Schaefer, Maximilian; Sapi, Geza; Lorincz, Szabolcs

Working Paper

The effect of big data on recommendation quality: The example of internet search

DIW Discussion Papers, No. 1730

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: Schaefer, Maximilian; Sapi, Geza; Lorincz, Szabolcs (2018) : The effect of big data on recommendation quality: The example of internet search, DIW Discussion Papers, No. 1730, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/178209

DIW BERLIN

1730

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2018

The Effect of Big Data on Recommendation Quality

The Example of Internet Search

Maximilian Schaefer, Geza Sapi and Szabolcs Lorincz

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

The Effect of Big Data on Recommendation Quality: The Example of Internet Search

Maximilian Schaefer* Geza Sapi† Szabolcs Lorincz‡

April 3, 2018

Abstract

Are there economies of scale to data in internet search? This paper is first to use real search engine query logs to empirically investigate how data drives the quality of internet search results. We find evidence that the quality of search results improve with more data on previous searches. Moreover, our results indicate that the type of data matters as well: personalized information is particularly valuable as it massively increases the speed of learning. We also provide some evidence that factors not directly related to data such as the general quality of the applied algorithms play an important role. The suggested methods to disentangle the effect of data from other factors driving the quality of search results can be applied to assess the returns to data in various recommendation systems in e-commerce, including product and information search. We also discuss the managerial, privacy, and competition policy implications of our findings.

Keywords: Big Data, Recommendation quality, Internet search, E-Commerce, Economies of Scale, Search engines

JEL Codes: C55, L81, L86, M15

*German Institute for Economic Research (DIW Berlin), 10117 Berlin, Germany, m.schaefer@diw.de

†European Commission DG COMP - Chief Economist Team and Düsseldorf Institute for Competition Economics, Heinrich-Heine-Universität Düsseldorf, 40204 Düsseldorf, Germany, sapi@dice.uni-duesseldorf.de

The views expressed in this article are solely those of the authors and may not, under any circumstances, be regarded as representing an official position of the European Commission. This is personal research based entirely on publicly available information and not related to any activity of the European Commission.

‡European Commission DG COMP - Chief Economist Team, 1210 Sint-Joost-ten-Noode, Belgium, szabolcs.lorincz@ec.europa.eu.

1 Introduction

The rise of the internet enables consumers to browse through and compare a broad range of products before they ultimately choose the one that is closest to their expectations. This also poses new challenges for firms in electronic commerce that seek to proactively engage with consumers and to suggest the most relevant product to facilitate purchasing. Firms like Google, Netfix, and Amazon are well known for their recommendation systems for product search and allegedly generate a substantial share of sales via these recommendations. In the words of Steve Jobs, "A lot of times, people dont know what they want until you show it to them" (Arora, 2016).

These recommendation systems consist of two main components. First, they rely on big data, to a large extent generated by customers in their previous purchases and product or information searches. Second, they make use of algorithms, such as those used in machine learning, to predict, from the underlying data, the most relevant products for the customers' Search. General purpose search engines like Bing, Google, and Yahoo! are perhaps the prime examples of such recommender systems. These service providers face the task of organizing and indexing the virtually infinite information available in the World Wide Web, then rank this content in order to provide the most relevant search results to each individual user.

In perhaps no other market has the question of the role of data stirred such a vivid discussion among industry participants, academic experts, and policy advocates than in general internet search. A wide spectrum of opinions about the potential impact of accumulated search data on search result quality. At one end of the spectrum, it is argued that data is a minor ingredient in the recipe to provide relevant search results. Other factors that are not related to the sheer amount of data are the key inputs of search result quality, such as the search engines ability to crawl and index the documents published in the internet, to store and organize the retrieved catalogue, as well as the algorithm used to retrieve the most relevant indexed documents to each search query (Varian, 2016). Commentators at opposing ends of the opinion scale argue that data on previous user search behavior is a crucial determinant of the quality of search results and search engines learn from more data (McAfee et al., 2015). This line of argument went on to highlight the role of data in strengthening Google's market position, alleging that "Google's overwhelming strength comes from its ownership of vast datasets" (Guardian, 2015).

Despite the prominence of data-driven businesses and the surrounding heated policy discussion, surprisingly little is known about the precise role of data in improving the quality of recommendation systems, such as internet search. How much data is needed for optimal quality? And what type of data? To our knowledge, this article is the first to address these important questions using real data on user behavior obtained from a general purpose internet search engine. We study the impact of accumulated search history data on the quality of internet search results using Yahoo! data. The main aim of this paper is to propose to managers and policy makers such as competition authorities - an empirical strategy that allows isolating the impact of data on service quality from other factors. The methods we bring forward are simple and straightforward to reproduce in contexts other than general internet search. The data requirements are extremely low, as user activity logs are available in every e-commerce firm. We apply a very simple and cheap measure to quantify the quality of search results (the click-through-rate of the first displayed URL), and show that our findings are robust to the use of a costly alternative quality measure based on editorial reference judgements.

Our analysis leads to two conclusions. First, there is learning from user data. Second, personalized information is crucial. In particular, our analysis reveals that the quality of results displayed to queries for which the search engine saw more personalized information (i.e. could track the identity of searchers for a longer sequence of previous searches) improves faster than that of queries with less personalized information available. This implies that the speed of learning from previous searches is amplified by more personalized information.

2 Related Literature

Despite the public policy attention surrounding the question, economic literature to quantify the effect of previously collected data on the quality of search results is surprisingly scarce. Chiou and Tucker (2017) rely on a browsing database to analyze how storage of personalized information affects the accuracy of search results. The authors exploit a policy change in European data retention rules as identification strategy, arguing that the shortening of query log retention time had no noticeable effect on the quality of search results. Our approach is different as our data allows a much more precise measurement of search quality. Furthermore, we study the impact

of personalized information on the quality of search results from an entirely different perspective: Instead of studying the direct impact of personalized information on search result quality, we study how personalized information mediates the speed of learning.

Bajari et al. (2018) examine the impact of data accumulation on the accuracy of retail forecasts. They find that additional data on the forecasts and the subsequent realization of retail quantities improves the accuracy of retail forecasts for a particular product. Observing more products in a specific product category, however, has almost no effect on forecast accuracy. Similarly to our approach, Bajari et al. (2018) account for non-data related improvements in forecast accuracy such as improved technology by controlling for elapsed time. As opposed to our data set, the data used by the authors do not allow to study the accumulation of different amount of data over the same time span, which is a crucial step in our identification strategy. Hence, while our approaches share the same underlying idea to control for non-data related factors, our dataset allows a different implementation of the concept

Other contributions approach the topic of scale economies in data from a policy perspective: Lerner (2014), Lambrecht and Tucker (2013) and Sokol and Comerford (2017) argue that economies of scale from data collection are low (even for tail queries) and that positive feedback loops between data collection and service quality (Pasquale, 2015; Bodapati, 2008; Stucke and Grunes, 2015) should be expected to be weak. Schepp and Wambach (2015) and Sokol and Comerford (2017) submit that the value of data is often of transitory nature and relevant only over a short time period. Argenton and Prüfer (2012) provide methods for search engine providers to share search logs with each other so they are better able to tab network externalities arising from more data. Rubinfeld and Gal (2017) and Chiou and Tucker (2017) discuss the antitrust policy relevance of big data and the extent to which data may become an entry barrier that hampers competition.

Our search engine data come from a particularly interesting market where the effect of data on search result quality is widely discussed by industry, academia and policy advocates. There is a spectrum of opinions about accumulated search datas potential impact on search result quality. At one end of the spectrum, it is argued that data is a minor ingredient in the recipe to provide relevant search results. Other factors that are not related to the sheer amount of data are the key inputs of search result quality, such as the search engines ability to crawl and index the documents published in the internet, to store and organize the retrieved catalogue, as well as the algorithm used to

retrieve the most relevant indexed documents to each search query (Varian, 2016). Commentators at the opposing end argue that data on previous user search behavior is a crucial determinant of the quality of search results. In short, search engines learn from more data (McAfee et al., 2015).

Our research is also related to the broad literature on the role of big data in driving firm performance, product development, and service quality. A vast amount of literature in various disciplines addresses this topic. Excellent literature reviews are provided by Chen et al. (2012) for information systems research, Chintagunta et al. (2016) for marketing and George et al. (2014) for management. Clearly, the emergence of big data revolutionized e-commerce and raised research questions that go beyond the borders of single disciplines: McAfee et al. (2012) discuss the overarching managerial challenges of the big data revolution.

Our article is further related to the literature on online advertising (Goldfarb and Tucker, 2014), in particular to the rapidly growing strand of research on targeted offers in marketing. A large share of this literature develops methods exploiting purchase history data to explain observed marketing outcomes, such as purchase probability (Moe and Fader, 2004; Bodapati, 2008; De et al., 2010) and clickthrough rates (Ansari and Mela, 2003). Dou et al. (2007) and Yoganarasimhan (2016) focus on the value of personal information for search quality and find that personal information can significantly increase search quality, but the effect may depend on other factors, such as the type of query or the length of user history. We contribute to this strand by highlighting once more the importance of personalized information for targeting accuracy.

The role of personalized information in online markets also attracts considerable interest from a public policy perspective, both from policy advocates and from academics. In Europe, for example, the General Data Protection Regulation (GDPR, 2016) introduced broad rights for consumers to control the data firms collect on them and granted public authorities the powers to issue significant fines for breaches, reaching up to 4% of global annual company turnover. In recent years the U.S. Federal Trade Commission, the US consumer protection watchdog, brought hundreds of cases against companies violating privacy legislation (FTC, 2017).

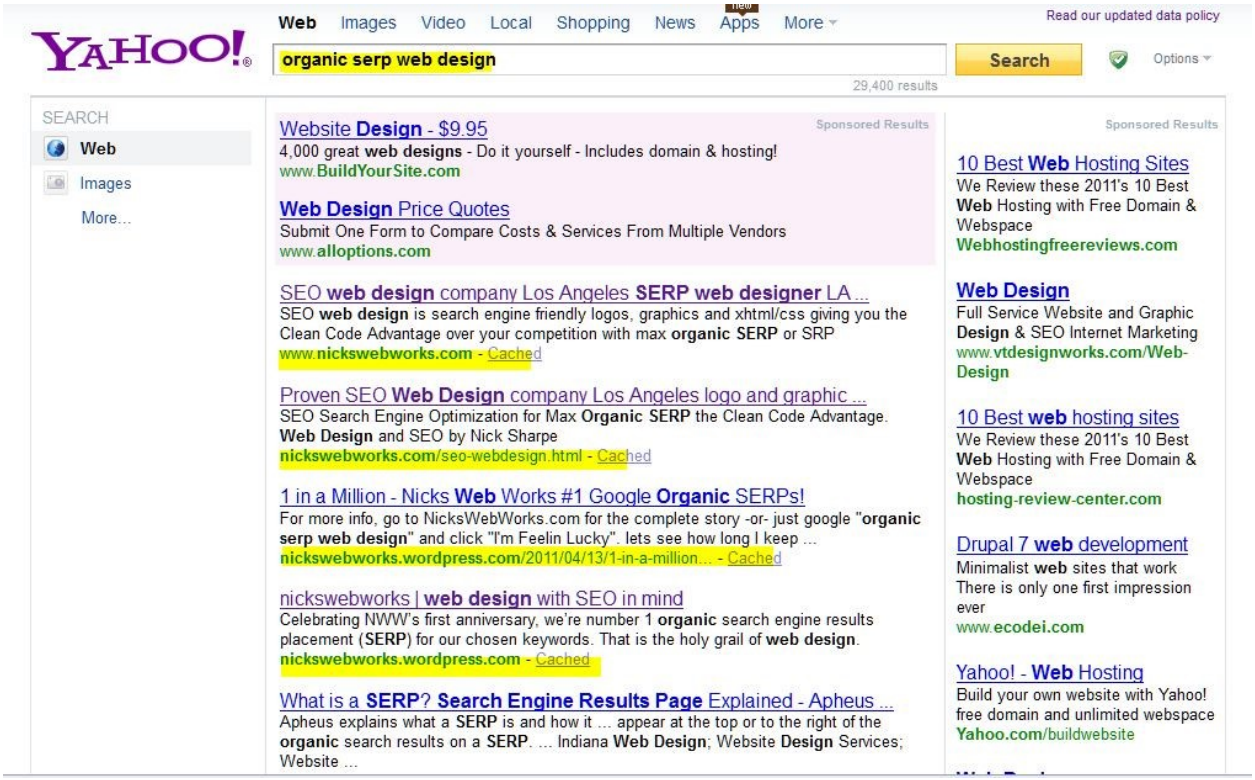


Figure 1 – Search Result Layout at Yahoo!, 2011

3 The Data

The data we use in this paper stem from Yahoo! and contain search logs spanning a period of 32 days from July 1, 2010, through August 1, 2010. A specific search log contains information about the search term, the time when the search term was submitted, the computer from which the search term originates, the returned list of results, and the clicking behavior of the user. Figure 1 illustrates the typical structure of the typical search results page (SERP) at Yahoo! at the time our dataset was gathered. The search term, which we also refer to as query in the remainder of this paper, is the sequence of characters a user types into the search bar of the search engine (in the middle in Figure 1) in her quest for information. The corresponding paid advertisements are displayed at the north and east edges of the search result lists. Our analysis focuses on organic results, which are highlighted in Figure 1, and how the quality of these non-paid result lists may improve due to more data.

The computer from which the search term originates is identified by a cookie. The returned list

of results we observe consists of the urls displayed on the first result page. The query, the cookie and the urls of the first result page were anonymized by Yahoo!, so we can only observe unique identifier codes instead of the actual values. The time the query was submitted is identified by a unix timestamp. From the data, we are able to identify the order of the displayed results, so that we know the positions of the urls on the result page.

