# Wikipedia usage survey results

Vipul Naik

2016-12-25

## Contents

# Summary

In 2016, Issa Rice and I conducted several surveys of Wikipedia usage. We collected survey responses from Slate Star Codex readers, Vipul's Facebook friends, and a few United States audiences through SurveyMonkey Audience and Google Surveys (known at the time as Google Consumer Surveys). Our survey questions measured how heavily people use Wikipedia, what sort of pages they read or expected to find, the relation between their search habits and Wikipedia, and other actions they took within Wikipedia.

The surveys are part of our work to understand the impact of contributing to

Wikipedia. Both of us regularly contribute to the site, and we are also getting more people to work on editing and adding content to Wikipedia. Therefore we want to understand how people use Wikipedia, how much they use Wikipedia, what types of people tend to use Wikipedia, and so on, so that we can direct efforts more strategically.

Our three main takeaways:

- Wikipedia consumption of heavily skewed toward a profile of "elite" people, and these people use the site in qualitatively different ways. (More)
- As a result, we've revised upward our estimate of the impact per pageview, and revised downward our estimate of the broad appeal and reach of Wikipedia. (More)
- The gap between elite samples of Wikipedia users and general United States Internet users is significantly greater than the gap between the different demographics within the United States that we measured.
  It is comparable to the gap between United States Internet users and Internet users in low-income countries. (More)

This post goes over the survey questions, the responses of participants, and other survey data (specifically, data from the New Readers surveys by the Wikimedia Foundation) and then explains the takeaways.

# Surveys

## First SurveyMonkey survey (S1)

At the end of May 2016, Vipul Naik[1] and I created a [Wikipedia usage survey](https://meta.wikimedia.org/wiki/Research:Wikipedia_usage_survey 'Vipul Naik and Issa Rice. "Research:Wikipedia usage survey". May 31, 2016. Wikimedia Meta Wiki.') on SurveyMonkey to gauge the usage habits of Wikipedia readers and editors.

### Audiences for S1

SurveyMonkey[2] allows the use of different "collectors" (i.e. survey URLs that keep results separate), so we circulated several different URLs among four locations to see how different audiences would respond.

The audiences were as follows:

[1] http://lesswrong.com/user/VipulNaik/overview/
[2] https://en.wikipedia.org/wiki/SurveyMonkey

- SurveyMonkey's United States audience with no demographic filters (62 responses, 54 of which are full responses). We will refer to this audience as SM, or S1SM if needed to avoid ambiguity.

  Acquisition cost: $100 ($2 per response for 50 responses, 4 extra responses given gratis)

- My Facebook timeline (post asking people to take the survey[3]; 70 responses, 69 of which are full responses). For background on the timeline audience, see my page on how I use Facebook[4]. We will refer to this audience as "Vipul's Facebook friends" or V for short.

  Acquisition cost: None

- The Wikipedia Analytics mailing list[5] (email linking to the survey[6]; 7 responses, 6 of which are full responses). Note that due to the small size of this group, the results below should not be trusted, unless possibly when the votes are decisive. We will refer to this audience as AM.

  Acquisition cost: None

- Slate Star Codex (post that links to the survey[7]; 618 responses, 596 of which are full responses). We will refer to this audience as SSC.

  While Slate Star Codex isn't the same as LessWrong, we think there is significant overlap in the two sites' audiences (see e.g. the recent LessWrong diaspora survey results[8]).

  Acquisition cost: None

- In addition, although not an actual audience with a separate URL, several of the tables we present below will include an "H" group; this is the heavy users group of people who responded by saying they read 26 or more articles per week on Wikipedia. This group has 179 people: 164 from Slate Star Codex, 11 from Vipul's timeline, and 4 from the Analytics mailing list.

We ran the survey from May 30 to July 9, 2016 (although only the Slate Star Codex survey had a response past June 1).

### Questions for S1

For reference, here are the survey questions for the first survey. A dummy/mock-up version of the survey can be found here: https://www.surveymonkey.com/

[3] https://www.facebook.com/vipulnaik.r/posts/10208540131276697
[4] http://vipulnaik.com/facebook/
[5] https://lists.wikimedia.org/mailman/listinfo/analytics
[6] https://lists.wikimedia.org/pipermail/analytics/2016-May/005219.html
[7] http://slatestarcodex.com/2016/06/02/links-616-linkandescence/
[8] http://lesswrong.com/lw/nor/2016_lesswrong_diaspora_survey_analysis_part_two/

r/PDTTBM8.

The survey introduction said the following:

> This survey is intended to gauge Wikipedia use habits. This survey has 3 pages with 5 questions total (3 on the first page, 1 on the second page, 1 on the third page). Please try your best to answer all of the questions, and make a guess if you're not sure.

And the actual questions:

1. How many distinct Wikipedia pages do you read per week on average?
    - less than 1
    - 1 to 10
    - 11 to 25
    - 26 or more
2. On a search engine (e.g. Google) results page, do you explicitly seek Wikipedia pages, or do you passively click on Wikipedia pages only if they show up at the top of the results?
    - I explicitly seek Wikipedia pages
    - I have a slight preference for Wikipedia pages
    - I just click on what is at the top of the results
3. Do you usually read a particular section of a page or the whole article? (Multiple options can be selected)
    - Particular section
    - Whole page
4. How often do you do the following? (Choices: Several times per week, About once per week, About once per month, About once per several months, Never/almost never.)
    - Use the search functionality on Wikipedia
    - Be surprised that there is no Wikipedia page on a topic
5. For what fraction of pages you read do you do the following? (Choices: For every page, For most pages, For some pages, For very few pages, Never. These were displayed in a random order for each respondent, but displayed in alphabetical order here.)
    - Check (click or hover over) at least one citation to see where the information comes from on a page you are reading
    - Check how many pageviews a page is getting (on an external site or through the Pageview API)
    - Click through/look for at least one cited source to verify the information on a page you are reading
    - Edit a page you are reading because of grammatical/typographical errors on the page

- Edit a page you are reading to add new information
- Look at the "See also" section for additional articles to read
- Look at the editing history of a page you are reading
- Look at the editing history solely to see if a particular user wrote the page
- Look at the talk page of a page you are reading
- Read a page mostly for the "Criticisms" or "Reception" (or similar) section, to understand different views on the subject
- Share the page with a friend/acquaintance/coworker

For the SurveyMonkey audience, there were also some demographic questions (age, gender, household income, US region, and device type). These questions were not filled by respondents at the time of the survey, but rather, are filled in by respondents in order to be able to participate in these surveys. You can learn more on the SurveyMonkey Contribute page[9].

## Second SurveyMonkey survey (S2)

After we looked at the survey responses on the first day, Vipul and I decided to create a second survey to focus on the parts from the first survey that interested us the most.

### Audiences for S2

The second survey was only circulated among SurveyMonkey's audiences:

- SurveyMonkey's US audience with no demographic filters (54 responses).

  Acquisition cost: $50 ($1 per response for 50 responses, 4 extra responses given gratis)

- SurveyMonkey's US audience with the following filters: ages 18–29 with a college or graduate degree (50 responses).

  Acquisition cost: $125 ($2.50 per response for 50 responses)

We first ran the survey on the unfiltered audience again because the wording of our first question was changed and we wanted to have the new baseline. We then chose to filter for young college-educated people because our prediction was that more educated people would be more likely to read Wikipedia. The SurveyMonkey demographic data does not include education, and we hadn't seen the Pew Internet Research surveys in the next section, so we were relying on our intuition and some demographic data from past surveys) for the "college-educated" part. Our selection of the age group was based on the fact that young

[9] https://contribute.surveymonkey.com/home

people in our first survey gave more informative free-form responses in survey 2 (SurveyMonkey's demographic data *does* include age).

**Questions for S2**

For reference, here are the survey questions for the second survey. A dummy/mock-up version of the survey can be found here: https://www.surveymonkey.com/r/28BW78V.

The survey introduction said the following:

> This survey is intended to gauge Wikipedia use habits. Please try your best to answer all of the questions, and make a guess if you're not sure.
>
> This survey has 4 questions across 3 pages.
>
> In this survey, "Wikipedia page" refers to a Wikipedia page in any language (not just the English Wikipedia).

