# Wikipedia Matters*

Marit Hinnosaar[†] Toomas Hinnosaar[‡] Michael Kummer[§]
Olga Slivko[¶]

September 13, 2017

**Abstract**

We document a causal influence of online user-generated information on real-world economic outcomes. In particular, we conduct a randomized field experiment to test whether additional information on Wikipedia about cities affects tourists' choices of overnight visits. Our treatment of adding information to Wikipedia increases overnight visits by 9% during the tourist season. The impact comes mostly from improving the shorter and incomplete pages in Wikipedia.

*JEL*: C93, H41, L17, L82, L83, L86
*Keywords*: field experiment, user-generated content, Wikipedia, tourism industry

## 1 Introduction

Asymmetric information can hinder efficient economic activity. In recent decades, the Internet and new media have enabled greater access to information than ever before. However, there is still unequal access created by the digital divide, language barriers, Internet censorship, technological constraints, and other factors. How much does it matter for economic outcomes?

In this paper, we analyze the causal impact of online information on real-world economic outcomes. In particular, we measure the impact of information on one of the primary economic decisions—consumption. As the source of information, we focus on Wikipedia. It is one of the most important online sources of information. It is the fifth most popular website in the world and receives about 18 billion direct page views per month.[1] However, the information available across Wikipedia's 299 language editions is

*We are grateful to Irene Bertschek, Avi Goldfarb, Shane Greenstein, Tobias Kretschmer, Thomas Niebel, Marianne Saam, Greg Veramendi, Joel Waldfogel, and Michael Zhang for valuable comments. Ruetger Egolf, David Neseer, and Andrii Pogorielov provided outstanding research assistance. Financial support from SEEK 2014 is gratefully acknowledged.

[†] Collegio Carlo Alberto and CEPR, marit.hinnosaar@gmail.com

[‡] Collegio Carlo Alberto, toomas@hinnosaar.net

[§] Georgia Institute of Technology, michael.kummer@econ.gatech.edu

[¶] ZEW, slivko@zew.de

[1] This does not include indirect uses like Apple's Siri or Google.

not the same. We analyze whether the differences in the available information have any impact on consumption.

To study the causal impact of information in Wikipedia on consumption choices, we conducted a randomized field experiment. Analyzing the impact of information using observational data would have been challenging, because of potential endogeneity. Products that are consumed more often tend to attract more attention and therefore, there is more information available about them. While an increase in information on Wikipedia tends to be correlated with the popularity of a product, the information is not necessarily causing consumption, instead it could be its byproduct. We overcome the identification problem using randomization.

We added information to randomly chosen Wikipedia pages in randomly chosen languages. We measured the outcome using data on tourists' overnight hotel stays in Spain. Spanish tourism sector is important in itself by accounting for almost 5% of Spain's GDP.[2] It also provided a good setting for the study, because the Spanish National Statistical Institute collects information about overnight stays in Spanish hotels at the level of city, month, and tourists' country of origin. Our treatment added text and photos to the Wikipedia pages of Spanish cities in different language editions of Wikipedia. The added text was translated mostly from the Spanish Wikipedia. The text was on topics relevant for tourists, such as city's main sights and culture. We focused our attention to cities with rather short Wikipedia pages. The randomization was done across city and language pairs. By varying the information available in various language editions of Wikipedia, we can tease out the causal impact on tourist choices.

We find that information in Wikipedia has a sizable impact on tourists' choices. Our estimates show that adding about 2000 characters (or about two paragraphs) and one photo to the Wikipedia page of a particular city increases the number of number of nights spent in this city by about 9% during the tourist season. The effect comes mostly from the pages that were initially relatively incomplete.In particular, in cities with initially very short pages in a particular language, the treatment increases hotel stays by about 33%, while there was no effect on city-language combinations, where the page was well developed.

Using the data on readership from Wikipedia page views and search activity from Google Trends, we can also shed some light on the mechanism. The added information does not have a significant impact on the search activity outside of Wikipedia but significantly increases the Wikipedia article readership. That is, more detailed Wikipedia articles gain more attention of potential readers. The size of this effect is similar in magnitude to the effect on tourist choices.