We also observe the user's actions after submitting their search request. Most of the time, these actions simply consist of clicking on urls.[1] The search log consists of the chronologically ordered series of clicks the user performs on the urls of the first result page until she performs a terminal action that ends the search log. There are three types of terminal actions: (i) the user clicks on a url and does not return to the result page within a pre-specified time span (i.e. the url fulfilled the informational need of the user); (ii) the user submits a new search term (which is encoded as "s"); or (iii) the user performs an action encoded as "o" (i.e. "other click") in the search log.[2]

In the remainder of this paper, we use the terms *session*, *search log*, and *search* interchangeably. Our dataset consists of roughly 81 million search logs for 67652 different queries performed by approximately 29 million different cookies. The queries we observe naturally differ in the number of times they are searched. The number of observed searches by query varies from a minimum of 4 to a maximum of 10000. The observed searches are only a proxy for the true number of searches performed on a query. Unfortunately, the data description offers no guidance on the relationship between our observed measure and the true number of searches carried out on a query. The sample is, however, a random sample from the totality of searches performed on the Yahoo! search engine. This allows us to derive conclusions on the relative traffic, as the relative popularity of queries should be adequately reflected by our proxy-measure: Due to the random sampling, queries should, on average, be included in our sample proportionately to their overall popularity.

Thus, our proxy measure preserves, on average, the relative popularity of the queries. However, the sample was constructed with the restriction that one query could not be sampled more than 10000 times. This implies that the average relative popularity between queries can only be precisely

[1] After clicking on a given url, the user can come back to the result page. If she does so, we also observe the subsequent clicks she performs after returning to the result page.

[2] The data description does not clearly specify the nature of the actions encoded as "o". If they appear in the search log, actions encoded as "o" are always the last recorded action, this is why we postulate their terminal nature. It is natural to assume that "o" summarizes behavior like closing the browser, long idle time after returning to the result page, or any other action that signalizes the search engine that it is very likely that the user stopped his search.

estimated when comparing queries with strictly less than 10000 searches. For instance, queries for which we observe 8000 sessions should have been searched, on average, 4 times more often than queries for which we observe 2000 sessions.[3] On the other hand, comparing queries we observe 10000 times with queries we observe 2000 times does not allow us to draw conclusions on the impact of a five-fold increase of popularity on quality. Rather, it only allows conclusions on the impact of, *at least*, a five-fold increase of popularity. We will point to this restriction whenever necessary; for now the reader should keep in mind that the dataset does not allow us to analyze quality changes based on the true absolute number of searches. Rather, we will have to rely on a relative measure of popularity for our analysis.

In the remainder of this paper we will refer to our proxy measure as number of searches/ sessions/ search logs; thus, we do not always emphasize the proxy nature of our variable.

3.1 Measurement of search result quality

The quality measure we use in this paper is based on the recorded clicking behavior of the user after submitting a query. Each query returns an ordered list of urls on the first result page of the search engine. This returned list of ordered urls can be considered as the output of the search engine's technology. Our aim is to analyze how the quality of this output evolves with the number of searches performed on a given query. For each query - url combination, we calculate the click through rate (CTR) on the top displayed url, which we define as:

$$CTR_{jl} = \frac{\sum_{s \in jl} \mathbb{1}\{clickpos = 1\}}{\sum_{s \in jl} \mathbb{1}\{clickpos \neq 0 \cup clickpos \in \emptyset\}} \tag{1}$$

Where j denotes the query, l the ordered url combination, s the session, and *clickpos* the position of the *last* click in a session. For all the sessions associated with a given query-url combination, we count the number of times the last click was performed on the url that appears on the top of the result page ($clickpos = 1$), and divide this statistic by the total number of sessions where either (i) the last click was not performed on the url in position 0; or (ii) there was no click ($clickpos \in \emptyset$). A user action encoded as click on position 0 is a click that was performed above the first displayed url, clicks above the first url are clicks on a spelling suggestion, clicks on an alternative proposed

[3] It should be clear that our proxy measure is subject to standard sampling error; we suppress this discussion here for ease of exposition.

formulation of the query, or clicks on an advertisement. The distribution of the last click position is displayed in Figures 8 and 9 of Appendix A. [4]

We decided to exclude clicks on position 0 because advertisements cannot be considered as "generic" search results. Advertisements are positioned as a result of processes that go beyond the objective of fulfilling the informational need of the user. Stated very roughly, their placement can be considered as the result of a bargaining process added on top of the algorithm of the search engine. As we are only interested in analyzing the performance of the algorithm in fulfilling the informational need of the customer, we consider it reasonable to exclude clicks on position 0.

Our quality measure can be seen as measuring the extent to which the search algorithm achieves what one might call "the ultimate goal of a search engine," namely to place the most relevant content for a search query at the top of the displayed result list.[5] The idea behind this quality measure is straightforward: A user having to click through several urls before finding the content that matches their informational need experiences a worse search quality than a user who immediately finds their desired content. Because users naturally inspect the proposed results from top to bottom, the search experience should be best when the most relevant content is placed on top of the result list.

Users might, however, differ in their search perseverance: Some users might visit different urls before returning to the first url for reasons which are unrelated to the quality of the first url. Some users might have a specific taste for variety or simply a different time constraint than other users confronted with the same ordered result list. Nonetheless, if they return to the first url, the search engine can be assumed to have fulfilled its goal; it is for this reason that we consider the last performed click of a session in our quality measure.

[4]The data do not allow for identify the exact nature of clicks on position 0. However, because advertisements are usually placed above the first url, it is reasonable to assume that a large share of the clicks on position 0 are likely to be clicks on advertisements. This presumption is particularly reasonable for so-called commercial queries, i.e. queries that are submitted with the likely intent to buy a product. As the queries are anonymized, we are not able to identify commercial queries (like, for instance, "Play Station 3"). Rather than relying on an ad-hoc heuristic to single out likely commercial queries (like defining a threshold of clicks on position 0 above which a query is likely to be commercial), we found it more reasonable to ignore clicks on position 0. The reason for this being that queries that are classified as commercial by the search engine might have been submitted for non-commercial reason by the user. In this case, the user should be interested in the "generic" content. By filtering out commercial queries by an ad-hoc heuristic, we believe that we would unnecessarily lose information relating to the performance of the generic algorithm.

[5]Displaying the most relevant search result at the top of the result page could be seen as an intermediate step in the process of implementing search engine providers' utopia: Delivering one single personalized search result for each query. In an interview with Charlie Rose in 2005, Eric Schmidt, Google CEO from 2011 to 2015, stated the following: "When you use Google, do you get more than one answer? [rhetorical question] Of course you do. Well, that's a bug. We should be able to give you the right answer just once. We should know what you meant. You should look for information. We should get it exactly right." (see Ferenstein, 2013)

A unique feature in our dataset allows us to test the robustness of our results using an alternative quality measure based on editorial judgements of query-url pairs for a subset of searches. An explanation of the measure and the corresponding robustness checks are provided in an Appendix C. The main limitation of this measure is that it rapidly reduces the number of observations; we can only construct it for about one-third of the observed searches. The main results of our analysis remain robust to the choice of the quality measure.

It is needless to point out that our quality measure is only a proxy for the ability of the search engine to achieve its "ultimate goal." For instance, we implicitly assume that a user who does not return to the result page after clicking a given url was satisfied with this url. Furthermore, some of the clicks on position 0, which we ignore, might be indicative of a high performing algorithm, such as a correct spelling suggestion or a better alternative formulation.[6]

For the remainder of the analysis, it is important to understand that our quality measure varies for the same query as the displayed content on the first result page is either permutated or a new content appears on the first result page. The ordered list of displayed results for a given query in the sample changes for several reasons: Most prominently, the search engine might display different result lists for the same query depending on the personalized information it has on the user.

Some query-url combinations are only rarely observed and might, therefore, provide unreliable estimates of the quality of the result page. To reduce the noise, we base our analysis on rolling window averages over 100 consecutive searches with a step size of 10 calculated for each query. To be more precise, for each query, starting with the first session, we define a window of length 100 (i.e. spanning the first 100 sessions) and calculate the simple average over the values of CTR_{jl} for these 100 sessions. After completion of this step we move the window by a step size of 10 and compute the average over the values of CTR_{jl} for the next 100 sessions (i.e. the window spanning $s \in [11, 110]$).[7] We repeat this procedure until there are fewer than 100 sessions remaining.

[6] While the first point is mitigated by the fact that our measure is averaged over several searches, the latter cannot be corrected due to lack of information on the precise nature of clicks on position 0.

[7] The choice of the window length reflects a trade off between granularity and reliability of the quality measure. A large window length would not allow properly analyzing quality changes for queries that are rarely searched, while a small window length in-sample results in a too noisy measure. The step size of 10 was mainly chosen to reduce the computing time.

4 Descriptives

This section provides some important insights into the data set, which helps motivate the approach we choose to study the impact of data on the quality of search results. We start by calculating the average quality as a function of observed searches in the sample period.[8]

$$\overline{CTR_s} = \frac{1}{N_{J_s}} \sum_{j \in J_s} CTR_{jl} \tag{2}$$

Where J_s denotes the set of queries for which we observe a particular number of sessions s and N_{J_s} is the number of queries in J_s, i.e. for each $s \in [1, 10000]$ we observe in our dataset, we calculate the average over CTR_{jl} for all the queries which we observe at least s times. More formally, let S_j denote the total number of sessions we observe for a given query j, then $J_s : \{j : S_j \geq s\}$ and $N_{J_s} = |J_s|$.

The solid line of Figure 2 displays the local polynomial regression smooth of the average CTR as a function of the observed sessions. The dashed line is the local polynomial regression smooth of the initial quality of all the queries in J_s:

$$\overline{ICTR_s} = \frac{1}{N_{J_s}} \sum_{j \in J_s} CTR_{jl'} \tag{3}$$

Where l' is simply the first observed result list of each $j \in J_s$. The statistics displayed in Figure 2 reveal that the average quality we observe for queries in J_s is largely explained by the initial quality for the queries in J_s. The positive relationship between the initial quality and the in-sample popularity measure of the queries begs the question whether in-sample popularity conveys information about the pre-sample popularity and whether the positive relationship we observe in Figure 2 can be seen as indicative for learning from data. We will elaborate more on this points in Subsection 6.2.

While Figure 2 suggests that the in-sample quality evolution is, on average, small, Figure 3 reveals that the magnitude of the in-sample quality evolution of a query heavily depends on its initial quality. Figure 3 displays the in-sample quality evolution (as measured by the rolling

[8] In this section only, some results are based on the original quality measure, namely the results displayed in Figures 2 and 4

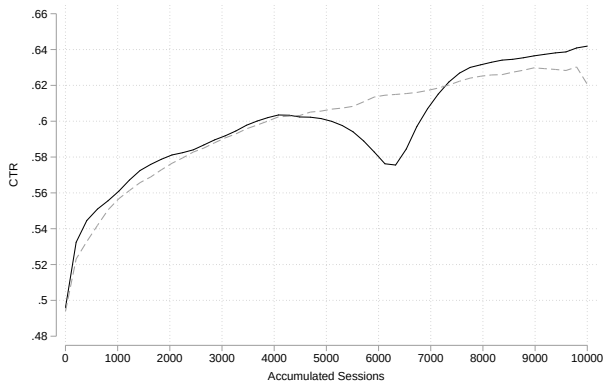


Figure 2 – Local Polynomial regression smooth of the average CTR and average initial CTR as a function of the observed sessions. The solid line refers to the average CTR, the dashed line refers to the average initial CTR. Calculations are based on original quality measure from Equation 1

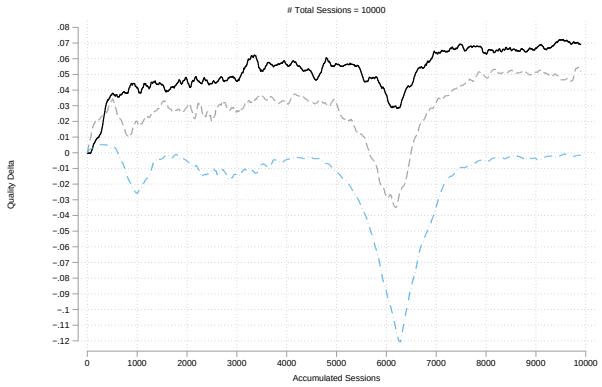


Figure 3 – In sample average quality evolution of the most popular queries for different initial quality categories. Solid line: $\overline{ICTR} \in [0, \frac{1}{3})$. Dashed line: $\overline{ICTR} \in [\frac{1}{3}, \frac{2}{3})$. Dash-Dotted light blue line: $\overline{ICTR} \in [\frac{2}{3}, 1]$. Calculations are based on the rolling window averages

window averages) for the most popular queries that accumulate evenly (Subsection 6.1 details the definition of even accumulation and provides the justification for their use, we decided to postpone this discussion for ease of exposition) for three different initial quality categories. More precisely, $\forall s$ we compute:

$$\overline{\Delta CTR}(s) = \frac{1}{N_{J_i}} \sum\nolimits_{j \in J_i} \left(\overline{CTR_j}(s) - \overline{CTR_j}(s=1) \right) \tag{4}$$

where s denotes the session that marks the left edge of each rolling window.[9] In our case s takes the values in $\{1, 11, 21, ..., 9901\}$. $\overline{CTR_j}(s)$ denotes the rolling window average over values of CTR_{jl} evaluated at s and $J_i : \left\{ j : \overline{CTR_j}(s=1) \in [a, b) \right\}$. In words, for the queries in J_i, we calculate the average difference between the rolling window values as s increases and the initial rolling window value ($s = 1$). The choice of a and b determines the initial quality category. In Figure 3, we chose $[a, b)$ to be $[0, \frac{1}{3})$ for the solid line, $[\frac{1}{3}, \frac{2}{3})$ for the dashed line and $[\frac{2}{3}, 1]$ for the dotted line. [10] As seen

[9] This choice is necessarily somewhat arbitrary, we could have chosen any arbitrary value inside the window. However, our choice allows interpreting the results displayed in figure 3 in a more natural way: By choosing the left edge of the window, the statistic in Equation 4 is computed and displayed right after a given number of searches elapsed.