And the actual questions:

1. How many distinct Wikipedia pages do you read (at least one sentence of) per week on average?
    - Fewer than 1
    - 1 to 10
    - 11 to 25
    - 26 or more
2. Which of these articles have you read (at least one sentence of) on Wikipedia (select all that apply)? (These were displayed in a random order except the last option for each respondent, but displayed in alphabetical order except the last option here.)
    - Adele
    - Barack Obama
    - Bernie Sanders
    - China
    - Donald Trump
    - Google
    - Hillary Clinton
    - India
    - Japan
    - Justin Bieber
    - Justin Trudeau
    - Katy Perry
    - Taylor Swift
    - The Beatles

- United States
- World War II
- None of the above

3. What are some of the Wikipedia articles you have most recently read (at least one sentence of)? Feel free to consult your browser's history.
4. Recall a time when you were surprised that a topic did not have a Wikipedia page. What were some of these topics?

As with the SurveyMonkey Audience responses for S1, the responses for S2 also came with demographic information that the respondents had previously filled in.

## Google Surveys survey (GS)

We ran a third survey on Google Surveys (known at the time as Google Consumer Surveys) with a *single question* that was a word-to-word replica of the first question from the second survey. The main motivation here was that on Google Surveys, a single-question survey costs only 10 cents per response, so it was possible to get to a large number of responses at relatively low cost, and achieve more confidence in the tentative conclusions we had drawn from the SurveyMonkey surveys.

### Audiences for GS

We bought 500 responses at 10 cents per response, for a total acquisition cost of $50. The responses were from a general United States audience.

GS uses a "surveywall" methodology to collect survey responses: the survey questions are shown to people who want to access a piece of content (article or video) and they need to answer the question to access it.

Overall, Google Surveys in the United States is reasonably close to representative of the voting US population and the Internet-using population. Also, the sample size of the survey was largest. Therefore, among the surveys we did, this survey comes closest to approximating the behavior of the Internet-using population in the United States.

You can learn more at the Wikipedia page for Google Surveys[10].

### Questions for GS

This survey had exactly one question. The wording of the question was exactly the same as that of the first question of the second survey.

[10] https://en.wikipedia.org/wiki/Google_Surveys

1. How many distinct Wikipedia pages do you read (at least one sentence of) per week on average?
    - Fewer than 1
    - 1 to 10
    - 11 to 25
    - 26 or more

One slight difference was that whereas in the second SurveyMonkey survey, the order of the options was fixed, the Google Surveys survey did a 50/50 split between that order and the exact reverse order. Such splitting is a best practice to deal with any order-related biases, while still preserving the logical order of the options.

You can read more on the questionnaire design page of the Pew Research Center[11].

The GS responses come with inferred demographic and geographic data (age, gender, income level, location). The geographic data is generally reliable because it is based on IP address, but inferred age and gender data is not as reliable as the self-reported data that we get from SurveyMonkey Audience. For more on the accuracy of the inferred data, see the Pew Research Center's comparison[12].

## Wikimedia Foundation New Readers survey (NR)

In late 2016, the Wikimedia Foundation's Global Reach team published the results of New Readers phone surveys[13]. The questions in these surveys have some overlap with the questions in our surveys, so we have updated our post to include a discussion of these surveys and how the results compare with ours.

### Audiences for NR

The NR surveys were conducted in the following five countries: Nigeria[14], India[15], Mexico[16], Brazil[17], and Egypt[18]. The surveys were conducted by phone.

For the first three countries (Nigeria, India, and Mexico), results of additional in-person surveys have also been published.

[11] http://www.pewresearch.org/methodology/u-s-survey-research/questionnaire-design/
[12] http://www.people-press.org/2012/11/07/a-comparison-of-results-from-surveys-by-the-pew-research-center-and-google-consumer-surveys/
[13] https://meta.wikimedia.org/wiki/Global_Reach/Insights
[14] https://meta.wikimedia.org/wiki/Global_Reach/Nigeria_survey
[15] https://meta.wikimedia.org/wiki/Global_Reach/India_survey
[16] https://meta.wikimedia.org/wiki/Global_Reach/Mexico_Survey
[17] https://meta.wikimedia.org/wiki/Global_Reach/Brazil_survey
[18] https://meta.wikimedia.org/wiki/Global_Reach/Egypt_survey

**Questions for NR**

We will compare the results of our surveys with the results of the New Readers surveys. To shed light on this comparison, we include below the list of questions in the New Readers phone survey.

Not all questions were presented in all surveys. The Egypt survey[19], which is the more recent, had the longest list of questions, and we provide this list below. The numbering is mostly based on the Egypt survey, though off by one for later questions due to a question missing from the Egypt survey.

Our later analysis will focus on the first, fourth, and seventh question, which are together comparable against the first question of S1, S2, and GS.

1. Do you use the Internet?
    - Yes
    - Said no, but uses Facebook
    - No
2. What do you use the Internet for the most? (for those who said Yes to Q1)
    - Look up info
    - Social media
    - Entertainment
    - News
    - Others
3. What's the biggest reason you don't use the Internet? (for those who said No to Q1)
    - Too expensive
    - Not sure it's useful
    - Not sure what it is
    - Other
4. Have you ever heard of Wikipedia?
    - Yes
    - No
5. Where did you find out about Wikipedia?
    - Internet
    - School
    - Friends and family
    - Radio or TV
    - Not sure
6. What do you use Wikipedia for?

[19] https://meta.wikimedia.org/wiki/Global_Reach/Egypt_survey

- School
- Work
- Entertainment
- Other

7. How often do you use Wikipedia?
    - Daily
    - Weekly
    - Monthly
    - Rarely
    - Never
8. How interested are you in reading Wikipedia? (for those who answered "Rarely" or "Never" to the previous question)
    - Not interested
    - Somewhat
    - Very interested
9. What's the largest barrier keeping you from reding Wikipedia? (for those who answere "Very interested" to Q8)
    - Don't trust content
    - Expensive data
    - Not interesting enough
    - Can't find it
    - Other
10. What would make you more likely to use Wikipedia? (for those who answered "Not interested" to Q8)
    - Trusted the content
    - Cheaper data
    - More interesting articles
    - Known how to find it
    - None
11. Do you have a mobile phone? (This question was in some other country surveys though not in the Egypt one. Hence the numbering for later questions is one more than the numbering in the actual Egypt survey)
    - Yes
    - No
12. Can you use the Internet with your phone?
    - Yes
    - No
13. How do you access the Internet on your phone?
    - Cellular

- Wifi and cell
- Wifi only
- No Internet
- Not sure

14. What is your usual network speed?
    - 2G / Edge
    - 3G
    - Better than 3G
    - Not sure
15. Do you download and use Apps?
    - Yes
    - No
16. What is your gender?
    - Male
    - Female
17. What is your age?
    - Under 18
    - 19–31
    - 31–50
    - over 50
    - Prefer not to say
18. What is your location?
    - Urban
    - Rural
    - Not sure
19. What is your geographical zone? (options specific to Egypt)

## Other surveys

Several demographic surveys regarding Wikipedia have been conducted, targeting both editors and users. The surveys we found most helpful were the following:

- The 2010 Wikipedia survey[20] by the Collaborative Creativity Group and the Wikimedia Foundation. The explanation before the bottom table on

[20] https://web.archive.org/web/20130717211630/http://wikipediastudy.org/

page 7 of the overview PDF[21] has "Contributors show slightly but significantly higher education levels than readers", which provides weak evidence that more educated people are more likely to engage with Wikipedia.
- The Global South User Survey 2014[22] by the Wikimedia Foundation
- Pew Internet Research's 2011 survey[23]: "Education level continues to be the strongest predictor of Wikipedia use. The collaborative encyclopedia is most popular among internet users with at least a college degree, 69% of whom use the site." (page 3)
- Pew Internet Research's 2007 survey[24].

There is also the New Readers survey mentioned earlier, that we examine in detail in this post.

## Motivation

Vipul and I ultimately want to get a better sense of the value of a Wikipedia pageview (one way to measure the impact of content creation), and one way to do this is to understand how people are using Wikipedia. As we focus on getting more people to work on editing Wikipedia[25] – thus causing more people to read the content we pay and help to create – it becomes more important to understand who is reading the content, and how they engage with it.