Our results have three policy implications. First, the results have implications on economic inequality and digital divide. Languages can pose barriers that hinder efficient economic activity. Language barriers have slowed innovation (Peri, 2005), decreased trade (Anderson and van Wincoop, 2004), and changed investments (Grinblatt and Keloharju, 2001). In particular, languages create a major obstacle to access to information. Large differences remain across languages in the information available online. Our results imply that these differences may lead to significant differences in economic behavior between

[2] Source: Tourism statistics. Eurostat. `http://ec.europa.eu/eurostat/statistics-explained/index.php/Tourism_statistics`, accessed June 21, 2017.

various groups.

Second, on the macroeconomic level we show that online user-generated content can have a significant causal impact on economic behavior and economic outcomes. The treatment increased the number of hotel visits by 9%. If we extend this to the tourism industry, then the impact is large. In 2015, international tourists stayed in Spain for 270 million nights. The same year the international travel receipts equaled 51 billion euros in Spain and 116 billion in the EU. [3] While we cannot say whether the online user-generated content is changing the size of expenditures or reallocating it, it could be responsible for affecting the choices in the order of billions of euros.

Third, on the microeconomic level, our results highlight the importance of online presence. Making sure that a city, a firm, or a product is accurately represented in Wikipedia and other online information sources in all the relevant languages is relatively cheap. In comparison, the 9% increase in consumption is very large. According to recent estimates (García-Sánchez, Fernández-Rubio, and Collado, 2013), on average each international tourist visiting Spain spends about 101 euros per day during the visit. This is a high return on investment.

The results of the paper pose a puzzle—why is the online presence so limited? Increasing the online presence is relatively inexpensive, while we show that the potential benefits are large. The online presence puzzle is opposite to most of the literature that studies contributions to online public goods. That literature finds that there is too much contribution compared to what the economic theory would suggest. While the public goods literature assumes that contributions are altruistic, we concentrate on a setting where there are parties who would benefit from making more information available.

Our paper makes three methodological contributions. First, it is among the first papers that uses Wikipedia as a treatment in a field experiment to study the impact on the behavior outside of Wikipedia.[4] Wikipedia provides a good ground for this, since anyone can freely improve it[5] and the whole process is automatically recorded in the form of revision histories. Moreover, the consumption of Wikipedia is well-recorded in the form of page views.

Second, we use a novel dataset on real-life outcomes—overnight hotel stays. In many European countries hotels are required to collect the identifications of all guests, including making a copy of their travel document. We were able to obtain this data from the Spanish National Statistical Institute aggregated to monthly level for each city and each country of origin. For example, we observe how many nights German tourists spent in a particular city in July 2015. We are using the fact that German tourists are more likely to get their information from Wikipedia in German language and Italian tourists from Italian language to map the tourist flows back to their potential information sources.

Finally, we make a technical contribution on how we analyze revision histories. As our treatment is adding information to particular Wikipedia pages, which can then be changed by other Wikipedia users, the first step of the analysis is to see how much of our additions are changed over time. For this, we are using diff algorithm that describes the shortest sequence of additions and deletions of characters to change the original text to

[3] Source: Tourism statistics. Eurostat.`http://ec.europa.eu/eurostat/statistics-explained/index.php/Tourism_statistics`, accessed June 21, 2017.

[4] There is a literature that studies the editing behavior in Wikipedia, which we will review below.

[5] Of course, following the terms and conditions.

the new one.[6] We apply the diff algorithm twice. First, to quantify which parts of the page were added by our experiment, and second, to see how much of this was still present in a revision a few months later. We see that our edits are very persistent, about 95% of our edits still existed four months after the treatment. This represents the fact that the information on the pages we edited was relatively scarce and (hopefully) our contributions were considered sufficiently valuable by the Wikipedia community.

