extent Cliqz produces the same top, top 3 and top 5 results as Google. Under this measures, the top x results are considered to be the same, when the respective elements are the same. The ordering is not taken into account. Figure E.9 shows for all 3 versions of this alternative measure that we obtain similar results as the ones in Figure 1 in the main text.

Figure E.8: Share mostly or completely satisfied assessments



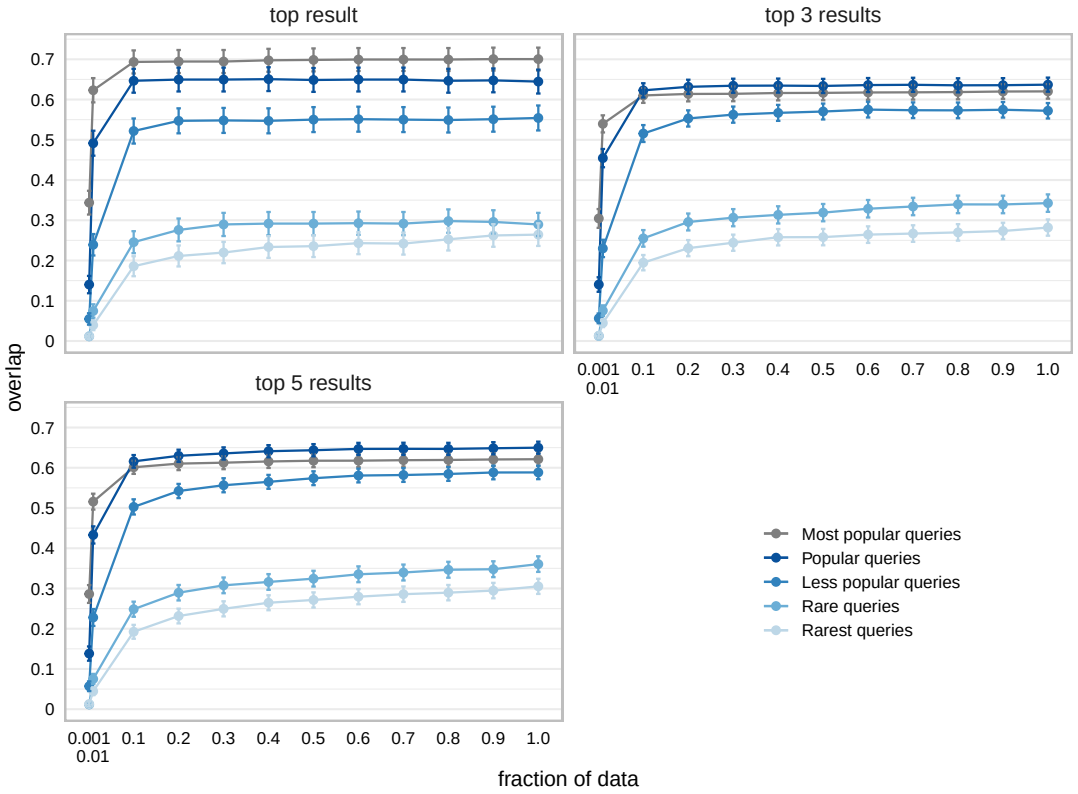
Notes: This figure shows the share of assessments that were mostly satisfied (rating of 6 on the Likert scale) or completely satisfied (rating of 7), by popularity of the query (bucket) and fraction of data that was used to produce the search results. Based on 10,260 assessments for a random sample of 493 queries and the corresponding 2,465 Cliqz result sets (see Appendix C for details).

Figure E.9: Position of best rated result



Notes: This figure shows how the position of the overall best result differs across search engines and depends on the amount of data that was used to obtain the search results for Cliqz. The overall best result was determined using the 493 mixed result sets for the 493 sampled queries (see Appendix C).

Figure E.10: Overlap with Google results



Notes: Fraction of Cliqz results for 493 queries at 12 different levels of data availability that are the same as Google results. Here, "same" means (a) the same top result, (b) the same top 3 results, (c) the same top 5 results. (b) and (c) are in the sense of an unordered set comparison, meaning that the set of results is the same and that the ordering does not matter.

References

1. D. He, *et al.*, *International Conference on Web and Internet Economics* (Springer, 2017), pp. 294–310.

2. M. Schäfer, G. Sapi, Learning from data and network effects: The example of internet search, DIW Working Paper, DIW Berlin, Germany (2020).

3. C. Silverstein, H. Marais, M. Henzinger, M. Moricz, *Acm sigir forum* **33**, 6 (1999).

4. Z. Guan, E. Cutrell, *Proceedings of the SIGCHI conference on Human factors in computing systems* (2007), pp. 417–420.