[10] By basing the allocation of a query to a specific initial quality group on the rolling window average of the first 100 values of the original quality measure, we ensure we do not miss allocate queries. If we base our allocation rule on a single value of CTR_{jl}, we would run the risk of doing so based on rarely observed query-url combinations that are uninformative of the true performance of the search engine on that query for the first observed sessions.

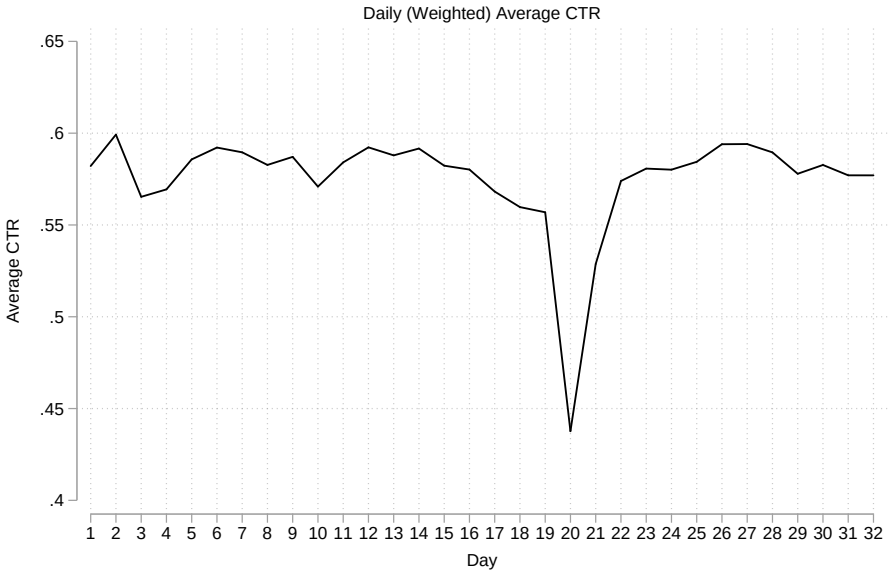


Figure 4 – Daily weighted average of the CTR, weights are calculated based on the daily search popularity of queries. Calculations are based on original quality measure from Equation 1

in Figure 3, the in-sample quality evolution ranges from 7% for the lowest initial quality group to 0% for the highest initial quality group. Accounting for the initial quality of a query is thus crucial when studying the in-sample quality evolution of queries. Figure 3 is representative and holds also for less popular queries.

Another apparent pattern in the data is the dip we observe around 6000 searches. This dip is likely due to the Yahoo! - Bing transition that occurred in the year 2010 and for which the testing phase happens to overlap with our sample period. The dip that we observe around 6000 searches largely coincides with a drop in quality on July $20th$, as most of the queries that were searched 10000 times reached this number of searches around this date. In fact, shortly before July 20, 2010, Yahoo! publicly announced its intention to begin testing Microsoft's search engine online, running the Bing algorithm on its real-life search traffic:

> *Though much of our testing is already happening offline, this month we'll also test the delivery of organic and paid search results provided by Microsoft on live Yahoo! traffic* -July 15, 2010 (Schwartz, 2010; Yahoo, 2010)

As seen in Figure 4, the queries experience a sudden and sharp quality reduction on July 20,

2010. The daily average search quality on that day is around 12.5 percentage points lower than the usual average. While it is extremely likely that the dip in quality is related to a timely limited algorithm transition for testing purposes, the exact course of events remains unclear. In our view, this sudden quality reduction is indicative of the importance of non-data related factors for the quality of search results because it is a sudden change in the employed technology that likely caused this event. In the remainder of the analysis, we ignore searches that occurred over the three days period between July 19 and 21, 2010.

5 Disentangling Data from Non-Data Related Factors

A natural starting point to analyze the impact of data in our dataset is to compare the quality evolution of queries depending on their popularity. If user-feedback data play an important role for the quality of search results, we would naturally expect to see that more popular queries experience a larger quality evolution than less popular queries.[11] Taking the difference in the quality evolution between more and less popular queries allows differencing out quality improving factors that are unrelated to data.

As a simple example, think of the case of two group of queries: Queries we observe 2000 times and queries that we observe 10000 times. If the first group experiences an average quality improvement of 5% and the latter a quality improvement of 8%, we would conclude that a minimum of a 5 fold increase in data in our sample period leads to a quality improvement of 3%. The differencing is sensible because it allows accounting for the fact that both groups might be subject to common quality driving factors. To better understand what these common factors could be, it is useful to revisit the major tasks of a search engine:

Crawling Search engines *crawl* the web to discover web pages available in the internet. This is done by programs called crawlers, spiders, or bots that move from site to site detecting and recording links to other web pages. These programs essentially map the internet by following the link structure on websites. Crawling the internet is costly as it requires computing power and storage space, with search engines differing in the size of their crawled database. It also

[11] User feedback is intrinsic in every action the user performs, thus a single search can be seen as "one unit of user-feedback."

seems that there is significant variation over time in the number of crawled web pages by various search engines.[12]

Indexing Search engines create a large database with keywords and other characteristics of the crawled web pages. Other characteristics may include the title of the webpage, words highlighted with different types of font, and appearing at different locations on the page. The extracted content is saved in a database that can be regarded as a stripped down and shortened version of the internet. Indexing involves significant costs for storing the index databases.

Ranking Search engines use a ranking algorithm to attach a relevance-weight of each crawled and indexed web page to various keywords. If a user enters a query into the search box, search engines go through their index to retrieve the web pages that they consider most relevant to the given keyword. Determining the relevance ordering of potentially millions of webpages to various keywords is the task of the ranking algorithm of the search engine.

From the above description, it should be clear that the technological requirements of a search engine offer ample room for non-data related quality factors. In our view, crawling and indexing have little to do with data on previous searches. Those tasks are essentially rather mechanic stocktaking and organization of the content of the web. The size of the web content inventory of a search engine is mainly determined by factors like storage space and computing capacity. These factors determine how frequently a search engine can crawl the web and the size of the database it stores.

The ranking task is probably the channel through which user feedback data are most likely to impact the quality of search results. Clicks performed by users and an assimilation of this clicking behavior to specific user types might help the search engine infer the preferences of other users when they search for a similar topic. However, it is also important to keep in mind that the ranking in modern search engine technology is also based on factors that have little or nothing to do with direct user feedback.[13]

[12] For an estimate of the number of web pages search engines know about, see http://www.worldwidewebsize.com/.

[13] One key determinant of the ranking is the link structure of the web. A specific website gains in relevance as an increasing number of other websites link to this specific website and as these other website also increase in relevance. This is the basic mechanism of Google's PageRank algorithm. While Yahoo's technology is undisclosed, industry experts agree that every search engine relies on similar principles in this regard.

In our view, the non-data related factors can best be accounted for by modeling them as time-dependent. Time might approximate or capture developments in the storage and computing capacities as well as the technological changes in ranking function, indexing methodology, and the crawler, among others. Thus, when comparing the quality evolution of queries with different levels of popularity, it is important to take time into account. Therefore, our analysis of the in-sample quality evolution of queries will focus on queries that we observe for a similar time span. More formally, we model the in-sample quality evolution of a query, ΔQ_i, as:

$$\Delta Q_i = f_i(TS_i, t_i) \tag{5}$$

Where f_i is the query specific function that determines how the total number of searches, TS_i, and elapsed time, t_i, translates into quality evolution during the sample period. TS_i and t_i are in-sample quantities. The main idea of our identification strategy is to fix t_i across all queries to $\bar{t}$. By doing so, differences in quality evolution should only be attributable to differences in TS and f. If f, which can be thought of as the "type" of a query, is orthogonal to TS, we can learn about the impact of data, by comparing the quality evolution of queries as TS varies. For the moment, we postpone any discussion about the pre-sample search history and how it might affect our empirical findings to the end of subsection 6.1.

6 Results

The result section is organized as follows: Subsection 6.1 presents the results based on the in-sample quality evolution for our one month sample period. Subsection 6.2 presents findings that we consider to be further corroborative evidence for the results from Subsection 6.1 and that highlight the long term perspective.

6.1 In - Sample Analysis

According to our hypothesis about the quality evolution process of queries formulated in Equation 5, we wish to focus on queries that we observe over the entire sample period to hold the impact of time, i.e. non-data related factors, constant. We achieve this by restricting the analysis to queries that we observe at least 2000 times and which accumulate evenly over the sample period. We base

our analysis on rolling window averages to obtain a reliable measure for the initial quality.[14] It is important to make sure that the time period that elapses for the calculation of one rolling window is sufficiently similar across and within queries. Otherwise, time related factors would affect the observed quality over the 100 searches used for one rolling window differently. This is undesirable if we wish to properly control for non-data related (i.e. time related) factors in the quality evolution process. It is for this reason that we focus on queries that accumulate evenly.

The definition of even accumulation is naturally somewhat arbitrary. In the context of our data, a query would accumulate perfectly evenly if each day it accumulates 1/32% of all its searches that we observe in the sample period. We base our criterion on a 3 day accumulation rate, i.e. 3/32%, and drop each query that deviates by more than 8 percentage point from this criterion in any three day period. From originally 10859 queries for which we observe more than 2000 searches, we are left with 9377 after applying the procedure. The interested reader is referred to the tables in Appendix A for query level summary statistics computed for different groups of queries.[15]

Another important concept in this section is the average amount of personalized knowledge that the algorithm had when confronted with a search on a given query. Each time we observe a search on a specific query, we know how often the cookie performing this search was observed in the sample before searching for that query. For each search on a given query, we record the number of times the cookie performing this search was observed previously and compute the average of this number for each query. In the remainder of this section, we will call this measure average cookie length (CL).

Remember from the discussion in the previous section that we are interested in the quality evolution of the queries during the sample period. Approximating the in-sample quality evolution as a linear function of the sessions, we can write:

$$Q_{is} = IQ_i + \beta_i \times s + \epsilon_{is} \tag{6}$$

The quality Q_{is} after a given number of sessions s is a linear function in s plus the initial

[14] As we show in Section 4, the initial quality is an important feature to explain quality evolution in-sample. A reliable measure for initial quality is only obtained by considering rolling window averages: Basing the initial quality on the first click or the first query-url combination that we observe does not provide a reliable estimate of the initial quality.

[15] The results of Figure 3 were computed based on the group of queries that accumulate evenly according to the above definition.

observed quality IQ_i. We are interested in estimating β_i given query level characteristics. Note that $\beta_i \times TS_i$ is the first order approximation of Equation 5 for $t_i \approx 32 \ \forall i$ (because we focus on queries that accumulate evenly). We estimate β_i by a varying coefficient model. More precisely, we estimate the following 4 specifications:

$$\begin{aligned}
(I) \quad & \beta_i = \alpha_0 + \alpha_1 \times CL_i + \alpha_2 \times TS_i + \alpha_3 \times IQ_i + u_i \\
(II) \quad & \beta_i = \alpha_0 + \alpha_1 \times CL_i + \alpha_2 \times TS_i + \alpha_3 \times IQ_i + \alpha_4 \times TS_i \times CL_i + u_i \\
(III) \quad & \beta_i = \alpha_0 + \alpha_1 \times CL_i + \alpha_2 \times TS_i + \alpha_3 \times IQ_i + \alpha_5 \times IQ_i \times CL_i + u_i \\
(IV) \quad & \beta_i = \alpha_0 + \alpha_1 \times CL_i + \alpha_2 \times TS_i + \alpha_3 \times IQ_i + \alpha_4 \times TS_i \times CL_i + \alpha_5 \times IQ_i \times CL_i + u_i
\end{aligned} \tag{7}$$

The coefficient α_1 tells us how the marginal impact of one additional session on the quality is affected by a longer average cookie length. We include the coefficient TS_i to allow for non-linearity in the in-sample quality evolution as a function of the number of total searches. It is only by including TS_i that we can capture scenarios in which queries with 2000 searches learn the same as queries with 10000 searches. Omitting TS_i would automatically imply that queries with 10000 searches learn more than queries with 2000 searches, unless $\beta_i = 0$. IQ_i captures the change of the learning slope as a function of the initial quality. Naturally, queries that start with a high initial quality have less scope for learning. In the descriptive part of the analysis, we show that the observed quality evolution depends heavily on the initial quality level of a query. The specifications $(II) - (IV)$ allow for potential interactions between CL and TS and/or IQ. Estimation of the α parameters is achieved by OLS after plugging the equations in 7 in Equation 6. Standard errors are clustered on the query level to account for the query specific error term u_i.

The regression results for the 4 specifications are presented in Table 1 from left to right. Before commenting on the results in more detail, we want to point out that our regression specification allows for an overall constant and a coefficient for the standalone initial quality variable that which is different from one. While the inclusion of an overall constant is innocuous and guarantees that the error has mean zero, the coefficient on the standalone initial quality variable departs from the model formulation in Equation 6. Restricting the coefficient to 1 does not change the qualitative implications of our results, quantitatively the results only change very marginally. The results for the restricted version are presented in Appendix B.1.