**Related literature** Our paper contributes to the media economics literature that studies the impact of media on economic outcomes (for an overview see DellaVigna and Ferrara (2016)). In particular, our paper adds to the studies about the impact of media on consumption. Most notably, Bursztyn and Cantoni (2015) use geographic variation in access to Western TV to study its long run impact on East German consumption choices. The paper also contributes to the studies on the impact of new media and online user-generated content.[7] Among others Chevalier and Mayzlin (2006); Luca (2011) study how product reviews affect sales. Enikolopov, Petrova, and Sonin (2017) analyze the impact of blog posts exposing corruption in state-controlled companies on their market returns. Xu and Zhang (2013) study the impact of Wikipedia on financial markets combining data of financial records, management disclosure records, news article coverage, and Wikipedia editing histories. Our paper adds to the literature by providing evidence of how Wikipedia affects consumption. It differs from these papers in terms of the research method. The above papers use either a natural experiment or detailed observational data, while we conduct a randomized field experiment which helps us to identify the effect.

Second, our paper relates to the emerging small branch of literature on information production in Wikipedia. Most of this literature analyzes contributions to Wikipedia (including Zhang and Zhu, 2011; Aaltonen and Seiler, 2015) and biases in Wikipedia (Greenstein and Zhu, 2012; Greenstein, Gu, and Zhu, 2016; Greenstein and Zhu, 2017). Our paper stresses the importance of understanding the Wikipedia production process and its biases by quantifying the impact of Wikipedia on offline economic behavior.

# 2 Background on Wikipedia

Wikipedia is a free-access Internet encyclopedia. It is the fifth most popular website in the world.[8] Wikipedia exists in 299 languages and altogether has 45 million articles (pages).[9]

[6] For a description of the algorithm, see Myers (1986). Although the method is standard in practice and computer science, it is not common in economic applications.

[7] More generally, it relates to the literature on how ICT by changing access to information affects economic outcomes. Among other topics this literature has studied the impact of Internet on economic growth (Czernich, Falck, Kretschmer, and Woessmann, 2011), on labor market outcomes (Forman, Goldfarb, and Greenstein, 2012; Akerman, Gaarder, and Mogstad, 2015), on airline industry (Dana and Orlov, 2014; Ater and Orlov, 2015), the impact of medical records on hospital costs (Dranove, Forman, Goldfarb, and Greenstein, 2014); the impact of electronic commerce on price dispersion (Overby and Forman, 2014).

[8] Only Google, Youtube, Facebook, and Baidu are more popular than Wikipedia. The popularity is measured by the web traffic measurement company Alexa Internet (`http://www.alexa.com/siteinfo/wikipedia.org`, accessed June 19, 2017).

[9] `https://meta.wikimedia.org/wiki/List_of_Wikipedias`, accessed June 19, 2017.

Wikipedia is written by volunteers. Anyone can create Wikipedia articles and edit almost any of its existing articles.

The amount of information available in Wikipedia differs across languages. English language Wikipedia is the largest with over five million articles. Only 13 language editions have more than a million articles each. All the language editions studied in the paper have each over one million articles (German: 2.0; French: 1.9; Italian: 1.4 million).[10]

Almost half of the population (46%) in the European Union does not speak any foreign language.[11] They can access the information only on the local language Wikipedia. This creates variation in access to information in Wikipedia. In eight EU countries (Poland, Czech Republic, Romania, Bulgaria, Spain, Portugal, Italy, and Hungary) the median person speaks only one (typically the local) language.[12] In all these countries, the local language Wikipedia is smaller than 30% of the English language Wikipedia (measured by the number of articles). Figure 1 shows the size of the local language Wikipedia (measuring the number of articles) and the percentage of population speaking more than one language. In the language editions studied in the paper, Wikipedia size varies from 25% (in Italian) to 38% (in German), while the percentage of population not speaking any foreign languages varies from 34% (in Germany) to 62% (in Italy).

It isn't only the topics covered that differ across languages, it's also the debt of coverage. Consider an example. Wikipedia keeps a list of 1000 most important topics.[13] The median length of text (relative to the corresponding page in English) across all topics varies from 5% in Latvian to 55% in French (see figure 2). Not all topics are covered equally (see figure 3). Overall, the worst covered topics are in categories like philosophy and religion (12%) and health and medicine (13%). The most covered topics are about people (22%) and geography (21%). Among the languages studied in the paper, pages in French Wikipedia are longer about most topics, in German Wikipedia about technology, society and social sciences, arts and culture, and in Italian about health and medicine. Importantly for this paper, pages about geography are shorter in Italian Wikipedia.