For some previous discussion, see also Vipul's answers to the following Quora questions:

- What are the various parameters that affect the value of a pageview?[26]
- What's the relative social value of 1 Quora pageview (as measured by Quora stats http://www.quora.com/stats) and 1 Wikipedia pageview (as measured at, say, Wikipedia article traffic statistics)?[27]

Wikipedia allows relatively easy access to pageview data (especially by using tools developed for this purpose, including one that Vipul made[28]), and there are some surveys that provide demographic data (see "Previous surveys" above). However, after looking around, it was apparent that the kind of information

[21] https://web.archive.org/web/20131209060146/http://wikipediastudy.org/docs/Wikipedia_Overview_15March2010-FINAL.pdf

[22] https://upload.wikimedia.org/wikipedia/commons/8/8a/Global_South_User_Survey_2014_-_Full_Analysis_Report.pdf

[23] http://www.pewinternet.org/files/old-media/Files/Reports/2011/PIP_Wikipedia.pdf

[24] http://www.pewinternet.org/2007/04/24/wikipedia-users/

[25] https://contractwork.vipulnaik.com

[26] https://www.quora.com/What-are-the-various-parameters-that-affect-the-value-of-a-pageview/answer/Vipul-Naik

[27] https://www.quora.com/Whats-the-relative-social-value-of-1-Quora-pageview-as-measured-by-Quora-stats-http-www-quora-com-stats-and-1-Wikipedia-pageview-as-measured-at-say-Wikipedia-article-traffic-statistics/answer/Vipul-Naik

[28] https://wikipediaviews.org/

our survey was designed to find was not available. This was before the New Readers survey results had been published.

## Results

In this section we present the highlights from each of the survey questions. If you prefer to dig into the data yourself, there are also some exported PDFs below provided by SurveyMonkey. Most of the inferences can be made using these PDFs, but there are some cases where additional filters are needed to deduce certain percentages.

For the SurveyMonkey surveys, we use the notation "S*n*Q*m*" to mean "survey *n* question *m*". The Google Surveys survey question is referred to as GS, and the New Readers questions are referred to with the notation "NRQ*m*" for question *m* of the survey.

### S1Q1: number of Wikipedia pages read per week

Here is a table that summarizes the data for Q1. Note that SMM and SMF don't add up to SM as some respondents did not specify their gender.

Table 1: How many distinct Wikipedia pages do you read per week on average? SM = SurveyMonkey audience, V = Vipul Naik's timeline, SSC = Slate Star Codex audience, AM = Wikipedia Analytics mailing list, SMM = SurveyMonkey males, SMF = SurveyMonkey females.

| Response | SM (N=62) | V (N=70) | SSC (N=618) | AM (N=7) | SMM (N=28) | SMF (N=26) |
|---|---|---|---|---|---|---|
| less than 1 | 42% | 1% | 1% | 0% | 25% | 58% |
| 1 to 10 | 45% | 40% | 37% | 29% | 46% | 42% |
| 11 to 25 | 13% | 43% | 36% | 14% | 29% | 0% |
| 26 or more | 0% | 16% | 27% | 57% | 0% | 0% |
| pgs/wk lower | 1.88 | 9.29 | 11.35 | 16.65 | 3.65 | 0.42 |
| pgs/wk upper | 8.17 | 22.76 | 26.21 | 34.90 | 12.10 | 4.70 |

The "pgs/wk lower" is obtained as the average pages read per week if everybody read at the lower end of their estimate (so the respective estimates are 0, 1, 11, and 26).

The "pgs/wk upper" is obtained as the average of pages read per week if everybody read at the upper end of their estimate, except the "26 or more" case where we assume a value of 50 (so the respective estimates are 1, 10, 25, and 50). We choose 50 as a reasonable upper bound on what the average person

who views more than 26 pages likely views, rather than a strict bound on every individual.

There are two reasons to compute the "pgs/wk lower" and "pgs/wk upper" numbers:

- Having these ranges makes it easier to quickly compare different audiences.
- The (very approximate) estimates of pages/week can be validated against known information about total pageviews.

The comments indicated that S1Q1 was flawed in several ways: we didn't specify which language Wikipedias count nor what it meant to "read" an article (the whole page, a section, or just a sentence?).

One comment questioned the "low" ceiling of 26; however, the actual distribution of responses suggests that the ceiling wasn't too low.

An interesting potential modification of the survey would be to ask further questions of people who selected an extreme response, to better bucket them.

## S1Q2: affinity for Wikipedia in search results

We asked Q2, "On a search engine (e.g. Google) results page, do you explicitly seek Wikipedia pages, or do you passively click on Wikipedia pages only if they show up at the top of the results?", to see to what extent people preferred Wikipedia in search results.

The main implication to this for people who do content creation on Wikipedia is that if people do explicitly seek Wikipedia pages (for whatever reason), it makes sense to give them more of what they want. On the other hand, if people don't prefer Wikipedia, it makes sense to update in favor of diversifying one's content creation efforts while still keeping in mind that raw pageviews indicate that content will be read more if placed on Wikipedia (see for instance Brian Tomasik's experience[29], which is similar to my own, or gwern's page comparing Wikipedia with other wikis[30]).

The following table summarizes our results. Wikipedia has been shortened to WP to conserve column width.

[29] http://reducing-suffering.org/the-value-of-wikipedia-contributions-in-social-sciences/#Readership

[30] http://www.gwern.net/Wikipedia%20and%20Other%20Wikis

Table 2: On a search engine (e.g. Google) results page, do you explicitly seek Wikipedia pages, or do you passively click on Wikipedia pages only if they show up at the top of the results? SM = SurveyMonkey audience, V = Vipul Naik's timeline, SSC = Slate Star Codex audience, AM = Wikipedia Analytics mailing list, H = heavy users (26 or more articles per week) of Wikipedia.

| Response | SM (N=62) | V (N=70) | SSC (N=618) | AM (N=7) | H (N=179) | SMM (N=28) |
|---|---|---|---|---|---|---|
| Explicitly seek WP | 19% | 60% | 63% | 57% | 79% | 25% |
| Slight preference for WP | 29% | 39% | 34% | 43% | 20% | 39% |
| Just click on top results | 52% | 1% | 3% | 0% | 1% | 35% |

An oversight on our part was not to include an option for people who avoided Wikipedia or did something else. This became apparent from the comments. For this reason, the "Just click on top results" options might be inflated. In addition, some comments indicated a mixed strategy of preferring Wikipedia for general overviews while avoiding it for specific inquiries, so allowing multiple selections might have been better for this question.

## S1Q3: section vs whole page

This question is relevant for us because the work we fund[31] is mainly whole-page creation. If people are mostly reading the introduction or a particular section like the "Criticisms" or "Reception" section (see S1Q5), then that forces us to consider spending more time on those sections, or to strengthen those sections on weak existing pages.

Responses to this question were fairly consistent across different audiences, as can be see in the following table.

Table 3: Do you usually read a particular section of a page or the whole article? SM = SurveyMonkey audience, V = Vipul Naik's timeline, SSC = Slate Star Codex audience, AM = Wikipedia Analytics mailing list, H = Heavy users (26 or more articles per week) of Wikipedia, SMM = SurveyMonkey males, SMF = SurveyMonkey females.

| Response | SM (N=62) | V (N=70) | SSC (N=618) | AM (N=7) | H (N=179) | SMM (N=28) | SMF (N=26) |
|---|---|---|---|---|---|---|---|
| Section | 73% | 80% | 74% | 86% | 70% | 68% | 73% |
| Whole | 34% | 23% | 33% | 29% | 37% | 39% | 31% |

[31] https://github.com/vipulnaik/contractwork/blob/master/new-article-pool.mediawiki

People were allowed to select more than one option for this question. The comments indicate that several people do a combination, where they read the introductory portion of an article, then narrow down to the section of their interest.

## S1Q4: search functionality on Wikipedia and surprise at lack of Wikipedia pages

We asked about whether people use the search functionality on Wikipedia because we wanted to know more about people's article discovery methods. The data is summarized in the following table.

Table 4: How often do you use the search functionality on Wikipedia? SM = SurveyMonkey audience, V = Vipul Naik's timeline, SSC = Slate Star Codex audience, AM = Wikipedia Analytics mailing list, H = heavy users (26 or more articles per week) of Wikipedia, SMM = SurveyMonkey males, SMF = SurveyMonkey females.

| Response | SM (N=62) | V (N=69) | SSC (N=613) | AM (N=7) | H (N=176) | SMM (N |
|---|---|---|---|---|---|---|
| Several times per week | 8% | 14% | 32% | 57% | 55% | 14% |
| About once per week | 19% | 17% | 21% | 14% | 15% | 21% |
| About once per month | 15% | 13% | 14% | 0% | 3% | 14% |
| About once per several months | 13% | 12% | 9% | 14% | 5% | 7% |
| Never/almost never | 45% | 43% | 24% | 14% | 23% | 43% |

Many people noted here that rather than using Wikipedia's search functionality, they use Google with "wiki" attached to their query, DuckDuckGo's "!w" expression, or some browser configuration to allow a quick search on Wikipedia.