Table 1 – Regression Results for Specifications (I) - (IV)

	(I)	(II)	(III)	(IV)
initial quality	9.417e-01***	9.389e-01***	9.395e-01***	9.379e-01***
	(3.052e-03)	(3.168e-03)	(3.132e-03)	(3.177e-03)
sessions	2.542e-06**	-6.260e-06**	-1.024e-05**	-1.517e-05***
	(8.455e-07)	(2.288e-06)	(3.424e-06)	(2.927e-06)
sessions x cookie (SC)	1.115e-06***	2.919e-06***	4.067e-06***	5.072e-06***
	(1.353e-07)	(4.345e-07)	(8.300e-07)	(6.967e-07)
sessions x total sessions	-3.686e-11	8.793e-10***	-8.492e-11	4.582e-10
	(1.020e-10)	(2.644e-10)	(1.019e-10)	(2.993e-10)
sessions x initial quality	-1.080e-05***	-1.046e-05***	7.555e-06	7.365e-06
	(1.132e-06)	(1.141e-06)	(4.113e-06)	(4.205e-06)
SC x total_sessions		-1.904e-10***		-1.127e-10*
		(4.802e-11)		(5.506e-11)
SC x initial quality			-3.893e-06***	-3.810e-06***
			(9.696e-07)	(1.002e-06)
Constant	3.998e-02***	4.174e-02***	4.128e-02***	4.229e-02***
	(2.083e-03)	(2.158e-03)	(2.149e-03)	(2.176e-03)
Observations	5066880	5066880	5066880	5066880
Adjusted R^2	0.883	0.883	0.883	0.883

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The coefficient α_1 (SC) is positive and significant across all presented specifications.[16] It reveals a positive relationship between the average cookie length and speed of the observed quality improvement. The coefficients on the interaction between the average cookie length and initial quality variables are always negative, which is natural: If a longer average cookie history implies a larger quality improvement it nevertheless remains true that queries starting from a high initial quality level cannot experience a large quality improvement. Thus, the slope for queries with a longer cookie history must adjust downward more sharply than the slope for queries with a shorter cookie history. The coefficients on the interaction between the cookie and the total searches variables are also negative, indicating that the positive impact of the cookie length is somewhat mitigated when allowing for non-linearities, as mentioned above. However, this mitigation of the cookie impact is very limited in magnitude, as will become clear in what follows.

The regression results are better interpreted and understood in terms of fit. The following

[16]This result is robust to the choice of the quality measure, the inclusion of further standalone variables in the above specification, and the restriction of the coefficient on the standalone variable on initial quality to the value of one. It can also be shown that the pattern emerges when performing a non-parametric analysis: Comparing the mean quality evolution of queries with an above median and below median average cookie length, we find that queries with an above median average cookie length learn faster; the results of this analysis are displayed in Appendix B.2

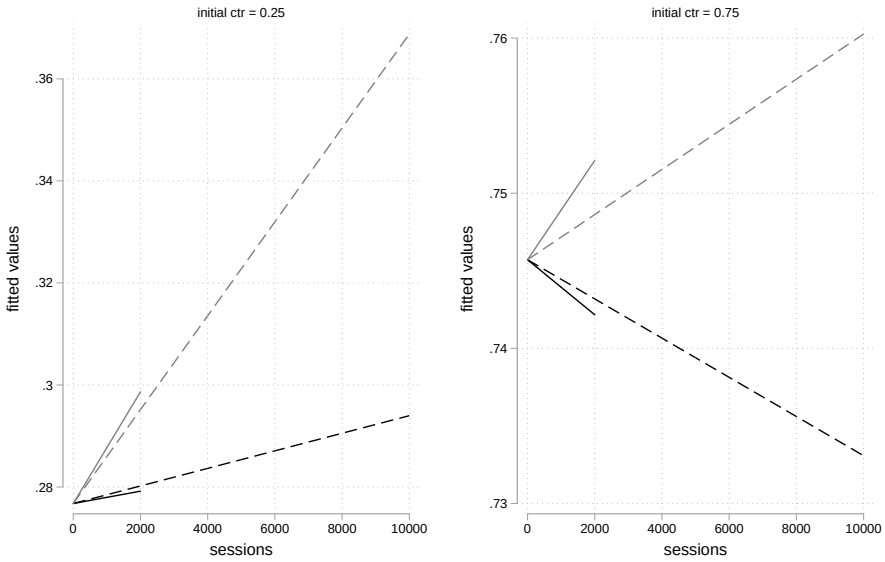


Figure 5 – Predicted quality evolution of model (IV). The left panel refers to queries with an initial CTR of 25%, the right panel to queries with an initial CTR of 75%. The black lines refer to queries with an average cookie length of 3.5, the gray lines to queries with an average cookie length of 6. The total number of searches of the queries is displayed on the x-axis.

discussion is based on the most flexible specification (IV), the implication of the regression results for the less flexible specifications should be easily understood from there. Figure 5 displays the quality evolution for queries with an initial quality of 25% in the left panel and 75% in the right panel.[17] The cookie length chosen for the exposition in Figure 5 correspond to the rounded values of the 25% and 75% percentile of the cookie length distribution of the queries used in the estimation (see tables in Appendix A).

As can be easily seen, a query with a longer average cookie history experiences a larger quality evolution than a query with a shorter cookie history but otherwise identical parameters. The negative coefficient on the interaction of the cookie length variable with the initial quality variable leads to a reduction of this spread as the initial quality increases. The interaction between the total searches variable and the cookie length variable allows the difference in slopes between queries with 2000 and queries with 10000 searches to vary across different cookie lengths. The negative sign indicates the mitigation of the positive cookie effect as the total number of searches increases.

[17] The fact that the origin of the curves in both panels is slightly different from 25% and 75% is due to the constant included and the unrestricted coefficient on the standalone initial quality variable in the model.

As suspected from the small magnitude of the respective coefficient in the regression results, this impact seems negligible.

As explained in the paragraph following Equation 6, the quality difference between the origin and the end point of each line can be seen as the average first order approximation of Equation 5. As elaborated in the discussion of Section 5, we propose to consider the difference in the quality evolution between queries with different popularity to account for non-data related factors. With this logic, the difference between the quality improvement described by the gray lines in the left panel of Figure 5 corresponds to the impact of a minimum increase in popularity of least 5 times for queries with an average cookie length of 6 and an initial quality of 25%. Analogously, the difference between the quality improvement described black lines of the left panels corresponds to the impact of an at least 5 times increase of popularity with an average cookie length of 3.5 and an initial quality of 25%. Furthermore, the difference between these differences can be considered as the reinforcement effect of the longer average cookie history on the data impact.

Table 2 displays the point estimates as well as the p-values of a corresponding analysis for 3 different initial quality levels for all specifications $(I) - (IV)$ ordered from top to bottom. The left column pair refers to the data impact for queries with an average cookie history of 6, the middle column pair refers to the data impact for queries with an average cookie history of 3.5 . The right pair refers to the reinforcement effect, which is always positive and significantly different from zero. Including an interaction effect between the cookie and the initial quality variables (which we argued is reasonable) changes the results substantially. Unsurprisingly, the interaction effect between the cookie and the total sessions variable only has a minor impact

According to our preferred specifications, the impact of a longer cookie history, as measured by the interquartile range of the cookie length distribution, on an at least 5 fold increase in data ranges from between 1.7% and 2.3% for queries with a high initial quality to between 5.5% and 6.2% for queries with a low initial quality. One reason for the faster learning process of queries with a longer average cookie history might be that better knowledge about the individuals searching for content enables the algorithm to assimilate this feedback to user-types. Learning from users who are well known to the algorithm, enables the algorithm to better predict identical or similar future search request by other users (who are also well known to the algorithm) by associating their requests to the respective type. Another potential channel might be the greater reliability of

Table 2 – Differences and Differences in Differences

	CL = 6		**CL = 3.5**		**Diff**	
Initial quality	Estimate	P-val	Estimate	P-val	Estimate	P-val
Specification (I)						
.25	0.049***	0.000	0.026***	0.000	0.022***	0.000
.5	0.027***	0.000	0.005	0.170	0.022***	0.000
.75	0.006*	0.022	-0.017***	0.000	0.022***	0.000
Specification (II)						
.25	0.044***	0.000	0.031***	0.000	0.013**	0.002
.5	0.023***	0.000	0.010*	0.013	0.013**	0.002
.75	0.002	0.422	-0.011*	0.015	0.013**	0.002
Specification (III)						
.25	0.074***	0.000	0.012*	0.034	0.062***	0.000
.5	0.042***	0.000	-0.000	0.905	0.042***	0.000
.75	0.010***	0.000	-0.013***	0.000	0.023***	0.000
Specification (IV)						
.25	0.070***	0.000	0.015*	0.021	0.055***	0.000
.5	0.039***	0.000	0.003	0.553	0.036***	0.000
.75	0.008**	0.009	-0.009*	0.016	0.017***	0.000

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

signals submitted by well-known or persistent users. A user repeatedly or persistently searching for a specific content might convey much more information than users who quickly give up on their quest for information. Furthermore, the search engine, by studying the browsing behavior of individuals, might learn about the quality of the signals specific user types send when interacting with the search engine in a specific way.

As mentioned at the end of Section 5, the pre-sample search history likely plays a role in the proper assessment of the data impact on the in-sample quality evolution. The queries that we compare in this section to learn about the impact of data should ideally be similar in terms of pre-sample characteristics. It is reasonable to suspect that queries with a larger in-sample popularity also had a larger pre-sample popularity. Given non-linearities in the learning process, this is problematic. For instance, in the scenario of a concave learning function, our empirical analysis would likely understate the impact of data, but in the scenario of a convex learning function, the impact is likely overstated. We note that if such processes are at work in our data, we would expect the variable TS to play a more noticeable role in our analysis. However, as we show, TS only seems to play a minor role. As far as the pre-sample exposure time of the queries to the algorithm is concerned, we assume that this latent factor is orthogonal to the observable characteristics we use in our analysis. Further research based on more comprehensive data sets is warranted in order to explicitly take into account those potential confounders.

Despite these caveats, we find overwhelming evidence in our data that the speed of quality evolution is significantly and positively impacted by the average amount of personalized information the algorithm has on the searchers. This result is robust to various specifications and the quality measure. Giving explicit consideration to non-data related factors in the quality evolution process, we find a statistically significant and positive impact of data that is reinforced by personalized information. The next section conceptualizes on the long run impact of data and provides further corroborative evidence of the positive impact that personalized information has on the speed of learning and the importance of data in general.

6.2 Long Run Considerations

If data-accumulation matters and if personalized information reinforces its impact on search result quality during the sample period, we would also expect to see these forces at play when considering the long run perspective. In Section 4, we already mention that the positive relationship between the total number of searches in-sample and the initial quality (see Figure 2) might be indicative of a positive relationship between the pre-sample popularity and quality.

In this section, we elaborate on the conditions under which the relative in-sample popularity of queries can be considered a good proxy for their relative pre-sample popularity and show that when we focus on queries that are likely to fulfill these conditions, we obtain a clearer relationship between in-sample popularity and initial quality. Remember that we think of the quality of a query i as a function of the accumulated searches TS and the time it was exposed to the algorithm t; in this section the reader should think about $\hat{t}$ and $\hat{TS}$ as unobserved pre-sample quantities. For the initial quality of a query, we write:

$$IQ_i = f_i(\hat{TS}_i, \hat{t}_i) \tag{8}$$

Note that $\hat{TS}_i = \frac{\hat{TS}_i}{\hat{t}_i} \times \hat{t}_i = \hat{\bar{s}}_i \times \hat{t}_i$. The accumulated search history pre-sample is simply the average number of searches by time unit, $\hat{\bar{s}}_i$, multiplied by the elapsed time, $\hat{t}_i$, since the query first appeared. For the remainder of the exposition, we assume that $\hat{t}_i$ is orthogonal to the average search frequency per time unit, $\hat{\bar{s}}_i$. We consider this assumption fairly innocuous: We simply claim that the average search frequency by month (or day, hour or second) is not systematically related to the first time a query appeared.

We wish to find queries for which the relative in-sample popularity is a good proxy for the relative pre-sample popularity, i.e. we wish to find queries where for arbitrary i, j, we have:

$$\mathbb{E}\left(\frac{TS_i}{TS_j}\right) \approx \mathbb{E}\left(\frac{\hat{TS}_i}{\hat{TS}_j}\right) \tag{9}$$

Where the quantities without a hat denote observed in-sample quantities. Note that, according to the above assumption about the orthogonality between time and average searches by unit of time and random sampling, this is equivalent to:

$$\mathbb{E}\left(\frac{s_i}{s_j}\right) \approx \mathbb{E}\left(\frac{\hat{\bar{s}}_i}{\hat{\bar{s}}_j}\right) \tag{10}$$

The approximation in Equation 10 becomes more accurate as the accuracy of the numerator and denominator on the left hand side of the equation increasingly approximates the numerator and denominator of the right hand side. In other words, if we focus on queries where the total search quantity over our one month sample period is a good estimator for the average search quantity per month before the sample period, the relative in-sample popularity should approximate the relative pre-sample popularity fairly well.

Which queries from our sample do we need to focus on to make the relative in-sample popularity a good proxy for the relative pre-sample popularity? We need to focus on queries where the monthly variance in popularity is small. This is likely the case for queries that converge to a constant level of popularity. Queries that experience an abrupt drop or surge in popularity during our sample period convey little information about their average pre-sample monthly popularity; the same is true for queries with a decreasing and increasing trend, which are likely to be subject to seasonal patterns.

To drop those queries, we elaborate on the idea of even accumulation introduced in the previous section. If we drop queries that deviate from our even accumulation criterion, we should get rid of trending queries or queries with erratic popularity changes. Of course, no query perfectly complies with our constant accumulation criterion of $\frac{3}{32}\%$ in any three day period. However, we can define bounds around this criterion that queries need to fulfill in order not to be dropped and we can shrink those bounds arbitrarily close to the perfect constant accumulation criterion.