The implication that is relevant for this paper, is that the amount of information available in each language edition of Wikipedia is not the same. It varies in terms of both which pages exist and the depth of the material on each topic. Figure 4 presents an example of information about a city. It describes pages of Murcia, a large Spanish city, across the different language editions of Wikipedia. The page of Murcia exists in 84 different language editions of Wikipedia.[14] The figure presents the length of the page in the 20 languages in which the page is the longest. Because it is a city in Spain, the page is the longest in Spanish Wikipedia. In Spanish Wikipedia, the page is more than five times as long as the same page in English Wikipedia (which holds the second place).

[10] `https://meta.wikimedia.org/wiki/List_of_Wikipedias`, accessed June 19, 2017.

[11] That is, they speak only their mother tongue. Data source for language skills is Eurobarometer (2012).

[12] The median person speaks one language in Ireland and UK too, but he can read English Wikipedia.

[13] `https://en.wikipedia.org/wiki/Wikipedia:Vital_articles`, accessed June 26, 2017.

[14] Wikipedia data on Murcia was accessed on June 20, 2017.

# 3 Experimental design

We conducted a field experiment in which we added information (text and photos) to the Wikipedia pages of Spanish cities in different language editions of Wikipedia. The randomization was done across city and language pairs. The outcome variable is the number of hotel nights stayed by the tourists from the countries where the population is speaking one of the treated languages. The experimental design is discussed below in detail.

**Sample** We restricted attention to four languages and tourists from the corresponding countries: Dutch (the Netherlands), German (Germany), French (France), Italian (Italy). Altogether we had hotel data in 135 Spanish cities. However, in many of these cities, the hotel data was missing for some months and some tourist country of origin. Hence, we expected to run into the problem that we cannot measure the effect of the treatment because the outcome (hotel) data is missing. We were also concerned that our fixed length treatment might not be strong enough in case of cities which Wikipedia pages were already long.

Therefore, we restricted attention to a sample of cities that satisfied two criteria. First, that the Wikipedia page is relatively short. Namely, that in each of the four languages the page is no longer than 24,000 characters. Second, that there is no missing hotel data. Specifically that data on hotel stays exists for each month from May until October in 2013 in case of all four countries. There are 60 cities that satisfied these two criteria. This gave us a sample of 240 Wikipedia pages (or city–tourist country of origin pairs).

**Randomization** We randomized across 240 Wikipedia pages: pages of 60 Spanish cities in four languages. Our goal was to treat each city equally. Therefore, for each city, we treated its page in two randomly chosen language editions of Wikipedia. In each language edition of Wikipedia, we treated 30 pages of cities. This resulted in a design where for each city some languages are assigned to the treatment and some to the control group. Similarly, in each language, some cities are in the treatment and some in the control group.

To ensure some balance in the treatment and control group, we used a stratified randomization design. We ordered the 60 cities by the total number of tourists. Then we divided the cities into ten groups, each of the size of six cities. Inside each group, we randomly assigned the city to one of the six treatments. The six treatments were: treat the city page in one of the six possible pairs of languages (Dutch & German; Dutch & French; Dutch & Italian; German & French; German & Italian; French & Italian). Hence, 120 city pages were treated and 120 pages remained as controls.

**Treatment** To the pages in the treatment group, text and photos were added. The added text and photos were on topics relevant for tourists like the main sights and culture. Added text was translated mostly from the corresponding Spanish or English language Wikipedia pages. Typically, the photos were also from these corresponding Wikipedia pages. The pages were treated in the second half of August in 2014.

Our goal was to improve the Wikipedia pages. We did not make Wikipedia pages worse by deleting the existing material. Following Wikipedia policies, we added material that according to our understanding was conveying already established and recognized knowledge.

**Measuring our treatment and its survival** We applied diff algorithm twice to quantify how much we added by our treatment and how much of it was preserved a few months later. In particular, for each page we compared three revisions that we took from the Wikipedia revision history: the last revision prior to our changes (which we call *pre-treatment* revision), the last revision created by our treatment (*post-treatment*), and version a few months later (*survived*). In the revision history, the text is always in the Wikitext format, which means that some of it is not visible for the viewer. We normalized all the three revisions as follows. We used Wikipedia's built-in parser to get the html-version of the content, which we then converted to plain text by removing the html commands, i.e. removed all pictures, links, etc. This gave us three texts.