To be more thorough about discovering people's content discovery methods, we should have asked about other methods as well. We did ask about the "See also" section in S1Q5.

Next, we asked how often people are surprised that there is no Wikipedia page on a topic to gauge to what extent people notice a "gap" between how Wikipedia exists today and how it *could* exist. We were curious about what articles people specifically found missing, so we followed up with S2Q4.

Table 5: How often are you surprised that there is no Wikipedia page on a topic? SM = SurveyMonkey audience, V = Vipul Naik's timeline, SSC = Slate Star Codex audience, AM = Wikipedia Analytics mailing list, H = heavy users (26 or more articles per week) of Wikipedia, SMM = SurveyMonkey males, SMF = SurveyMonkey females.

| Response | SM (N=62) | V (N=69) | SSC (N=613) | AM (N=7) | H (N=176) | SMM (N |
|---|---|---|---|---|---|---|
| Several times per week | 2% | 0% | 2% | 29% | 6% | 4% |
| About once per week | 8% | 22% | 18% | 14% | 34% | 14% |
| About once per month | 18% | 36% | 34% | 29% | 31% | 18% |
| About once per several months | 21% | 22% | 27% | 0% | 19% | 29% |
| Never/almost never | 52% | 20% | 19% | 29% | 10% | 36% |

Two comments on this question (out of 59) – both from the SSC group – specifically bemoaned deletionism, with one comment calling deletionism "a cancer killing Wikipedia".

## S1Q5: behavior on pages

This question was intended to gauge how often people perform an action for a specific page; as such, the frequencies are expressed in page-relative terms.

The following table presents the scores for each response, which are weighted by the number of responses. The scores range from 1 (for every page) to 5 (never); in other words, the lower the number, the more frequently one does the thing.

Table 6: For what fraction of pages you read do you do the following? Note that the responses have been shortened here; see the "Survey questions" section for the wording used in the survey. Responses are sorted by the values in the SSC column. SM = SurveyMonkey audience, V = Vipul Naik's timeline, SSC = Slate Star Codex audience, AM = Wikipedia Analytics mailing list, H = heavy users (26 or more articles per week) of Wikipedia, SMM = SurveyMonkey males, SMF = SurveyMonkey females.

| Response | SM (N=54) | V (N=69) | SSC (N=596) | AM (N=7) | H |
|---|---|---|---|---|---|
| Check $\geq 1$ citation | 3.57 | 2.80 | 2.91 | 2.67 | 2.6 |
| Look at "See also" | 3.65 | 2.93 | 2.92 | 2.67 | 2.7 |
| Read mostly for "Criticisms" or "Reception" | 4.35 | 3.12 | 3.34 | 3.83 | 3.1 |
| Click through $\geq 1$ source to verify information | 3.80 | 3.07 | 3.47 | 3.17 | 3.3 |
| Share the page | 4.11 | 3.72 | 3.86 | 3.67 | 3.7 |
| Look at the talk page | 4.31 | 4.28 | 4.03 | 3.00 | 3.8 |

| Response | SM (N=54) | V (N=69) | SSC (N=596) | AM (N=7) | H |
|---|---|---|---|---|---|
| Look at the editing history | 4.35 | 4.32 | 4.12 | 3.33 | 3.9 |
| Edit a page for grammatical/typographical errors | 4.50 | 4.41 | 4.22 | 3.67 | 4.0 |
| Edit a page to add new information | 4.61 | 4.55 | 4.49 | 3.83 | 4.3 |
| Look at editing history to verify author | 4.50 | 4.65 | 4.48 | 3.67 | 4.7 |
| Check how many pageviews a page is getting | 4.63 | 4.88 | 4.96 | 3.17 | 4.9 |

The table above provides a good ranking of how often people perform these actions on pages, but not the distribution information (which would require three dimensions to present fully). In general, the more common actions (scores of 2.5–4) had responses that clustered among "For some pages", "For very few pages", and "Never", while the less common actions (scores above 4) had responses that clustered mainly in "Never".

One comment (out of 43) – from the SSC group, but a different individual from the two in S1Q4 – bemoaned deletionism.

## S2Q1: number of Wikipedia pages read per week

Note the wording changes from S1Q1: "less" was changed to "fewer", the clarification "at least one sentence of" was added, and we explicitly allowed any language. (The explicit allowing of any language was in the introduction to the survey and not part of the question itself). We have also presented the survey 1 results for the SurveyMonkey audience in the corresponding rows, but note that because of the change in wording, the correspondence isn't exact. For more, see the S1Q1 explanation.

Table 7: How many distinct Wikipedia pages do you read (at least one sentence of) per week on average? SM = SurveyMonkey audience with no demographic filters, CEYP = College-educated young people of SurveyMonkey, S1SM = SurveyMonkey audience with no demographic filters from the first survey, SMM = SurveyMonkey males, SMF = SurveyMonkey females, CEYPM = College-educated young males of SurveyMonkey, CEYPF = College-educated young females of SurveyMonkey.

| Response | SM (N=54) | CEYP (N=50) | S1SM (N=62) | SMM (N=25) | SMF (N=26) | CEYPM (N=24 |
|---|---|---|---|---|---|---|
| Fewer than 1 | 37% | 32% | 42% | 32% | 42% | 29% |
| 1 to 10 | 48% | 64% | 45% | 40% | 54% | 67% |
| 11 to 25 | 7% | 2% | 13% | 16% | 0% | 4% |
| 26 or more | 7% | 2% | 0% | 12% | 4% | 0% |
| pgs/wk lower | 3.07 | 1.38 | 1.88 | 5.28 | 1.58 | 1.11 |
| pgs/wk upper | 9.02 | 7.82 | 8.17 | 11.92 | 7.02 | 7.99 |

The "pgs/wk lower" is obtained as the average pages read per week if everybody read at the lower end of their estimate (so the respective estimates are 0, 1, 11, and 26). The "pgs/wk upper" is obtained as the average of pages read per week if everybody read at the upper end of their estimate, except the "26 or more" case where we assume a value of 50 (so the respective estimates are 1, 10, 25, and 50).

Comparing SM with S1SM, we see that probably because of the wording, the percentages have drifted in the direction of more pages read. It might be surprising that the young educated audience seems to have a smaller fraction of heavy users than the general population. However note that each group only had ~50 responses, and that we have no education information for the SM group.

## S2Q2: multiple-choice of articles read

Our intention with this question was to see if people's stated or recalled article frequencies matched the actual, revealed popularity of the articles. Therefore we present the pageview data[32] along with the percentage of people who said they had read an article.

Table 8: Which of these articles have you read (at least one sentence of) on Wikipedia (select all that apply)? SM = SurveyMonkey audience with no demographic filters, CEYP = College-educated young people of SurveyMonkey. Columns "2016" and "2015" are desktop pageviews in millions. Note that the 2016 pageviews only include pageviews through the end of June. The rows are sorted by the values in the CEYP column followed by those in the SM column.

| Response | SM (N=54) | CEYP (N=50) | 2016 | 2015 |
|---|---|---|---|---|
| None | 37% | 40% | — | — |
| World War II | 17% | 22% | 2.6 | 6.5 |
| Barack Obama | 17% | 20% | 3.0 | 7.7 |
| United States | 17% | 18% | 4.3 | 9.6 |
| Donald Trump | 15% | 18% | 14.0 | 6.6 |
| Taylor Swift | 9% | 18% | 1.7 | 5.3 |
| Bernie Sanders | 17% | 16% | 4.3 | 3.8 |
| Japan | 11% | 16% | 1.6 | 3.7 |
| Adele | 6% | 16% | 2.0 | 4.0 |
| Hillary Clinton | 19% | 14% | 2.8 | 1.5 |
| China | 13% | 14% | 1.9 | 5.2 |
| The Beatles | 11% | 14% | 1.4 | 3.0 |
| Katy Perry | 9% | 12% | 0.8 | 2.4 |

[32] https://web.archive.org/web/20160714023739/http://wikipediaviews.org/displayviewsformultipleyears.php?tag=Pages%20in%20SurveyMonkey%20second%20survey

| Response | SM (N=54) | CEYP (N=50) | 2016 | 2015 |
|---|---|---|---|---|
| Google | 15% | 10% | 3.0 | 9.0 |
| India | 13% | 10% | 2.4 | 6.4 |
| Justin Bieber | 4% | 8% | 1.6 | 3.0 |
| Justin Trudeau | 9% | 6% | 1.1 | 3.0 |

Below are four plots of the data. Note that r_s denotes Spearman's rank correlation coefficient[33]. Spearman's rank correlation coefficient is used instead of Pearson's *r* because the former is less affected by outliers. Note also that the percentage of respondents who viewed a page counts each respondent once, whereas the number of pageviews does not have this restriction (i.e. duplicate pageviews count), so we wouldn't expect the relationship to be entirely linear even if the survey audiences were perfectly representative of the general population.