In the next paragraphs, we show that the stricter the bounds, the clearer the relationship

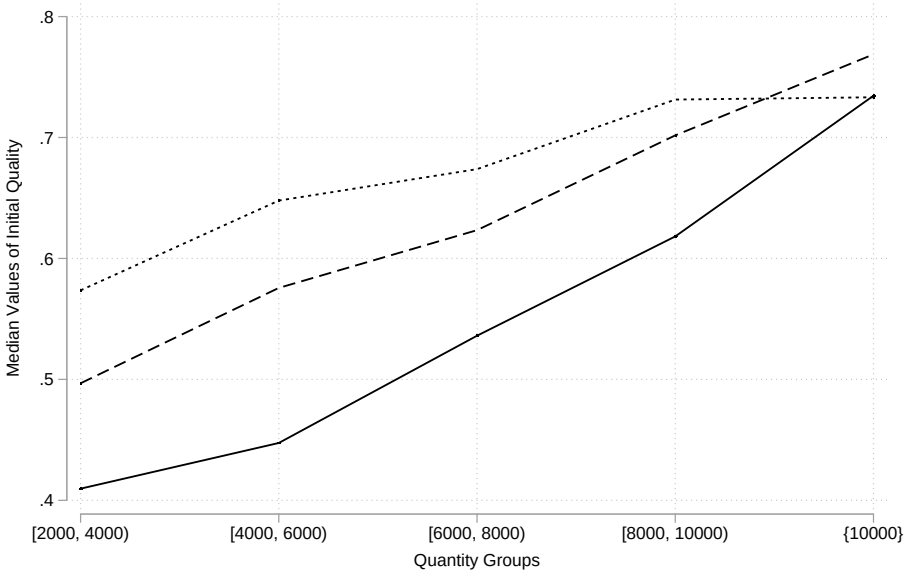


Figure 6 – Relationship between the median initial quality and the number of total searches. Each median value is calculated for the queries falling within a specific quantity group, as specified on the x-axis. The dotted line displays the median values calculated for all queries. The dashed line is calculated based on queries that deviate no more than 4 percentage points from our constant frequency definition. The solid line is calculated based on queries that deviate no more than 3 percentage points from our constant frequency definition.

between the initial quality we observe and the in-sample popularity (wish proxies pre-sample popularity). This is exactly what we would expect to happen if accumulation of data (and hence feedback) matters because in-sample relative popularity becomes a better proxy for relative pre-sample popularity. Note that the orthogonality assumption between $\hat{t}_i$ and $\hat{\hat{s}}_i$ implies that the elapsed exposure time before the sample period started is - on average - constant across different search quantities.

Figure 6 displays the medians for queries belonging to different quantity groups. As shown, the relationship between the initial quality and the in-sample popularity becomes clearer, the stricter we define the bounds around the constant frequency definition. Under the hypothesis that data accumulation does not matter, we would not expect to see this pattern emerge. The confidence intervals around the medians of the dotted and solid lines do not overlap for the first four quantity groups (see Appendix B.3). It is also remarkable that the differences between the median value

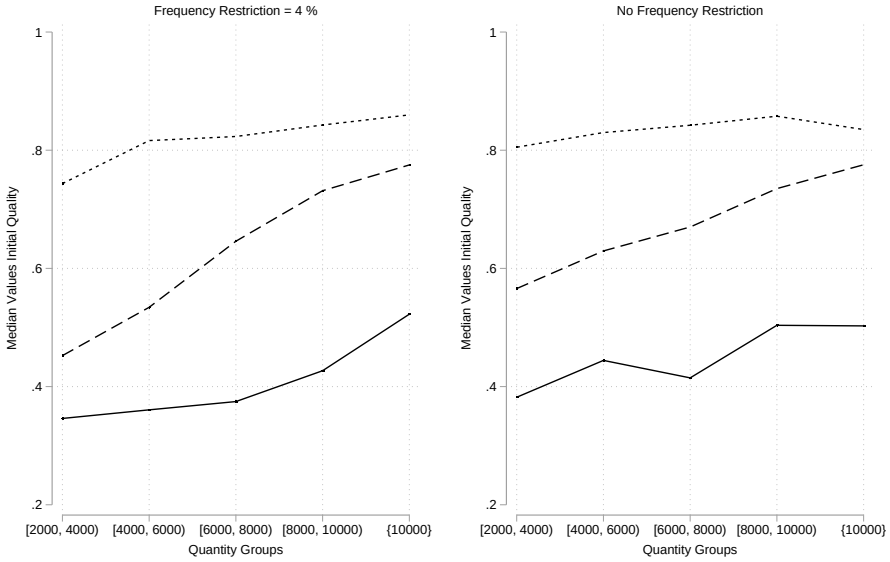


Figure 7 – Relationship between the median initial quality and the number of total searches. The dotted lines refer to queries with an average cookie length exceeding the 66% percentile. The dashed lines refer to queries with an average cookie length between the 33% and 66% percentile. The solid lines refer to queries below the 33% percentile. The percentiles vary across group of total searches and frequency restrictions. The left panel displays the result for the queries that deviate no more than 3% from our constant frequency definition. The right panel displays the results for all the queries in the sample (no restriction on the search frequency)

of smallest and largest quantity group gradually increases as the constant frequency requirement becomes stricter.[18]

To assess the impact of the average amount of personalized information on the long run impact of data, we split the queries into three groups based on the 33% and 66% percentile of the average cookie length distribution. Figure 7 displays the relationship between the in-sample quality and the initial quality for each of these three groups. The left panel refers to queries that deviate no more than 4 percentage points from the perfect constant frequency definition. The right panel refers to all queries with more than 2000 searches in our sample (i.e. no restriction on the type of accumulation). Again, focusing on queries fulfilling a stricter constant frequency requirement seems to strengthen the relationship between in-sample total searches and initial quality for each cookie length group.

[18] We decided to only show three lines for ease of exposition, the graph for seven different thresholds is displayed in Appendix B.3

Furthermore, the results are consistent with the results from the in-sample analysis in Subsection 6.1, which provided strong evidence for the hypothesis that the speed of learning seems to be reinforced by a longer average cookie history. Two observations point in this direction: firstly, for the same number of in-sample searches (which proxy pre-sample searches) the initial quality is higher for queries with longer average cookie history. Secondly, the relationship between the in-sample searches and the initial quality is highest for queries between the 33% and 66% percentile and comparatively flat for the lower and upper 33% of the cookie length distribution. This is consistent with a scenario in which a longer average cookie history increases the speed of learning to a point where a majority of queries belonging to the upper 33% of the cookie length distribution are already out learned at the beginning of our sample period (also for comparatively small values of pre-sample popularity).

Of course the conclusions of the above analysis should be taken with a grain of salt: A potential important confounding factor, like the "type" of a query (difficult vs. easy), cannot be accounted for due to data limitations. However, the above analysis combined with the in-sample analysis from Subsection 6.1 (where a confounding factor like the "type" of a query could to some extent be controlled for by accounting for the initial quality) provide strong indication in favor of the hypothesis that user feedback data matter in both the long- and short-runs. Furthermore, the speed of learning is positively influenced by the average knowledge the algorithm has about the searcher.

7 Conclusion

In this paper, we empirically examine the question of to what extent does user data about previous searches drive the quality of subsequent internet search results. To do so, we provide a method to disentangle the effects of user data from other factors, such as the size of the search engine's indexed content or the quality of the search engine's ranking algorithm.

We find that additional data on previous searches on the same keyword tends to improve the quality of search results. Moreover, our results highlight that the type of data matters. In particular, having a larger quantity of personalized information about the clicking behavior of users increases the speed of learning from previous data. Additional personalized information systematically leads

to quicker quality improvement. At the same time, we find indications that factors not directly related to user data play an important role.

Our insights have important management implications and can be generalized to any search recommendation technology in e-commerce, including information and product search. Such technologies are critical elements of e-commerce and are widely used by leading firms, such as Netflix and Amazon. Anecdotal evidence suggest that the quality of recommendation technology improves significantly with additional data. This paper suggests methods to disentangle the effect of data from other factors driving the quality of recommendations. The data requirements are very low: simple activity logs routinely retained by these e-commerce services suffice. Managers need to consider carefully which types of data to invest in and where to allocate resources for data analysis. A key factor in this decision should be which type of data improves the quality of results the most. Our results show that personalized data is the most valuable. Therefore, closely tracking the activity of a few users over time may be more valuable for businesses than collecting non-personalized data covering a large number of users.

Our results may also inform competition policy by qualifying, in a detailed manner, how data may give rise to economies of scale. We find evidence that data matters for internet search quality. Our results also call for awareness from antitrust policy regarding potentially anti-competitive firm behavior that seeks to lock in relatively few users for a longer period of time. Since information about the individual user is particularly important in triggering economies of scale, capturing users for a longer time period may grant firms a data advantage that is difficult for competitors to overcome. Assessing how market demand depends on search result quality (and, therefore, on data) and to what extent accumulated data impose an entry barrier are interesting questions for further research.

Finally, our insights are also highly relevant for consumer and privacy policy. They lend support to initiatives that enable users of IT services to easily carry their personal data to other service providers, including Article 20 of the General Data Protection Regulation, which provides EU citizens the right to data portability. As it is individual data that drives recommendation quality the most, allowing users to control and carry their data to competitors may be a smart way of mitigating potential market power. However, at the same time, our results are also consistent with personalized data on one user enabling the service provider to offer better results to other

users. Thus, there might be externalities from data across users. These externalities, in turn, may even result in suboptimal levels of switching justifying consumer policy attention. A systematic theoretical and empirical assessment of user switching in light of externalities from data across users would be a valuable contribution to future research.

References

Ansari, A. and Mela, C. F. (2003). E-customization. *Journal of marketing research*, 40(2):131–145.

Argenton, C. and Prüfer, J. (2012). Search engine competition with network externalities. *Journal of Competition Law and Economics*, 8(1):73–105.

Arora, S. (2016). Recommendation engines: How Amazon and Netflix are winning the personalization battle. (28 June 2016), MTA, Martech Advisor, San Francisco, CA. https://www.martechadvisor.com/articles/customer-experience-2/recommendation-engines-how-amazon-and-netflix-are-winning-the-personalization-battle/, (last accessed: 19 March 2018).

Bajari, P., Chernozhukov, V., Hortaçsu, A., and Junichi, S. (2018). The impact of big data on firm performance: An empirical investigation. NBER Working Paper No. 24334, http://www.nber.org/papers/w24334, (last accessed 19 March 2018).

Bodapati, A. V. (2008). Recommendation systems with purchase data. *Journal of marketing research*, 45(1):77–93.

Chen, H., Chiang, R. H., and Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. *MIS quarterly*, pages 1165–1188.

Chintagunta, P., Hanssens, D. M., and Hauser, J. R. (2016). Marketing science and big data. *Marketing Science*, 35(3):341–342.

Chiou, L. and Tucker, C. (2017). Search engines and data retention: Implications for privacy and antitrust. NBER Working Paper No. 23815, http://www.nber.org/papers/w23815, (last accessed 19 March 2018).

De, P., Hu, Y., and Rahman, M. S. (2010). Technology usage and online sales: An empirical study. *Management Science*, 56(11):1930–1945.

Dou, Z., Song, R., and Wen, J.-R. (2007). A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web*, pages 581–590. ACM.

Ferenstein, G. (2013). Google, competition and the perfect result. The Washington Post (4 January 2013), https://www.washingtonpost.com/national/on-innovations/google-competition-and-the-perfect-result/2013/01/04/fc3eceda-4551-11e2-9648-a2c323a991d6_story.html?utm_term=.a8eb817d3f99, (last accessed 19 March 2018).

FTC (2017). Privacy & data security update (2016). FTC Report on Privacy and Data Security, https://www.ftc.gov/reports/privacy-data-security-update-2016, (last accessed 19 March 2018).

GDPR (2016). General data protection regulation of the European parlament and council, regulation 2016/679. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679, (last accessed 19 March 2018).

George, G., Haas, M. R., and Pentland, A. (2014). Big data and management. *Academy of management Journal*, 57(2):321–326.

Goldfarb, A. and Tucker, C. E. (2014). Standardization and the effectiveness of online advertising.

Management Science, 61(11):2707–2719.

Guardian, The. (2015). Google dominates search. But the real problem is its monopoly on data. (19 April 2015), https://www.theguardian.com/technology/2015/apr/19/google-dominates-search-real-problem-monopoly-data, (last accessed: 19 March 2018).

Lambrecht, A. and Tucker, C. (2013). When does retargeting work? Information specificity in online advertising. *Journal of Marketing Research*, 50(5):561–576.

Lerner, A. V. (2014). The role of 'big data' in online platform competition. Available on SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2482780, (last accessed 19 March 2018).

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., and Barton, D. (2012). Big data: the management revolution. *Harvard business review*, 90(10):60–68.

McAfee, P., Rao, J., Kannan, A., He, D., Tao, Q., and Liu, T.-Y. (2015). Measuring scale economies in search. PowerPoint Presentation at LearConference2015 (June 25–26), Learlab, Rome, Italy. http://www.learconference2015.com/wp-content/uploads/2014/11/McAfee-slides.pdf, (last accessed 19 March 2018).

Moe, W. W. and Fader, P. S. (2004). Dynamic conversion behavior at e-commerce sites. *Management Science*, 50(3):326–335.

Pasquale, F. (2015). The black box society: the secret algorithms that control money and information. (Cambridge: Harvard University Press).

Rubinfeld, D. L. and Gal, M. S. (2017). Access barriers to big data. *Arizona Law Review*, 59:339–381.

Schepp, N.-P. and Wambach, A. (2015). On big data and its relevance for market power assessment. *Journal of European Competition Law & Practice*, 7(2):120–124.

Schwartz, B. (2010). Official: Yahoo testing Bing powered results in July (aka now). Search Engine Roundtable (15 July 2010), https://www.seroundtable.com/archives/022555.html, (last accessed 19 March 2018).

Sokol, D. and Comerford, R. (2017). Does antitrust have a role to play in regulating big data? *The Cambridge Handbook of Antitrust, Intellectual Property, and High Tech (Cambridge University Press)*, pages 293–316.

Stucke, M. E. and Grunes, A. P. (2015). Debunking the myths over big data and antitrust. CPI Antitrust Chronicle (May), Available on SSRN : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2612562 (last accessed 19 March 2018).