The length of pre-treatment is our page length measure. To quantify the content added by our treatment, we used a diff algorithm. It computes the smallest number of character additions and deletions from pre-treatment to post-treatment. The algorithm outputs which characters stayed the same, which ones were deleted, and which ones added. The total length of the added text is our measure of treatment length. Finally, to compute how much of the text survived after the editing process a few months later we computed diff from the added text to the survived text.[15] See figure 5 for illustration.

**Survival of added material** While editing German, French, and Italian Wikipedia was not problematic, we were not successful in editing Dutch Wikipedia. Wikipedia allows anyone to edit it. This also means that anyone can delete an article or some parts of it, or undo the latest changes by reverting to a previous version to the article. All our additions to Dutch Wikipedia were deleted in less than 24 hours. That is, all Dutch Wikipedia pages were essentially untreated from the point of view of a person reading these Wikipedia pages or accessing these indirectly using Apple's Siri or Google information box. Therefore, we exclude all Dutch Wikipedia articles from our analysis. Note that the results won't change much if we consider all Dutch articles as non-treated.

Table 1 shows that in German, French, and Italian Wikipedias our added text and photos survived well. Of the added text on average 96 percent had survived by the beginning of the following month after treatment and 93 percent by the beginning of the following year after treatment. We interpret this in two ways. First, the edits were sufficiently persistent to give us hope that sufficiently many people saw the information our treatment added. Strictly speaking, it is not necessary that the precise wording of our treatment added survives—it is to be expected that the other Wikipedia editors improve any added contributions over time in terms of wording, references, or content. But measuring the preserved content is more difficult than measuring the actual text.

[15] It is slightly imperfect measure, as there could be some text that was deleted, but the algorithm is unable to differentiate it from the other parts of the page (that were unrelated to our treatment), but in examples we checked by hand the results were accurate within a reasonable margin.

Second, we hope that the additions of our treatment were considered useful by the fellow Wikipedia editors, otherwise they would have either reversed the edits or changed more.

**Descriptive statistics** Table 2 shows that assignment into the treatment group was random in terms of the covariates.

Table 3 shows descriptive characteristics of treatment. The median treatment added about 2000 characters of text and one photo. The treatment added relatively more to pages that were initially shorter (see Figure 6).

Figure 7 presents the histogram of the logarithm of the number of hotel nights. There is a large variation in the number of hotel nights (as seen also in Table 4). Figure 8 presents the percentage of missing data by calendar month. It describes seasonality with slightly above ten percent of missing data from May to October and up to 40 percent in December and January.

# 4 Results

**Empirical strategy** Our goal is to estimate the impact of additional information in Wikipedia on hotel stays in the corresponding city by tourists from the corresponding country. The main outcome variable is the logarithm of the number of hotel nights stayed in city $i$ by tourists from country (exposed to language) $j$ during month $t$. In our main analysis, we estimate the following difference-in-differences regression:

$$log(Nights_{ijt}) = \alpha + \beta Treatment_{ijt} + \gamma X_{ijt} + CityLanguageFE_{ij} + \varepsilon_{ijt} \tag{1}$$

The variable of interest $Treatment$ equals one for the treated city-language pairs during the months after treatment and equals zero otherwise. The regression includes fixed effects for city-language pairs $CityLanguageFE_{ij}$ and time varying control variables, $X_{ijt}$. The time varying control variables include: first, an indicator for period after treatment interacted with language fixed effects to take into account tourist country of origin specific trend; second, an indicator for period after treatment interacted with city fixed effects to take into account city specific trend; third, logarithm of number of tourists from Spain interacted with language fixed effects to take into account events in the city which lead to an overall increase in tourism. We cluster the standard errors by city-language pair. Due to the missing data problem discussed above, in the main analysis, we restrict the sample to May - October during each year 2010–2015.