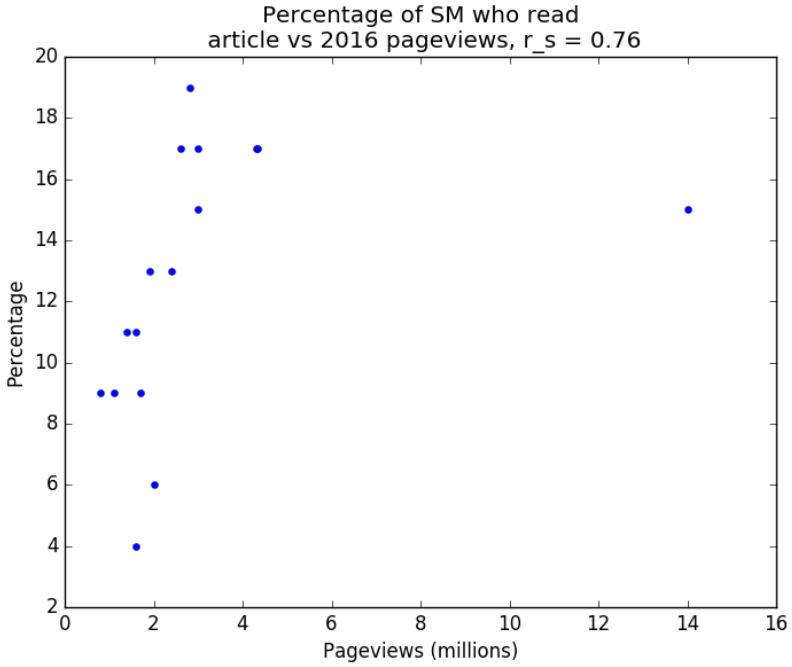


Figure 1: SM vs 2016 pageviews

[33]https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

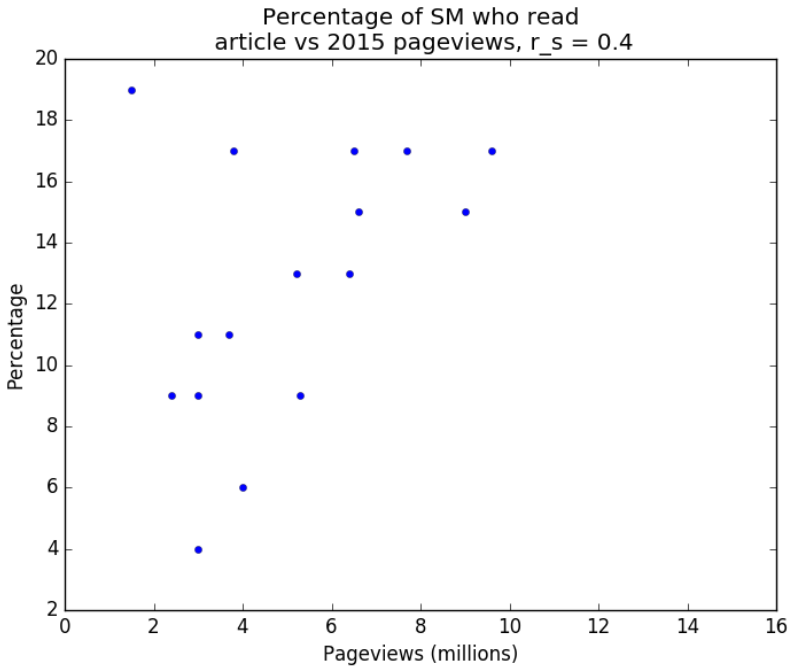


Figure 2: SM vs 2015 pageviews

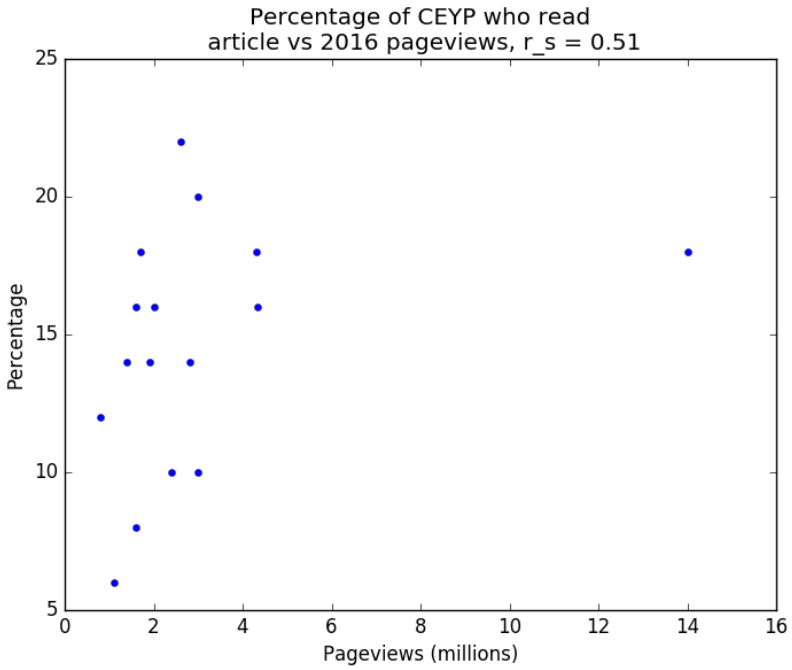


Figure 3: CEYP vs 2016 pageviews

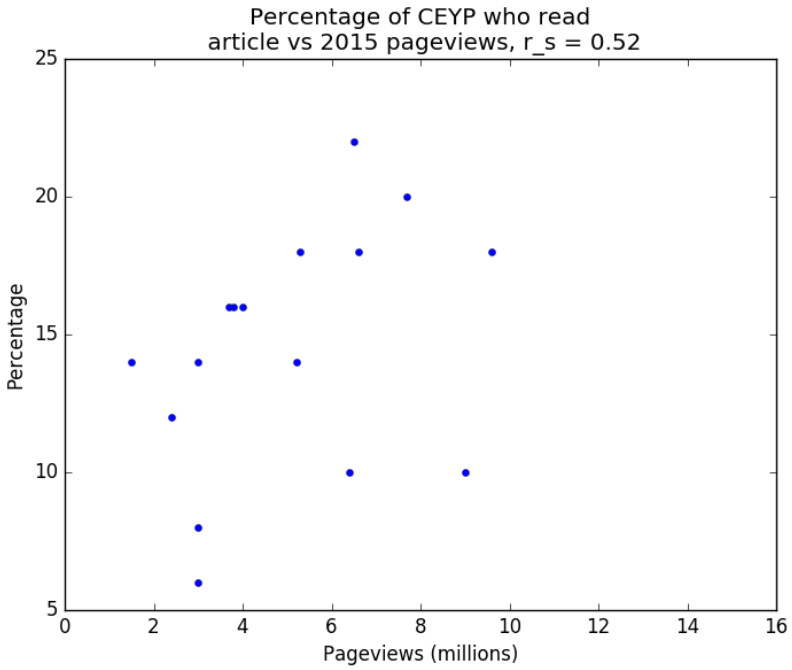


Figure 4: CEYP vs 2015 pageviews

## S2Q3: free response of articles read

The most common response was along the lines of "None", "I don't know", "I don't remember", or similar. Among the more useful responses were:

- News stories (e.g. Death of Harambe[34], "WikiLeaks scandal" – unclear which page this is, since there are several pages on various aspects of WikiLeaks)
- Popular culture:
  - People including Megan Fox[35], LeBron James[36], Rita Hayworth[37]
  - Works including Aladdin and the King of Thieves[38], X-Men: Apocalypse[39]
- More traditional encyclopedic information (e.g. Emerald ash borer[40], Spain[41], Siphonophorae[42], Scolopendra gigantea[43])

## S2Q4: free response of surprise at lack of Wikipedia pages

As with the previous question, the most common response was along the lines of "None", "I don't know", "I don't remember", "Doesn't happen", or similar.

The most useful responses were classes of things: "particular words", "French plays/books", "Random people", "obscure people", "Specific list pages of movie genres", "Foreign actors", "various insect pages", and so forth.