Varian, H. (2016). Bruegel conference on big data, digital platforms and market competition. Video Recording at 15:17 (October 3), Brussels European and Global Economic Laboratory, Brussels, Belgium. http://bruegel.org/events/big-data-digital-platforms-and-market-competition/6/, (last accessed 19 March 2018).

Yahoo (2010). As we continue to work closely with Microsoft to implement our search alliance, we wanted to provide you an update on our progress. Web page archieved on 15 July 2010, https://web.archive.org/web/20100718193504/http://ebm.yahoo-email.com/c/tag/hBMPzDXAdp9-oB8OqGyAku8rdWn/doc.htm, (last accessed 19 March 2018).

Yoganarasimhan, H. (2016). Search personalization using machine learning. Available on SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2590020, (last accessed 19 March 2018).

Appendices

Appendix A

Table 3 – Summary statistics all queries

	mean	sd	min	p25	p50	p75	max	count
total searches	1194.04	2467.43	3	19	115	851	10000	67652
average cookie length	4.06	2.18	1	2.79	3.56	4.78	67.3	67652
average pos. last click*	2.18	1.19	1	1.21	1.84	2.85	10	66370
share generic clicks	0.56	0.25	0	0.37	0.57	0.78	1	67652
share clicks on pos 0	0.25	0.18	0	0.096	0.24	0.38	1	67652
share clicks on other**	0.09	0.09	0	0.018	0.058	0.12	1	67652
share no clicks	0.10	0.11	0	0.038	0.067	0.12	1	67652

Note: All click measures refer to the position of the last click

* only generic clicks are considered

** other clicks sumarize clicks below last url, new searches and clicks encoded as *o*

Table 4 – Summary statistics queries with more than 2000 searches

	mean	sd	min	p25	p50	p75	max	count
initial CTR on pos 1	0.58	0.29	0	0.31	0.66	0.84	1.00	10859
total searches	6041.51	2998.11	2001	3132	5471	9653	10000	10859
average cookie length	4.98	1.99	1.08	3.54	4.51	5.95	19.6	10859
average pos. last click*	1.79	0.84	1	1.14	1.41	2.30	7.52	10858
share generic clicks	0.62	0.24	0	0.44	0.65	0.84	0.97	10859
share clicks on pos 0	0.22	0.16	0	0.077	0.19	0.35	0.83	10859
share clicks on other**	0.06	0.06	0	0.021	0.042	0.086	0.56	10859
share no clicks	0.10	0.11	0.0079	0.047	0.062	0.095	1	10859

Note: All click measures refer to the position of the last click

* only generic clicks are considered

** other clicks sumarize clicks below last url, new searches and clicks encoded as *o*

Table 5 – Summary statistics queries used in regression analysis

	mean	sd	min	p25	p50	p75	max	count
initial CTR on pos 1	0.59	0.29	0	0.34	0.68	0.84	1.00	9377
total searches	6060.99	2986.75	2001	3162	5509	9638	10000	9377
average cookie length	4.93	1.89	1.08	3.53	4.48	5.93	19.0	9377
average pos. last click*	1.76	0.84	1	1.14	1.38	2.23	7.52	9376
share generic clicks	0.63	0.23	0	0.47	0.67	0.84	0.97	9377
share clicks on pos 0	0.22	0.16	0	0.076	0.19	0.35	0.83	9377
share clicks on other**	0.06	0.06	0	0.021	0.040	0.081	0.56	9377
share no clicks	0.09	0.10	0.010	0.047	0.061	0.088	1	9377

Note: All click measures refer to the position of the last click

* only generic clicks are considered

** other clicks sumarize clicks below last url, new searches and clicks encoded as *o*

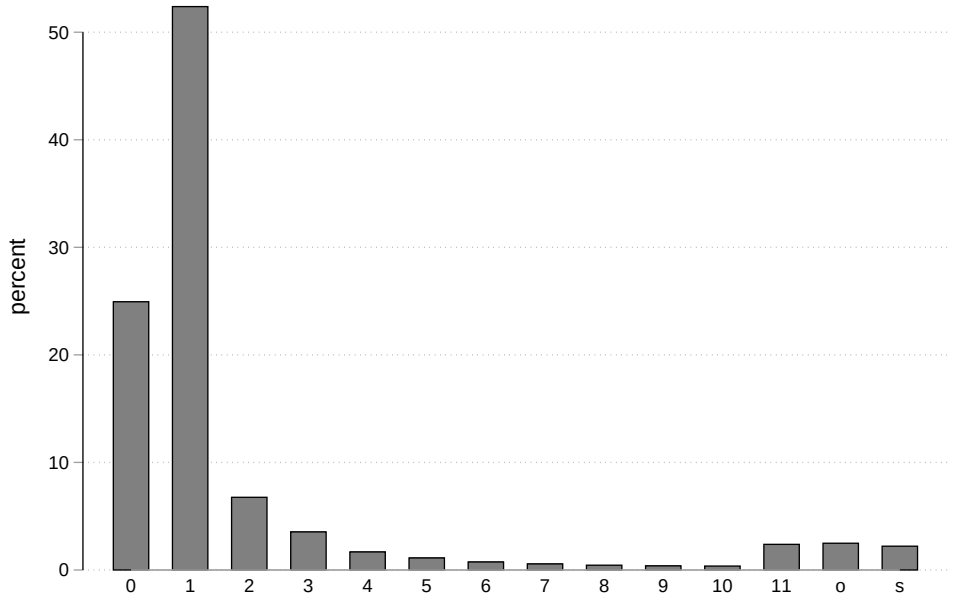


Figure 8 – Distribution of the position of the last click. Searches that end with no click are discarded from the calculation. Positions 1 to 10 describe generic urls. Position 0 is extensively discussed in Section 3.1. Position 11 describes an url below the 10 generic urls (such as "go to next page"). "s" stands for a new search and " o" for another click (see Section 3).

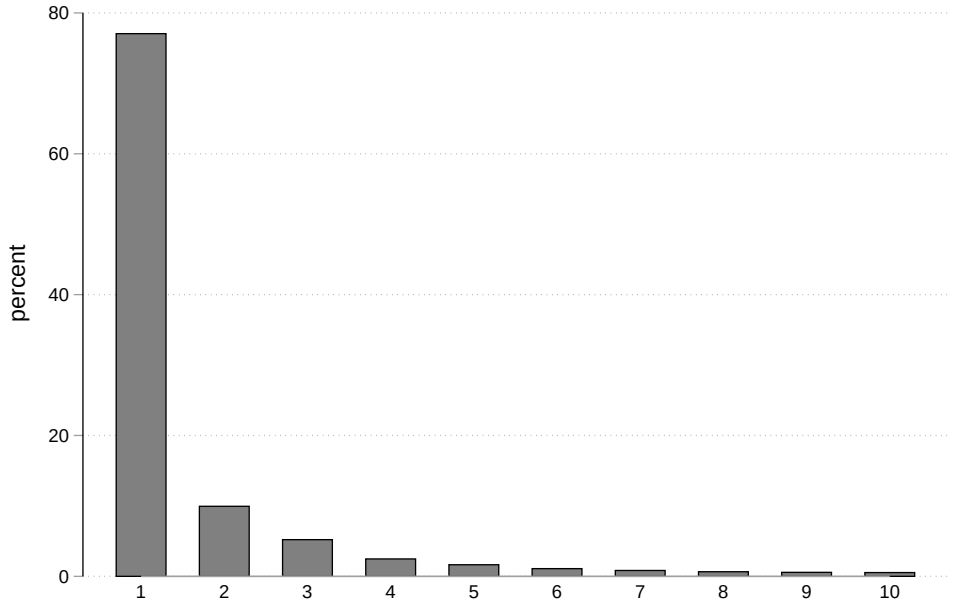


Figure 9 – Distribution of the position of the last click conditional on the last click being performed on a generic url

Appendix B

B.1 Regression tables restricted vs. unrestricted estimation

Table 6 – Restricted Results for Specifications (I) - (IV)

	(I)	(II)	(III)	(IV)
initial quality	1.000e+00	1.000e+00	1.000e+00	1.000e+00
	(.)	(.)	(.)	(.)
sessions	9.804e-06***	8.459e-06***	-1.969e-06	3.320e-08
	(9.094e-07)	(2.205e-06)	(3.364e-06)	(2.833e-06)
sessions x cookie (SC)	1.160e-06***	1.447e-06***	3.940e-06***	3.515e-06***
	(1.343e-07)	(4.228e-07)	(8.308e-07)	(6.817e-07)
sessions x total sessions	-1.256e-10	1.942e-11	-1.740e-10	-4.034e-10
	(1.028e-10)	(2.594e-10)	(1.029e-10)	(2.968e-10)
sessions x initial quality	-2.150e-05***	-2.153e-05***	-4.606e-06	-4.390e-06
	(1.143e-06)	(1.143e-06)	(3.997e-06)	(4.132e-06)
SC x total_sessions		-3.028e-11		4.781e-11
		(4.662e-11)		(5.423e-11)
SC x initial quality			-3.664e-06***	-3.702e-06***
			(9.697e-07)	(9.909e-07)
Constant	4.159e-03***	4.169e-03***	4.117e-03***	4.101e-03***
	(9.585e-04)	(9.591e-04)	(9.572e-04)	(9.572e-04)
Observations	5066880	5066880	5066880	5066880
Adjusted R^2				

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

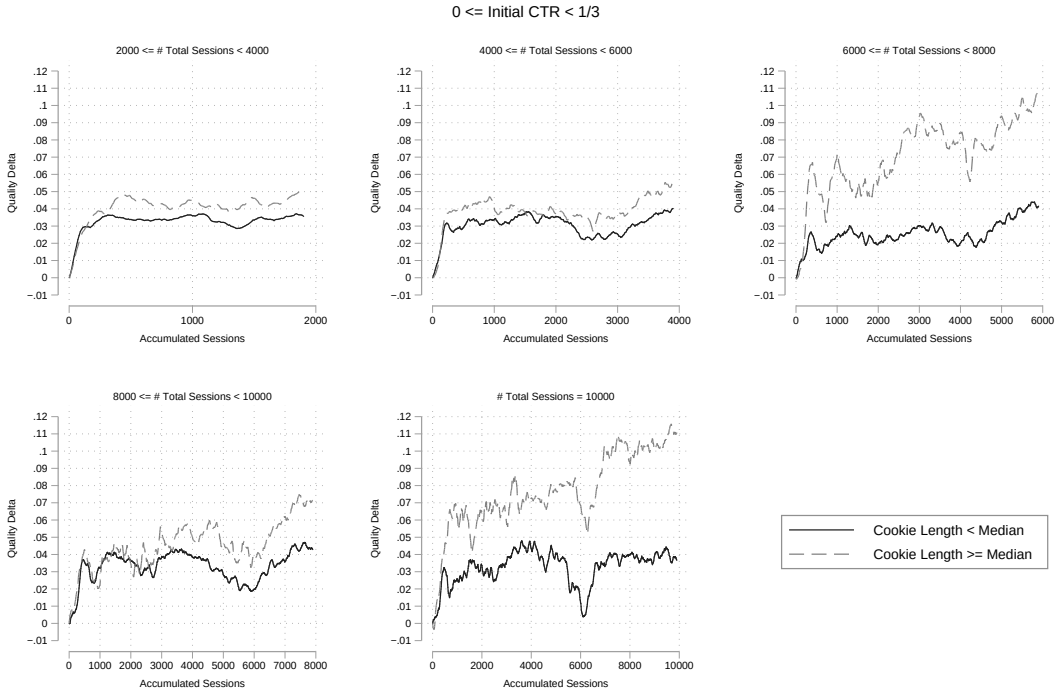
Table 7 – Original Results for Specifications (I) - (IV)

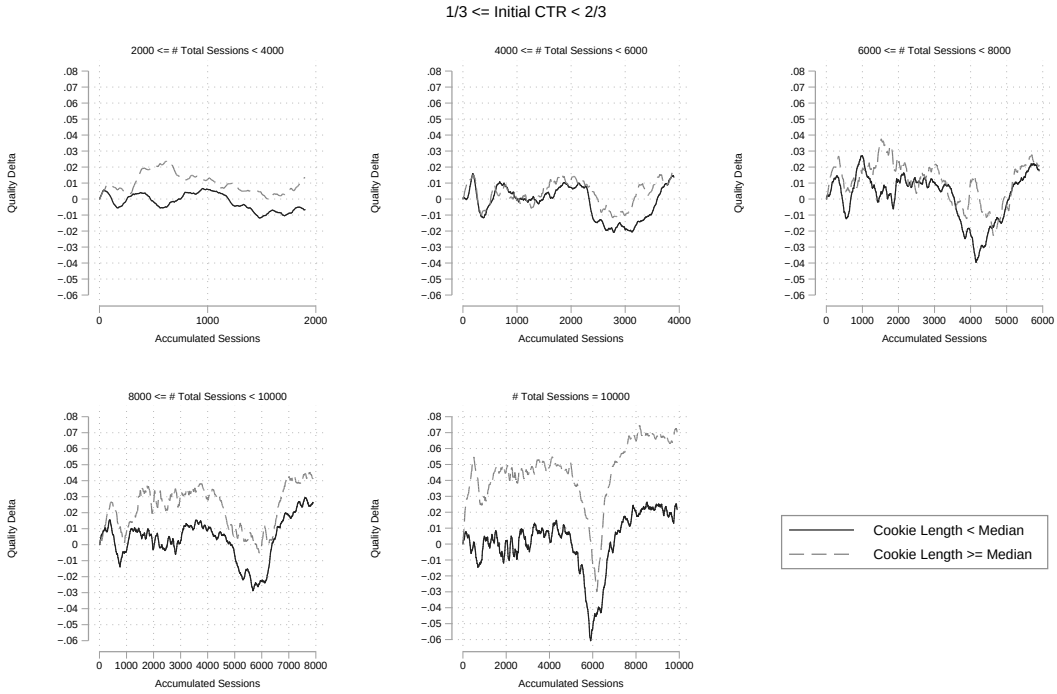
	(I)	(II)	(III)	(IV)
initial quality	9.417e-01***	9.389e-01***	9.395e-01***	9.379e-01***
	(3.052e-03)	(3.168e-03)	(3.132e-03)	(3.177e-03)
sessions	2.542e-06**	-6.260e-06**	-1.024e-05**	-1.517e-05***
	(8.455e-07)	(2.288e-06)	(3.424e-06)	(2.927e-06)
sessions x cookie (SC)	1.115e-06***	2.919e-06***	4.067e-06***	5.072e-06***
	(1.353e-07)	(4.345e-07)	(8.300e-07)	(6.967e-07)
sessions x total sessions	-3.686e-11	8.793e-10***	-8.492e-11	4.582e-10
	(1.020e-10)	(2.644e-10)	(1.019e-10)	(2.993e-10)
sessions x initial quality	-1.080e-05***	-1.046e-05***	7.555e-06	7.365e-06
	(1.132e-06)	(1.141e-06)	(4.113e-06)	(4.205e-06)
SC x total_sessions		-1.904e-10***		-1.127e-10*
		(4.802e-11)		(5.506e-11)
SC x initial quality			-3.893e-06***	-3.810e-06***
			(9.696e-07)	(1.002e-06)
Constant	3.998e-02***	4.174e-02***	4.128e-02***	4.229e-02***
	(2.083e-03)	(2.158e-03)	(2.149e-03)	(2.176e-03)
Observations	5066880	5066880	5066880	5066880
Adjusted R^2	0.883	0.883	0.883	0.883