**Main results** Table 5 presents the main results. According to the estimate in column 1, the treatment increases the number of hotel nights on average by nine percent. Column 2, adds an interaction of the treatment variable and an indicator for Wikipedia pages that were initially relatively short. The estimates in column 2 show that in the case of the cities which pages in a particular language were initially very short, the treatment increases hotel stays by about 33% more visits, while there was no effect on others. Column 3, tries to explain the result by interacting the treatment variable and an indicator for the Wikipedia pages to which we added relatively longer text compared to the initial text length. Recall that since the length of text added was about the same, the treatment was relatively larger on pages that were initially short (Figure 6). Results in column 3 confirm that the effect is larger on pages where the treatment was relatively larger.

**Robustness** Table 6 presents a number of robustness checks. Columns 1–5 repeat regression in column 1 in table 5, so the magnitudes of the estimates are comparable.

Column 1 substitutes missing observations by zeros (only for city-year pairs, when data exists for some month and tourist country of origin). It excludes the variables that measure the number of tourists from Spain because the number of tourists from Spain is also missing. The results are very similar.

Column 2 adds observations for tourists from the Netherlands and considers these all as non-treated. The results are very similar. Recall that half of the city pages in Dutch Wikipedia were assigned to treatment, but editing Dutch Wikipedia proved impossible (24h after treatment all the pages remained untreated). We could estimate the same regression and adding a separate indicator variable that equals one for months after treatment only for Dutch pages assigned to treatment. The results regarding the main treatment effect remain the same.

Column 3 and 4 add remaining months and column 4 substitutes missing observations by zeros (only for city-year pairs, when data exists for some month and tourist country of origin). Again, the variables that measure the number of tourists from Spain are excluded. The results are similar, but in column 3, less statistically precise.

Column 5 adds additional controls, namely, the logarithm of the number of tourists from UK interacted by language. the variables that measure the number of tourists from Spain are excluded. Results are similar.

In column 6, the dependent variable is the number of tourists from country $j$ divided by the number of tourists from country $j$ plus from Spain and UK. Again, the variables that measure the number of tourists from Spain are excluded. While the results are not comparable in magnitude, the treatment effect is positive and statistically significant.

**Mechanism** What was the mechanism that led from information on Wikipedia to an increase in tourism? To understand the mechanism, we analyze two additional outcome variables. First is the page views of Wikipedia articles, which measures the number of visits to a particular article. Studying the page views of Wikipedia articles by language tells us whether our treatment changed the amount of attention the articles received. Second, we study Google Trends that is a measure of Google Search volume. It measures how often a particular city is searched for on Google by a population in a particular country. Studying Google Trends shows us whether there were changes in the interest for these cities.

Table 8 presents estimates of analogous regressions as equation 1. In columns 1–3, the outcome variable is the logarithm of the number of page views of a Wikipedia page of city $i$ in language $j$ during month $t$. In columns 4–6, the outcome variable is Google Trend of city $i$ from country $j$ during month $t$. Estimates in column 1 show that the treatment increased page views by about 14 percent. Regression in column 2 separates the effect by the length of the article (before treatment), showing that the treatment effect is larger on shorter pages. Relatedly, regression in column 3 show that the treatment effect is larger on pages where treatment added a relatively larger share of text (these tended to be shorter pages). Estimates in columns 4–6 show that our treatment had no effect on Google Trends (Google Search volume). The robustness of these estimates is studied in Table 9.

How to interpret the results? First, we would not expect that our treatment has any direct effect on search volume because search volume measures general interest that is outside of Wikipedia. However, there could be indirect effects where additional information on Wikipedia leads to an increased interest, for example, via word-of-mouth. The fact that search activity was not affected by the treatment indicates that our treatment worked only through the Wikipedia readership itself and not through the external channels. People started to read the Wikipedia articles more often without more often searching for them on Google. This could be because search engines showed the improved Wikipedia pages more prominently in the search results. This could also be because more informative articles lead readers to return to the pages.[16] In both cases, we can conclude that our treatment increased the readership significantly.

The size of the effect on readership is similar in magnitude than the effect on hotel nights. This is consistent with the interpretation that added information on Wikipedia increased demand mostly through added readership. As we do not observe unique page views, we cannot distinguish between higher conversion rate from readers to visitors and larger audience.