## GS

The survey was circulated to a target size of 500 in the United States (no demographic filters), and received 501 responses.

Since there was only one question, but we obtained data filtered by demographics in many different ways, we present this table with the columns denoting responses and the rows denoting the audience segments.

We also include the S1Q1SM, S2Q1SM, and S2Q1CEYP responses for easy comparison. Note that S1Q1SM did not include the "at least one sentence of" caveat. We believe that adding this caveat would push people's estimates upward.

[34] https://en.wikipedia.org/wiki/Death_of_Harambe
[35] https://en.wikipedia.org/wiki/Megan_Fox
[36] https://en.wikipedia.org/wiki/LeBron_James
[37] https://en.wikipedia.org/wiki/Rita_Hayworth
[38] https://en.wikipedia.org/wiki/Aladdin_and_the_King_of_Thieves
[39] https://en.wikipedia.org/wiki/X-Men:_Apocalypse
[40] https://en.wikipedia.org/wiki/Emerald_ash_borer
[41] https://en.wikipedia.org/wiki/Spain
[42] https://en.wikipedia.org/wiki/Siphonophorae
[43] https://en.wikipedia.org/wiki/Scolopendra_gigantea

If you view the Google Surveys results online[44] you will also see the 95% confidence intervals for each of the segments. Note that percentages in a row may not add up to 100% due to rounding or due to people entering "Other" responses. For the entire GS audience, every pair of options had a statistically significant difference, but for some subsegments, this was not true.

Table 9: How many distinct Wikipedia pages do you read (at least one sentence of) per week, on average? SM = SurveyMonkey Audience, GS = Google Surveys, SMM = SurveyMonkey males, SMF = SurveyMonkey females.

| Audience segment | Fewer than 1 | 1 to 10 | 11 to 25 | 26 or more | pgs/wk range |
|---|---|---|---|---|---|
| S1Q1SM (N=62) | 42% | 45% | 13% | 0% | 1.88–8.17 |
| S1Q1SMM (N=28) | 25% | 46% | 29% | 0% | 3.65–12.10 |
| S1Q1SMF (N=26) | 58% | 42% | 0% | 0% | 0.42–4.70 |
| S2Q1SM (N=54) | 37% | 48% | 7% | 7% | 3.07–10.42 |
| S2Q1SMM (N=25) | 32% | 40% | 16% | 12% | 5.28–14.32 |
| S2Q1SMF (N=26) | 42% | 54% | 0% | 4% | 1.58–7.82 |
| S2Q1CEYP (N=50) | 32% | 64% | 2% | 2% | 1.38–8.22 |
| S2Q1CEYPM (N=24) | 29% | 67% | 4% | 0% | 1.11–7.99 |
| S2Q1CEYPF (N=26) | 35% | 62% | 0% | 4% | 1.66–8.55 |
| GS all (N=501) | 47% | 35% | 12% | 6% | 3.23–9.73 |
| GS male (N=205) | 41% | 38% | 16% | 5% | 3.44–10.71 |
| GS female (N=208) | 52% | 34% | 10% | 5% | 2.74–8.92 |
| GS 18–24 (N=54) | 33% | 46% | 13% | 7% | 3.71–11.68 |
| GS 25–34 (N=71) | 41% | 37% | 16% | 7% | 3.95–11.61 |
| GS 35–44 (N=69) | 51% | 35% | 10% | 4% | 2.49–8.51 |
| GS 45–54 (N=77) | 46% | 40% | 12% | 3% | 2.50–8.96 |
| GS 55–64 (N=69) | 57% | 32% | 7% | 4% | 2.13–7.52 |
| GS 65+ (N=50) | 52% | 24% | 18% | 4% | 3.26–9.42 |
| GS Urban (N=176) | 44% | 35% | 14% | 7% | 3.71–10.94 |
| GS Suburban (N=224) | 50% | 34% | 10% | 6% | 3.00–9.40 |
| GS Rural (N=86) | 44% | 35% | 14% | 6% | 3.45–10.44 |
| GS $0–24K (N=49) | 41% | 37% | 16% | 6% | 3.69–11.11 |
| GS $25–49K (N=253) | 53% | 30% | 10% | 6% | 2.96–9.03 |
| GS $50–74K (N=132) | 42% | 39% | 13% | 6% | 3.38–10.57 |
| GS $75–99K (N=37) | 43% | 35% | 11% | 11% | 4.42–12.18 |
| GS $100–149K (N=11) | 9% | 64% | 18% | 9% | 4.78–15.49 |
| GS $150K+ (N=4) | 25% | 75% | 0% | 0% | 0.75–7.75 |

The "pgs/wk range" is obtained as follows. The lower bound is obtained as the average pages read per week if everybody read at the lower end of their

[44] https://www.google.com/insights/consumersurveys/view?survey=o3iworx2rcfixmn2x5shtlppci&question=1&filter=&rw=1

estimate (so the respective estimates are 0, 1, 11, and 26). The upper bound is obtained as the average of pages read per week if everybody read at the upper end of their estimate, except the "26 or more" case where we assume a value of 50 (so the respective estimates are 1, 10, 25, and 50). For more, see the S1Q1 explanation.

We can see that the overall GS data vindicates the broad conclusions we drew from SurveyMonkey data. Moreover, most GS segments with a sufficiently large number of responses (50 or more) display a similar trend as the overall data. One exception is that younger audiences seem to be slightly less likely to use Wikipedia very little (i.e. fall in the "Fewer than 1" category), and older audiences seem slightly more likely to use Wikipedia very little.

## Data validation using known total United States Wikipedia pageviews

Using the country breakdown data[45] for traffic to Wikipedia, we see that Wikipedia received 3.54 billion views in the United States for a recent 30-day period, which translates to about 827 million weekly pageviews.

Estimates for the number of active Internet users in the United States vary, based on definition, between 150 million and 290 million. With these estimates, we get a range of 2.85–5.51 for the number of pageviews per week for a United States user. We see that this range is loosely within the range for the SurveyMonkey surveys as well as Google Surveys. In other words, the survey data is loosely plausible and consistent with known facts.

## NRQ1: Do you use the Internet? and NRQ4: Have you ever heard of Wikipedia?

Both questions were asked in the New Readers phone survey for all five countries. NRQ1 was the same across all countries (though for Egypt, the "No" responses were further split to separate people who used Facebook). NRQ4 was asked as Q3 in Nigeria, Mexico, and Brazil.

We additionally want to know the percentage of Internet users who have heard of Wikipedia, as this will be useful later when making estimates of total pages/week read by people. We don't directly know this number. However, if we assume that the people who have heard of Wikipedia are a subset of the people who use the Internet, then we can compute this percentage as the ratio of the percentage of Yes responses to NRQ4 and NRQ1. This assumption

[45] https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdownHuge.htm

is a reasonable proxy for reality, so we will use the ratio as a stand-in for the percentage of Internet users who have heard of Wikipedia.

Table 10: New Readers question responses. NRQ1 = Do you use the Internet? NRQ4 = Have you heard of Wikipedia?

| Country | NRQ4 Yes | NRQ1 Yes | Ratio |
|---|---|---|---|
| Nigeria (N=2768) | 23% | 65% | 35% |
| India (N=9235) | 25% | 64% | 39% |
| Mexico (N=2568) | 45% | 80% | 56% |
| Brazil (N=5343) | 32% | 77% | 42% |
| Egypt (N=3976) | 17% | 59% | 29% |

An interesting note of comparison: for the surveys we circulated, we did not even ask people if they had heard of Wikipedia. The implicit assumption was that people had heard of Wikipedia. This assumption was probably reasonable in the contexts we operated in, and it didn't make sense to waste a question (and the underlying survey costs) on getting that information.

## NRQ7: How often do you use Wikipedia?

This question was in all country surveys, though at different positions.

The respondents to this question appear to have been selected as only the ones who had heard of Wikipedia.

Table: How often do you use Wikipedia? N values represent respondents to the question. Country Daily Weekly Monthly Rarely Never pgs/wk range ——— ——— —— ——— ———- ——— —— ———— Nigeria (N=610) 20% 24% 15% 17% 24% 1.07–11.35 India (N=2270) 22% 26% 16% 20% 16% 1.17–12.46 Mexico (N=1169) 18% 33% 19% 17% 14% 1.09–10.84 Brazil (N=1736) 13% 33% 23% 20% 11% 0.89–8.38 Egypt (N=665) 11% 23% 23% 24% 19% 0.72–6.88

The pgs/wk range is calculated as follows. For daily use, we assume between 4 and 50 views a week. For weekly use, we assume between 1 and 5 views a week. For monthly use, we assume between 0.2 and 1 view a week. We do not count any contribution for "Rarely" and "Never".