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B.2 Descriptive analysis – personalized information





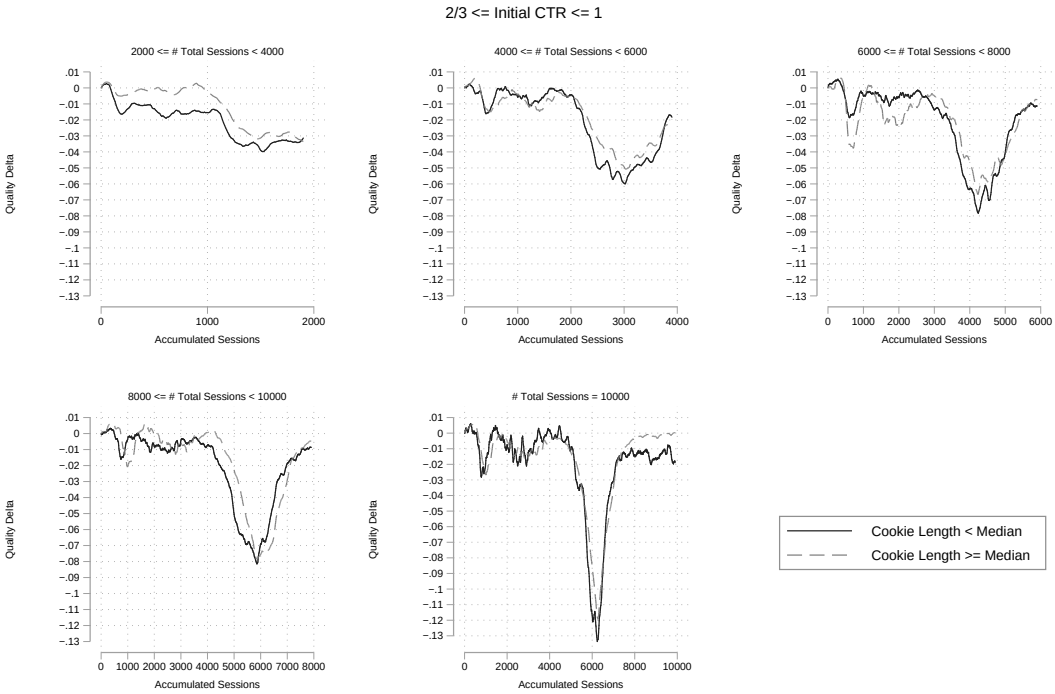


Figure 10 – Average quality evolution of queries with an average cookie history length above the median vs. quality evolution of queries with an average cookie history length below the median. Each panel displays the results for a specific initial quality category. Each figure within a panel displays the results for a specific quantity group. There are 5 quantity groups: queries with total searches $\in$ (i) $[2000, 4000)$; (ii) $[4000, 6000)$; (iii) $[6000, 8000)$; (iv) $[8000, 10000)$; and (v) $\{10000\}$. The lines display simple averages based on the rolling window values of the quality measure. The lines expand until the left edge of each quantity group to prevent attrition. The median value of the average cookie length is 4.48. The queries considered are the same as the ones used in the main analysis of Subsection 6.1.

B.3 Appendix Long Run Considerations

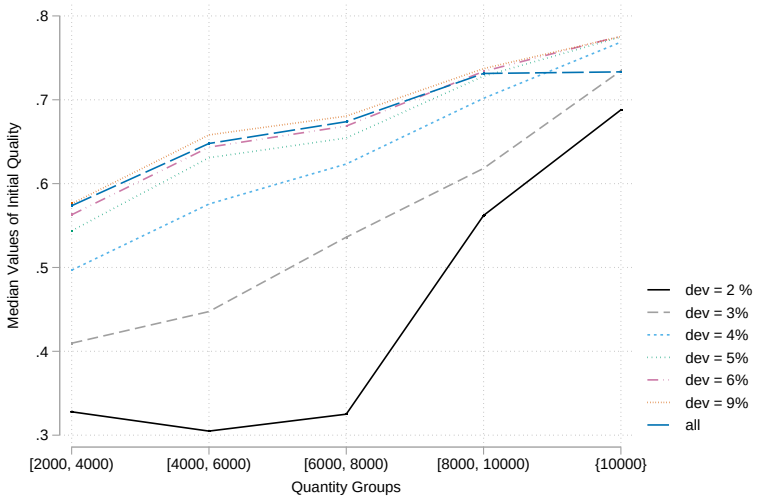


Figure 11 – Relationship between the median initial quality and the number of total searches. Each median value is calculated for the queries falling within a specific quantity group, as specified on the x-axis. The different constant frequency criterions are noted in the legends. The lines for a deviation of no more than 7% and 8% are omitted for ease of exposition; these do not differ substantially from the lines for a deviation of no more than 9%.

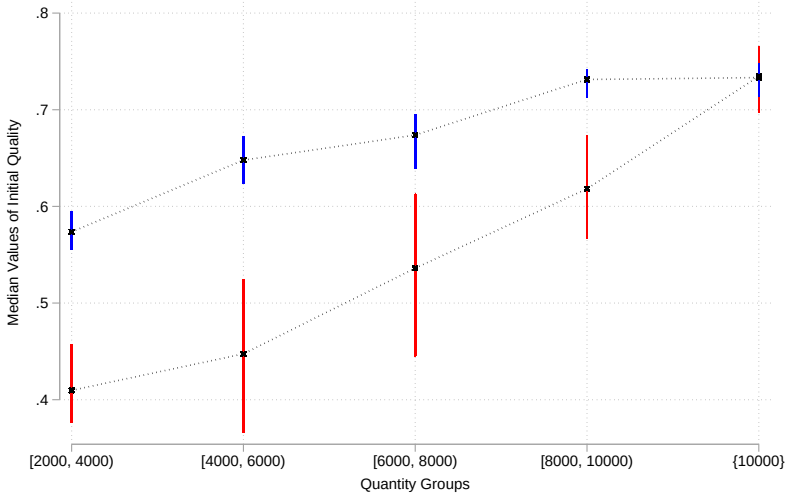


Figure 12 – Relationship between the median initial quality and the number of total searches. Each median value is calculated for the queries falling within a specific quantity group, as specified on the x-axis. The red vertical lines denote the 95% confidence interval for queries deviating no more than 3% from the constant frequency requirement. The blue vertical lines denote the confidence interval of the medians calculated for all queries with more than 2000 total searches.

Appendix C Robustness checks based on the editorial quality measure

Our data set comes with 659,000 editorial quality judgments collected from human experts on query-url pairs. The editorial quality judgments assess the relevance of a url for a specific query by a numerical grade ranging from 0 (not at all relevant) to 4 (highly relevant). By aggregating the editorial quality judgments of multiple urls displayed on the first result page, it is possible to obtain an overall "grade" for the quality of the result page.

As it requires considerable human resources, collecting editorial quality judgments is a very expensive process. The information retrieval literature on the advantages and disadvantages of editorial based quality measures is abundant and it is beyond the scope of this appendix to provide an in depth discussion or review of it. Our approach is to provide a very brief description of the editorial quality measure and, subsequently, to repeat the analysis provided in the paper based on the editorial quality measure.

C.1 The Editorial Quality Measure

The discounted cumulative gain is a widely used quality measure in the information retrieval (IR) literature. It is based on editorial quality judgments of query-url pairs based on the assessment of human experts. For example, assume that a specific url that was shown as a consequence of a specific query was rated with a relevance grade of 4 by a human expert. Furthermore, assume that this url was shown on position 2 of the corresponding result page. Then, we say that the **discounted gain** of this url with relevance (rel) 4 on position $j = 2$ is given by:

$$DG = \frac{2^{rel_j} - 1}{\log_2(j+1)} = \frac{2^4 - 1}{\log_2(2+1)}, \tag{11}$$

The numerator captures the informational "gain" that the searcher obtains by being provided this url. The denominator discounts for the fact that the url is displayed in position 2: The searcher had to "scan" through the search result page to be provided with this url. Note that by applying the logarithm of base 2 to the denominator, the gain of a document displayed on the first position is not discounted. To assess the quality of the entire result page, one possibility is to add up the discounted gain of all the documents displayed on the first result page. Assume for convenience that all 10 documents on the first result page are assigned a relevance judgment, then the discounted cumulative gain is given by:

$$DCGp = \sum_{j=1}^{p=10} \frac{2^{rel_j} - 1}{\log_2(j+1)}, \tag{12}$$

Two criteria determine the value of the DCG: (I) the general relevance of the documents available on the result page and (II) the ranking of the documents. (I) simply captures the idea that providing documents with relevant content is generally desirable (i.e. a lot of documents rated 4 are better than a lot of documents rated 1). (II) captures the idea that, given a specific set of documents with a given relevance, it is desirable to display the most relevant documents at the top of the result page (the ordering 4,3,2 is better than the ordering 2,3,4). The DCG captures both dimensions.

Obviously, to be able to compute the DCG for the entire result page, we need relevance judgments for all the urls presented on the result page. This is only very rarely the case in our dataset. Furthermore, in order to be able to compare different result pages based on the DCG criteria , we need to restrict our attention to result pages where the same number of consecutive urls come with relevance judgments. Thus, in the remainder of this appendix, we focus on result pages where the first 3 urls come with a relevance judgment. Each url has 5 possible relevance judgments (from 0 to 5). Hence, there are $5^3 = 125$ different possible combinations of relevance judgments.

We decided to restrict our attention on result pages with 3 consecutive relevance judgments for the first 3 urls to obtain a quality measure with greater depth than the one we use for our main analysis (which has a depth of one). We stopped at a depth of 3 in order to not lose too many result pages: compared to our initial sample of $\sim$ 80 Million result pages, we are left with $\sim$ 29 Million result pages when using our DCG based quality measure. Furthermore, a depth of 3 seems reasonable because, conditional on seeing a click on a generic url, $\sim$ 90% of all clicks are performed on one of the first three urls (see figure 8).

Each of the 125 possible combinations of relevance judgments is associated with a DCG-value. This measure is convex in the sense that if we order the 125 DCG values from lowest to highest, we

obtain a convex shaped curve. For example, the absolute increase in DCG moving from a result list with the grades $1, 1, 1$ to $2, 1, 1$ is smaller than the absolute increase in DCG for moving from $3, 3, 3$ to $4, 3, 3$. This is due to the fact that the relevance judgment enters the DCG-formula exponentially. For the purpose of our analysis, we find this property undesirable. In the above example, it is not clear which kind of change is valued more by the consumer. Instead of relying directly on the DCG for our quality measure, we rely on the ordering dictated by the DCG. Therefore, our quality measure ranges from 1 to 125. If two result combination result in a tie with respect to the DCG measure, they are also tied with respect to our quality measure. Furthermore, we normalize our quality measure to be between 0 and 1. For convenience, we label the quality measure "rank," reminiscent of the fact that our result pages can be ranked from 1 to 125 according to our quality measure.

C.2 Descriptive Analysis

Figures 13 and 14 reproduce the results from Figures 2 and 3 of the descriptive part of our main analysis. As can be seen, the findings remain similar. The initial quality level increases as a function of the total searches that we measure in-sample (Figure 13). Furthermore, the quality evolution that we measure in-sample decreases with the initial quality level (Figure 14).

As opposed to the results of our main analysis, in Figure 13 the average quality is slightly below the average *initial* quality. As can be seen from Figure 14, this is due to the fact that queries starting from a high initial quality level seem to experience a slight decrease in quality. Furthermore, the negative relationship between the initial quality and the in-sample quality evolution is not as clearly pronounced as for our click based quality measure. The two lowest initial quality groups experience roughly the same increase in quality, with the middle initial quality group even performing slightly better than the lowest initial quality group.

Interestingly, the average daily quality as measured by the rank does not pick up the dip in quality that we observe with the click based quality measure on the 20th of July. This is not due to the fact that we observe fewer result pages with 3 consecutive relevance judgments for the first 3 urls, as one might suspect. A closer analysis also reveals that the group of queries for which we observe 3 consecutive ranks (hence for which we can calculate our rank quality measure) experience a sharp decrease in quality as measured by our click based quality measure. This finding is troubling as it suggests that the DCG does not pick up a well documented event that resulted in a decrease in quality measured by different quality indicators: The number of searches for which we record no click at all experiences a 5 fold increase compared to its long term average. The average click though rate on the first 3 urls also decreases by roughly 15 percentage points compared to its long term average.