# 5 Discussion

We found a significant causal impact of user-generated content in Wikipedia on real-life choices. The impact is large. A well-targeted two-paragraph improvement may lead to a 9% increase in the visits by tourists. This has significant implications both in macroeconomic and microeconomic scale.

Overall, the estimated effect suggests that additional information on Wikipedia leads to a significant increase in the number of tourists in the city. The baseline estimates show an increase of nine percent. The median monthly number of hotel nights by tourists from the three countries to the cities in the control group was about 3000 (during the six months from May–October). This implies an increase of about 270 nights per month. Even conservatively, assuming no tourists in the remaining 6 month, it means about 1,600 additional hotel nights per year.

What are the implications for the local economy? According to recent estimates (García-Sánchez, Fernández-Rubio, and Collado, 2013), on average each international tourist visiting Spain spends about 101 EUR per day during his or her visit. Back-of-the-envelope calculations suggest that improving a city's Wikipedia page leads to about 160,000 euros. additional revenue per year. This implies considerable impact on the local hotels and overall local tourist industry.

Our results highlight the importance of online presence. Making sure that a city, firm, or product is accurately represented in online information sources in all the relevant languages is relatively cheap, i.e. almost free or costs a few hundred dollars in mostly

[16] Unfortunately, Wikipedia did not collect unique page views prior to 2015, therefore we cannot distinguish between new and returning readers.

one-time costs. In comparison, the 9%-increase in demand is rather large. This is very high return to investment.

The amount of information available in different languages is different. Our results imply that this may lead to large differences in economic decisions as well.More generally, the results pose questions about economic inequality and digital divide across cultural and ethnic groups.

We must note that the results might be specific to the languages and types of pages in which the experiment was conducted. In these languages, the city pages were typically not too long and had room for improvement. However, these language editions of Wikipedia are still among the largest, each with over one million articles. This reflects that these language editions receive a relatively large number of viewers. It is not clear what would be the impact of additional information in the case of smaller language editions of Wikipedia.

# References

Aaltonen, A., and S. Seiler (2015): "Cumulative Growth in User-Generated Content Production: Evidence from Wikipedia," *Management Science*, 62(7), 2054–2069.

Akerman, A., I. Gaarder, and M. Mogstad (2015): "The Skill Complementarity of Broadband Internet," *The Quarterly Journal of Economics*, 130(4), 1781–1824.

Anderson, J. E., and E. van Wincoop (2004): "Trade Costs," *Journal of Economic Literature*, 42(3), 691–751.

Ater, and E. Orlov (2015): "The Effect of the Internet on Performance and Quality: Evidence from the Airline Industry," *The Review of Economics and Statistics*, 97(1), 180–194.

Bursztyn, L., and D. Cantoni (2015): "A Tear in the Iron Curtain: The Impact of Western Television on Consumption Behavior," *The Review of Economics and Statistics*, 98(1), 25–41.

Chevalier, J. A., and D. Mayzlin (2006): "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43(3), 345–354.

Czernich, N., O. Falck, T. Kretschmer, and L. Woessmann (2011): "Broadband Infrastructure and Economic Growth," *The Economic Journal*, 121(552), 505–532.

Dana, J., and E. Orlov (2014): "Internet Penetration and Capacity Utilization in the US Airline Industry," *American Economic Journal: Microeconomics*, 6(4), 106–137.

DellaVigna, S., and E. L. Ferrara (2016): "Economic and social impacts of the media," in *Handbook of media economics*, ed. by S. Anderson, D. Stromberg, and J. Waldfogel. Elsevier, Amsterdam.

Dranove, D., C. Forman, A. Goldfarb, and S. Greenstein (2014): "The Trillion Dollar Conundrum: Complementarities and Health Information Technology," *American Economic Journal: Economic Policy*, 6(4), 239–270.

Enikolopov, R., M. Petrova, and K. Sonin (2017): “Social media and corruption,” *American Economic Journal: Applied Economics*, forthcoming.

Eurobarometer (2012): “Europeans and their Languages Report,” Special Report 386, European Commission.

Forman, C., A. Goldfarb, and S. Greenstein (2012): “The Internet and Local Wages: A Puzzle,” *American Economic Review*, 102(1), 556–575.