We also calculate the percentages relative to the set of all survey respondents (so that the denominator now includes people who have never heard of Wikipedia) and add all the ones who didn't respond to the Never column:

Table: How often do you use Wikipedia? N values represent respondents to the survey. Those who did not respond to the question are placed in the Never category. Country Daily Weekly Monthly Rarely Never pgs/wk range ——— ———- —— ——— ———- ——— —— ———— Nigeria (N=2768) 4.4% 5.3% 3.3% 3.7% 83.3% 0.24–2.50 India (N=9235) 5.4% 6.4% 3.9% 4.9% 79.3% 0.29–3.06 Mexico (N=2568) 8.2% 15.0% 8.7% 7.7% 61.0% 0.50–4.94 Brazil (N=5343) 4.2% 10.7% 7.5% 6.5% 71.1% 0.29–2.71 Egypt (N=3976) 1.8% 3.8% 3.8% 4.0% 86.5% 0.12–1.13

Next, we do the same calculation, but now use our denominator as the number of people who use the Internet. This is the closest in spirit to the audience for SurveyMonkey Audience and Google Surveys in the United States, though the selection dynamic does differ quite a bit.

Table: How often do you use Wikipedia? N values represent respondents to the survey who use the Internet? Those who use the Internet but did not respond to this question are placed in the Never category. Country Daily Weekly Monthly Rarely Never pgs/wk range ————————- —— ——— ——— ——— —— ——— Nigeria 7.1% 8.5% 5.3% 6.0% 73.1% 0.38–4.03 India 8.6% 10.2% 6.3% 7.8% 67.3% 0.46–4.87 Mexico 10.1% 18.6% 10.7% 9.6% 51.6% 0.61–6.09 Brazil 5.4% 13.7% 9.6% 8.3% 63.0% 0.37–3.48 Egypt 3.2% 6.6% 6.6% 6.9% 76.7% 0.21–2.00

## Comparison against United States audiences

The combin can be compared with S1Q1, S2Q1, and GS. However, the buckets presented to users were very different. The potential correspondence is below.

1. How many distinct Wikipedia pages do you read (at least one sentence of) per week on average?
    - Fewer than 1: This corresponds to Monthly, Rarely, and Never.
    - 1 to 10: This corresponds to Weekly and a subset of Daily.
    - 11 to 25: This mostly corresponds to Daily.
    - 26 or more: This mostly corresponds to Daily.

The data show that the people surveyed read Wikipedia less than the SurveyMonkey Audience and Google Surveys audiences. The total of the Monthly, Rarely, and Never columns for each of the five countries is over 70%, and it is over 80% for all countries other than Mexico. The corresponding "Fewer than 1" percentage for each iteration of urveyMonkey Audience and Google Surveys is less than 50%, and even on subsegments it is less than 60%.

In other words, the surveys suggest that Wikipedia use is less in the five countries than in the United States.

## Data validation against known total country traffic

We get the estimate for weekly traffic by scaling from 30 days to 7 days the country breakdown data[46]. Data was captured on December 23, 2016.

We get Internet-using population estimates from the Wikipedia page[47], which in turn relies on Internet Live Stats. Estimates were captured on December 23,

[46]https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerCountryBreakdownHuge.htm

[47]https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users

2016. We use this data rather than the data from stats.wikimedia.org since this data is more up to date, and includes extrapolated estimates rather than the most recent confirmed estimate.

Internet user and weekly pageview counts are in the millions. The range is the one computed based on the Internet-using population.

We see that the pgs/wk number is a little lower than the range for the case of Nigeria and India, with the gap particularly huge in Nigeria. Otherwise, however, the ranges are plausible and so the pageview data loosely validates the survey results.

Table: Comparison of known data on Internet users and Wikipedia pageviews against previous estimates of pages/week from survey Country Internet users Weekly pageviews pgs/wk pgs/wk range ——- ————— ————— ——- ————— Nigeria 86.2 7.88 0.09 0.38–4.03 India 462.1 127.20 0.28 0.46–4.87 Mexico 68.3 71.01 1.04 0.61–6.09 Brazil 120.1 71.77 0.59 0.37–3.48 Egypt 42.3 9.59 0.23 0.21–2.00

## Summaries of responses (exports for SurveyMonkey, weblink for Google Surveys)

SurveyMonkey allows exporting of response summaries. Here are the exports for each of the audiences.

- Survey 1, SurveyMonkey's audience[48]
- Survey 1, Vipul's timeline[49]
- Survey 1, Wikipedia Analytics mailing list[50]
- Survey 1, Slate Star Codex[51]
- Survey 1, Heavy users[52]
- Survey 1, SurveyMonkey's audience by gender[53]
- Survey 2, SurveyMonkey's audience, no demographic filters[54]
- Survey 2, SurveyMonkey's audience of college-educated young people[55]
- Survey 2, SurveyMonkey's audience, no demographic filters, by gender[56]
- Survey 2, SurveyMonkey's audience of college-educated young people, by gender[57]

[48] https://files.vipulnaik.com/surveys/SurveyMonkey.pdf
[49] https://files.vipulnaik.com/surveys/Vipul_timeline.pdf
[50] https://files.vipulnaik.com/surveys/Wikipedia_analytics_mailing_list.pdf
[51] https://files.vipulnaik.com/surveys/Slate_Star_Codex.pdf
[52] https://files.vipulnaik.com/surveys/Heavy_users.pdf
[53] https://files.vipulnaik.com/surveys/SurveyMonkey_gender.pdf
[54] https://files.vipulnaik.com/surveys/S2_unfiltered.pdf
[55] https://files.vipulnaik.com/surveys/S2_educated.pdf
[56] https://files.vipulnaik.com/surveys/S2_unfiltered_gender.pdf
[57] https://files.vipulnaik.com/surveys/S2_educated_gender.pdf

The Google Surveys survey results are available online at https://www.google.com/insights/consumersurveys/view?survey=o3iworx2rcfixmn2x5shtlppci&question=1&filter=&rw=1.

## Takeaway: Huge gap between heavy users and general US audience, plus predictors of heavy use

The most striking finding to us was just how wide the gap is between audiences such as Vipul's Facebook friends and Slate Star Codex on the one hand, and general US Internet users (as measured through SurveyMonkey Audience and Google Surveys) on the other.

### Confirming the gap with numbers

Here are three different ways to slice the data to confirm the gap between the audiences.

- Percentage of respondents who view less than 1 Wikipedia page per week: For Vipul's Facebook friends, Slate Star Codex, and the Analytics mailing list, this was 0% or 1%.

In contrast, for all the SurveyMonkey Audience and Google Surveys segments considered, this was 25% or higher, with the most general US audiences and largest sample sizes giving numbers between 40% and 60%.

- Estimated pages/week range: For Vipul's friends, Slate Star Codex, and the Analytics mailing list, the lower end was 9 or higher, and the upper end was 19 or higher.

In contrast, for all the SurveyMonkey Audience and Google Surveys segments considered, the lower end was less than 5 and the upper end was less than 15.

- Percentage of respondents who view 26 or more Wikipedia pages per week: For Vipul's friends, Slate Star Codex, and the Analytics mailing list, the number was 16%, 27%, and 57% respectively. In contrast, for all the SurveyMonkey Audience and Google Surveys segments, this percentage was less than 13%, and for most of the larger segments it was less than 7%.

### Qualitative differences in other aspects of Wikipedia engagement

Through the additional questions in S1, we got evidence for these statements, true both for heavy users, and for audiences that have a larger proportion of heavy users:

- They tend to explicitly seek Wikipedia in search results (S1Q2).
- They are more likely to be surprised at the absence of a Wikipedia page (S1Q4).
- They are more likely to use the search functionality witihn Wikipedia (S1Q4).
- They are considerably more likely to engage with page content in various ways, including looking at the See Also section, sharing the page, focusing on Criticisms and Reception, checking citations, and checking the talk page (S1Q5).

However:

- They are not too different from the general US audience in terms of the extent to which they read a section versus the whole page (S1Q3).
- They are not noticeably more likely to engage in editing actions on Wikipedia (in other words, active Wikipedia editors constituted a small fraction of heavy users) (S1Q5).