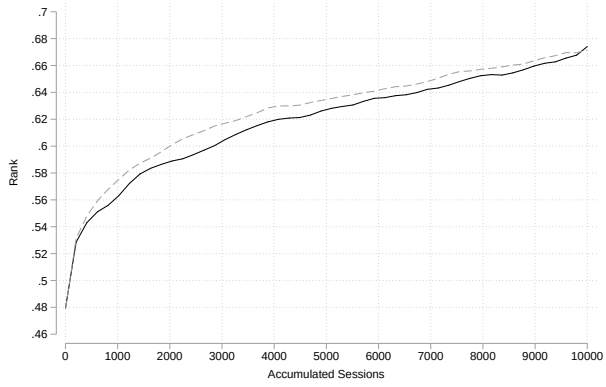


Figure 13 – Local Polynomial regression smooth of the average rank and average initial rank as a function of the observed sessions. The solid line refers to the average rank, the dashed line refers to the average initial rank

Figure 14 – In sample average quality evolution of the most popular queries for different initial quality categories. Solid line: $\overline{Irank} \in [0, \frac{1}{3})$. Dashed line: $\overline{Irank} \in [\frac{1}{3}, \frac{2}{3})$. Dash-dotted light blue line: $\overline{Irank} \in [\frac{2}{3}, 1]$. Calculations are based on the rolling window averages

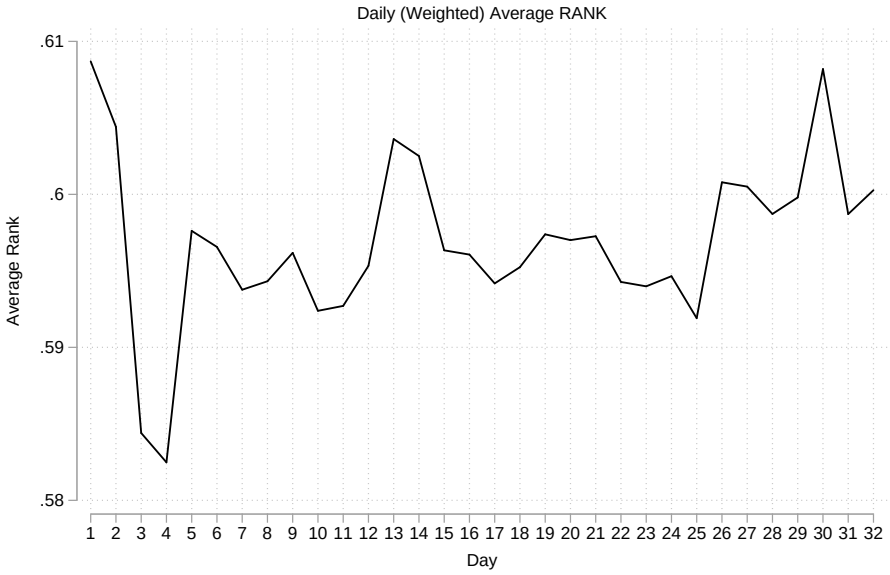


Figure 15 – Daily weighted average of the Rank, weights are calculated based on the daily searches for which the editorial quality measure can be computed, i.e. for result pages where the first 3 urls are rated

C.3 In-sample Analysis

Table 1 displays the regression results based on the rank quality measure. As for the analysis based on the click based quality measure, the dependent variable is calculated based on rolling window averages for 100 consecutive searches. Figure 16 displays the resulting fit for specification (IV). The results confirm that our in-sample findings are robust to the rank quality measure. The impact of the cookie length variable is positive across all specifications. For specification (II) the coefficient is not significant. As explained in the corresponding section of the paper, it is reasonable to interact the cookie length variable with the initial quality variable to account for the fact that queries starting from a higher initial quality level have less scope for quality improvement. For the specifications that include this interaction, we find a significant and positive effect of the cookie length variable.

The left column pair of Table 2 presents the estimated differences (and p-values) in quality evolution between queries we observe 10000 times and queries we observe 2000 times when the average cookie length is equal to 6. Each row stands for a different initial quality level under one of the 4 specifications (I) -(IV) (from top to bottom). The middle column pair repeats the same analysis for queries with an average cookie length of 3.5. The right column pair presents the estimates differences between the point estimates of the left and middle column along with the p-values. As for our main results, we find that a longer average cookie history reinforces economies of scale from data. For low initial quality levels, the impact is significant across all specification except (II).

Table 8 – Results for Specifications (I) - (IV) for editorial quality Judgments

	(1) (I)	(2) (II)	(3) (III)	(4) (IV)
initial quality	9.755e-01***	9.751e-01***	9.741e-01***	9.741e-01***
	(2.847e-03)	(2.912e-03)	(2.927e-03)	(2.938e-03)
sessions	1.450e-06	-8.100e-07	-4.085e-06	-3.679e-06
	(1.050e-06)	(2.984e-06)	(2.595e-06)	(3.016e-06)
sessions x cookie (SC)	4.811e-07***	9.750e-07	1.641e-06**	1.552e-06*
	(1.354e-07)	(6.001e-07)	(5.451e-07)	(6.206e-07)
sessions x total sessions	8.013e-11	3.140e-10	7.665e-11	3.172e-11
	(1.199e-10)	(3.281e-10)	(1.196e-10)	(3.658e-10)
sessions x initial quality	-6.372e-06***	-6.312e-06***	1.811e-06	1.840e-06
	(1.008e-06)	(1.020e-06)	(2.886e-06)	(2.990e-06)
SC x total_sessions		-5.147e-11		9.884e-12
		(6.560e-11)		(7.541e-11)
SC x initial quality			-1.625e-06*	-1.633e-06*
			(6.326e-07)	(6.637e-07)
Constant	1.050e-02***	1.075e-02***	1.140e-02***	1.136e-02***
	(2.008e-03)	(2.039e-03)	(2.082e-03)	(2.074e-03)
Observations	2707043	2707043	2707043	2707043
Adjusted R^2	0.928	0.928	0.928	0.928

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9 – Differences and Differences in Differences

	CL = 6		CL = 3.5		Diff	
Initial quality	Estimate	P-val	Estimate	P-val	Estimate	P-val
Specification (I)						
.25	0.030***	0.000	0.020***	0.000	0.010***	0.000
.5	0.017***	0.000	0.007	0.078	0.010***	0.000
.75	0.004	0.170	-0.005	0.188	0.010***	0.000
Specification (II)						
.25	0.028***	0.000	0.021***	0.000	0.007	0.134
.5	0.016***	0.000	0.008	0.067	0.007	0.134
.75	0.003	0.382	-0.004	0.380	0.007	0.134
Specification (III)						
.25	0.038***	0.000	0.013*	0.017	0.025**	0.002
.5	0.022***	0.000	0.005	0.229	0.017**	0.001
.75	0.006	0.071	-0.003	0.503	0.008***	0.000
Specification (IV)						
.25	0.038***	0.000	0.013*	0.039	0.025*	0.017
.5	0.022***	0.000	0.005	0.327	0.017*	0.025
.75	0.006	0.122	-0.003	0.527	0.009	0.077

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

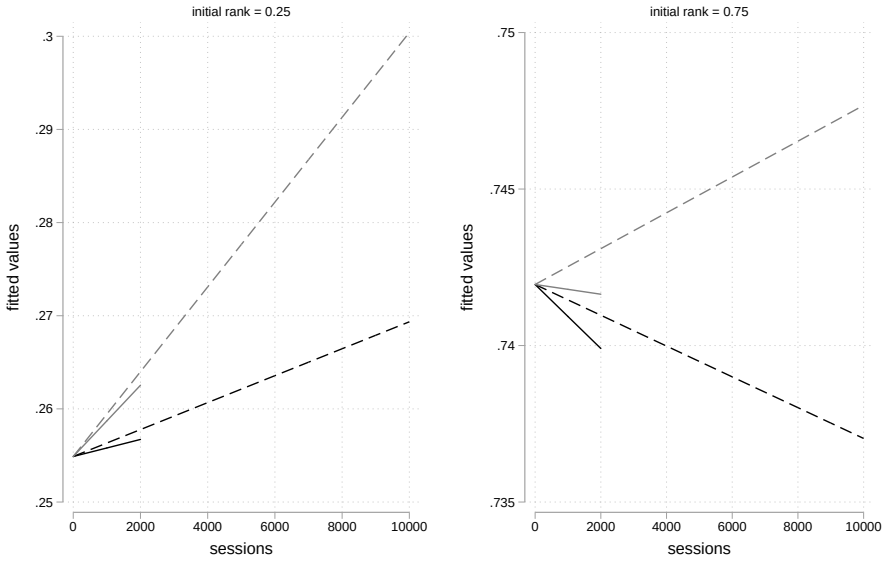


Figure 16 – Predicted quality evolution of model (IV). The left panel refers to queries with an initial CTR of 25%, the right panel to queries with an initial CTR of 75%. The black lines refer to queries with an average cookie length of 3.5, the gray lines to queries with an average cookie length of 6. The total number of searches of the queries is displayed on the x-axis.

C.4 Long Run Considerations

An analysis based on the concepts developed in the corresponding section of the main paper delivers similar result. Here, the use of the editorial based quality measure also does not substantially alter our results. The results from Figure 17 are not as clear as the corresponding results based on the click based quality measure. Figure 18 suggests that this might be mainly due to composition effects. Once the average cookie length is taken into account, the result based on the rank quality measure closely match the results based on the click based quality measure: Focusing on queries for which the in-sample popularity is likely to be a good proxy for the pre-sample popularity reveals a stronger relationship between the in-sample searches and the initial quality measure.

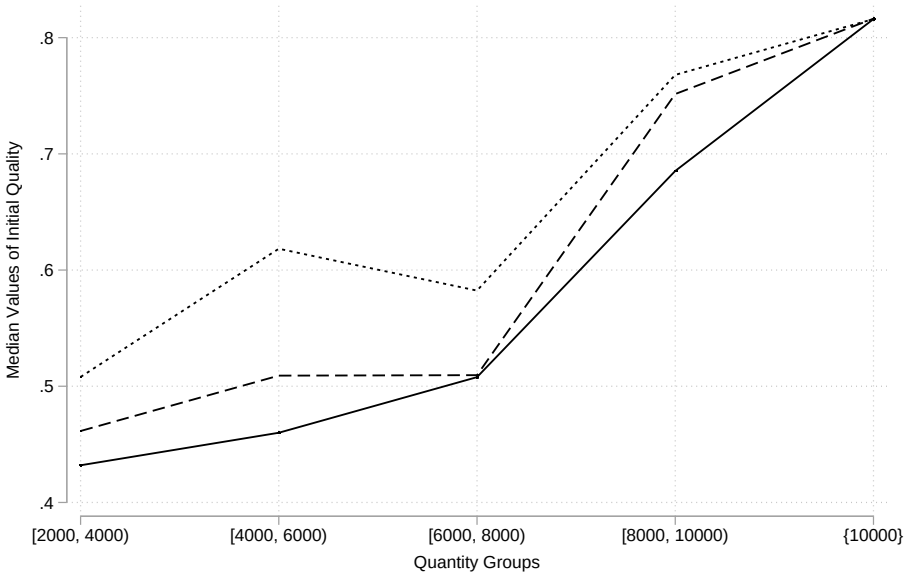


Figure 17 – Relationship between the median initial quality and the number of total searches. Each median value is calculated for the queries falling within a specific quantity group as specified on the x-axis. The dotted line displays the median values calculated for all queries. The dashed line was calculated based on queries that deviate no more than 4 percentage points from our constant frequency definition. The solid line is calculated based on queries that deviate no more than 3 percentage points from our constant frequency definition.

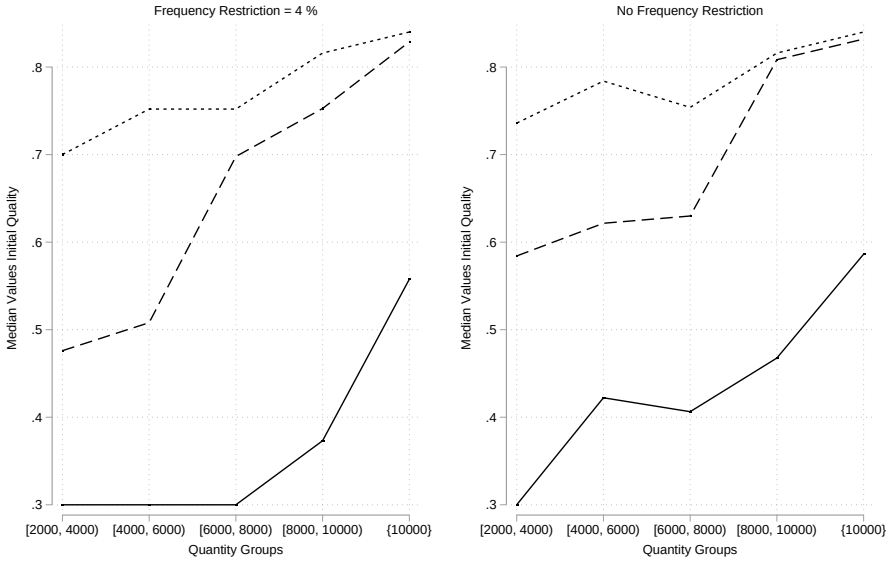


Figure 18 – Relationship between the median initial quality and the number of total searches. The dotted lines refer to queries with an average cookie length exceeding the 66% percentile. The dashed lines refer to queries with an average cookie length between the 33% and 66% percentile. The solid lines refer to queries below the 33% percentile. The left panel displays the result for the queries which deviate no more than 3% from our constant frequency definition. The right panel displays the results for all the queries in the sample (no restriction on the search frequency).