### Predictors of audiences with high proportions of heavy users

Of the three audiences with a high proportion of heavy Wikipedia users: Vipul's Facebook friends, Slate Star Codex, and the Wikimedia Analytics mailing list, only the third has an obvious connection with Wikipedia. The first two audiences are not directly linked to Wikipedia, and this is evidenced somewhat by the low rate of Wikipedia editing in these audiences. This suggests that visiting a specific website or being in a specific friend group on social media can be good predictors of heavy Wikipedia use without necessarily predicting Wikipedia editing.

It would be interesting to run this survey among audiences of different websites and people in different friend networks to get a better sense of what attributes predict high Wikipedia use.

## Takeaway: Effect on impact estimates for pageviews

As described in the Motivation section, our interest in the topic stems partly from a desire to quantify the value of individual Wikipedia pageviews. The results we obtained caused us to revise our estimate upward, but with the important caveat of downgrading the reach of Wikipedia.

### Upgrading estimate of impact based on reader quality

For some pages, the main way it is impactful is if the right set of people read it. For such a page, getting 1,000 pageviews from the right people (the ones with the information, authority, and skill to act on it) is more valuable than getting 1,000 pageviews from people who happen to visit the page accidentally.

The qualities we have identified for heavy Wikipedia users around their explicit seeking of Wikipedia as well as their use of advanced features on Wikipedia to verify facts and learn more, give us a little more confidence that pages are being read by the right people who are equipped to take action on them.

Additionally, other information we have about the audiences with high proportions of heavy users (specifically, that they are friends with Vipul, read Slate Star Codex, or are on the Analytics mailing list) also give us reason to be optimistic about these readers relative to general Internet users.

### Potentially downgrading estimate of impact through reach

For some pages and Wikipedia use cases, the impact pathway crucially depends on a lot of people in diverse contexts and life situations reading it. The results we have obtained suggest, very tentatively, that the views on a given page are likely to come from a less diverse audience than we might naively think.

For instance, let's say we go on a spree to significantly improve Wikipedia's coverage of 100 pages related to healthy living habits, and we then see that the pages we've improved got 10 million pageviews collectively.

Naively, we might have thought that we were reaching millions of users. However, if a lot of Wikipedia's pageviews come from heavy users, there's a good chance that those 10 million pageviews came from a few hundred thousand of these heavy users.

For any given page or set of pages, we can only speculate.
Therefore, this downgrading is only potential, and is accompanied by considerable uncertainty.

## Takeaway: Comparison with demographic gaps in the US and worldwide

### Gender within the United States

For S1, S2, and GS, males in the general US audience used Wikipedia a bit more than females in the general US audience. The biggest and clearest gap was in

the percentage of people who view less than 1 Wikipedia page. The gaps were as follows:

- S1Q1: 25% of males and 58% of females view less than 1 Wikipedia page per week.
- S2Q1: 32% of males and 42% of females view fewer than 1 Wikipedia page per week.
- GS: 41% of males and 52% of females view fewer than 1 Wikipedia page per week.

The gaps at the higher end of pages/week were less statistically robust because the percentages were too small, and therefore easily affected by outlier individuals. With that said, the overall pages/week ranges for men were mostly higher than those for women, as expected.

The gender gap is consistent with past research on gender differences in Wikipedia reading[58].

Note, however, that people talking of Wikipedia's gender gap are usually referring to a gender gap in Wikipedia editing, and the two gender gaps are both conceptually different and very different in magnitude (the gender gap in editing is much stronger than the gender gap in reading).

The gender gap in reading is small compared to the difference with Vipul's Facebook friends, Slate Star Codex, and the Analytics Mailing List, all of which had 0% or 1% of people viewing less than 1 Wikipedia page per week.

Unfortunately, for the three audiences (Vipul's Facebook friends, Slate Star Codex, and the Analytics mailing list), we do not have gender data for individual respondents. The audiences from which the respondents were drawn are between 60% and 80% male, but it's plausible that the actual respondents had a gender proportion outside this range.

## Age within the United States

For S1 and S2, the number of people within each age bucket was too small to draw any conclusion. We did notice that on S2, older people were less likely to enter optional comments, but we don't know why (it could be because of greater difficulty typing rather than anything specific to Wikipedia).

For GS, we saw a clear age gradient. In particular, older pepole were more likely to select the "Fewer than 1" option for the number of Wikipedia pages they read per week. Here's a snippet from the Google Surveys results table showing that:

[58] http://www.sciencedirect.com/science/article/pii/S0740818810000356

Table 11: How many distinct Wikipedia pages do you read (at least one sentence of) per week on average?

| Audience segment | Fewer than 1 |
|---|---|
| GS 18–24 (N=54) | 33% |
| GS 25–34 (N=71) | 41% |
| GS 35–44 (N=69) | 51% |
| GS 45–54 (N=77) | 46% |
| GS 55–64 (N=69) | 57% |
| GS 65+ (N=50) | 52% |

These age differences pale in comparison with the differences with Vipul's Facebook friends, Slate Star Codex, and the Analytics mailing list, all of which had 1% or less of their users viewing fewer than 1 page per week.

Unfortunately, for the three audiences (Vipul's Facebook friends, Slate Star Codex, and the Analytics mailing list), we do not have age data for individual respondents. The audiences from which the respondents were drawn are mostly in the 18–24, 25–34, and 35–44 age groups.

## Cross-country comparison in perspective

Here's an ordering by Wikipedia use:

Low-income countries (India, Nigeria, Brazil, Egypt) < Mexico < United States < Audiences such as Vipul's Facebook friends, Slate Star Codex, Analytics mailing list < Heavy users

Here are the estimates:

- Low-income countries: 0.05 to 0.6 pages/week per Internet user (based on actual pageview data)
- Mexico: Around 1 page/week per Internet user (based on actual pageview data)
- United States: Between 2.85 and 5.51 pages/week per Internet user (based on actual pageview data)
- Vipul's Facebook friends and Slate Star Codex: Between 9 and 26 pages/week per Internet user
- Heavy users: At least 26 pages/week per Internet user.

Thus, we see that the gap from the United States average to a heavy user is about the same as the gap from a low-income country to the United States.

Here's another way of thinking about it. Wikipedia as a whole got about 16 billion pageviews over a recent 30-day period. If Internet users everywhere

used it as much as they do in the United States (even at current Internet penetration rates) Wikipedia would get around double that many pageviews, or about 32 billion a month. If Internet users everywhere used Wikipedia as much as Slate Star Codex readers, Wikipedia would get between 150 billion and 300 billion pageviews a month (a number comparable to the total number of Google searches performed worldwide). If everybody in the world had Internet connectivity and used Wikipedia as much as Slate Star Codex readers do, Wikipedia would get between 400 billion and 800 billion monthly pageviews.

# Further reading

- "The great decline in Wikipedia pageviews (condensed version)"[59] by Vipul Naik
- "In Defense Of Inclusionism"[60] by gwern

# The making of this post

## Document source

The document and all sources used to compile it are available as a GitHub Gist[61].

The document is also available as a PDF.

## Original version and revision history

This post is a fork of an earlier post of Issa Rice available at http://lesswrong.com/r/discussion/lw/nru/wikipedia_usage_survey_results/ and as a PDF at https://files.issarice.com/wikipedia-survey-results.pdf/

The source files used to compile the earlier document are available in a GitHub Gist[62].

The earlier post has the following major revision history:

- 2016-07-14: Initial public version.
- 2016-08-27: A summary is added to the top of the post.
- 2016-10-05: Google Surveys (then Google Consumer Surveys) results are added.

[59] http://lesswrong.com/lw/lxc/the_great_decline_in_wikipedia_pageviews/
[60] http://www.gwern.net/In%20Defense%20Of%20Inclusionism
[61] https://gist.github.com/vipulnaik/c79a10aea048d6d2ceeb871b90a8e3f2
[62] https://gist.github.com/riceissa/7296b85303f96f2dc1dfb4f9a8ea44d9

The current version of the post has been written by me (Vipul Naik), with some feedback from Issa Rice. All errors and imperfections are mine.

### Survey cost

The survey response collection cost was $325, broken down as follows:

- S1, to SurveyMonkey Audience: $100, for 50 responses at $2 per response
- S2, to SurveyMonkey Audience: $50, for 50 responses at $1 per response
- S2, to SurveyMonkey Audience with filters (college-educated, age 18–29): $125, for 50 responses at $1.25 per response
- GS (Google Surveys): $50, for 500 responses at 10 cents per response

## License



[63] https://creativecommons.org/publicdomain/zero/1.0/