García-Sánchez, A., E. Fernández-Rubio, and M. D. Collado (2013): “Daily expenses of foreign tourists, length of stay and activities: evidence from Spain,” *Tourism Economics*, 19(3), 613–630.

Greenstein, S., Y. Gu, and F. Zhu (2016): “Ideological Segregation among Online Collaborators: Evidence from Wikipedians,” Working Paper 22744, National Bureau of Economic Research.

Greenstein, S., and F. Zhu (2012): “Is Wikipedia Biased?,” *American Economic Review: Papers and Proceedings*, 102(3), 343–348.

——— (2017): “Do Experts or Crowd-based Models Produce More Bias? Evidence from Encyclopedia Britannica and Wikipedia,” *MIS Quarterly*, forthcoming.

Grinblatt, M., and M. Keloharju (2001): “How Distance, Language, and Culture Influence Stockholdings and Trades,” *The Journal of Finance*, 56(3), 1053–1073.

Luca, M. (2011): “Reviews, Reputation, and Revenue: The Case of Yelp.com,” *manuscript*.

Myers, E. W. (1986): “AnO(ND) difference algorithm and its variations,” *Algorithmica*, 1(1-4), 251–266.

Overby, E., and C. Forman (2014): “The Effect of Electronic Commerce on Geographic Purchasing Patterns and Price Dispersion,” *Management Science*, 61(2), 431–453.

Peri, G. (2005): “Determinants of Knowledge Flows and Their Effect on Innovation,” *The Review of Economics and Statistics*, 87(2), 308–322.

Xu, S. X., and X. Zhang (2013): “Impact of Wikipedia on Market Information Environment: Evidence on Management Disclosure and Investor Reaction,” *MIS Q.*, 37(4), 1043–1068.

Zhang, X., and F. Zhu (2011): “Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia,” *The American Economic Review*, 101(4), 1601–1615.

# Tables and figures

Table 1: Survival over time of text and photos which we added to Wikipedia

| | France | Germany | Italy | Total |
|---|---|---|---|---|
| % text survived: 24h | 100.0 | 94.7 | 100.0 | 98.2 |
| % text survived: next month | 98.7 | 90.2 | 99.9 | 96.3 |
| % text survived: next year | 95.1 | 86.7 | 97.5 | 93.1 |
| % photos survived: 24h | 100.0 | 96.2 | 100.0 | 98.8 |
| % photos survived: next month | 100.0 | 92.3 | 96.4 | 96.4 |
| % photos survived: next year | 100.0 | 88.5 | 92.9 | 94.0 |
| Number of observations | 30 | 30 | 30 | 90 |

Note: Unit of observation is a city page in a given language Wikipedia. Percentage of text survived is calculated as described in section 3. % of text or photos survived is calculated over three time periods: 24 hours, by the beginning of the next calendar month after treatment, by the beginning of the next calendar year after treatment.

Table 2: Ability of covariates to predict treatment status

| | Coef. | p-value |
|---|---|---|
| Log(Sum of tourists in 2013) | -0.002 | 0.958 |
| Log(Number of tourists) | -0.012 | 0.527 |
| Tourist data missing | 0.045 | 0.556 |
| Log(Initial text length) | -0.000 | 0.994 |

Note: Dependent variable is the treatment group (an indicator that equals one if a city-language pair is assigned to the treatment group and zero if it is assigned to the control group). Each row presents estimates from a separate regression of the form: $TreatmentGroup_i = Constant + \beta Variable_i + \varepsilon_i$, where $Variable$ is listed in the first column. In rows 1 and 4, a unit of observation is a city-language pair. In rows 2 and 3, a unit of observation is a city-language-month triplet and the sample covers time period until treatment.

Table 3: Descriptive statistics of treatment

| | mean | sd | p25 | p50 | p75 | count |
|---|---|---|---|---|---|---|
| Length of text added | 2047.2 | 697.2 | 1671 | 2082 | 2377 | 90 |
| Number of photos added | 1.2 | 1.1 | 1 | 1 | 1 | 90 |
| % of text added | 43.2 | 37.9 | 18 | 29 | 56 | 90 |

Note: Unit of observation is a Wikipedia page in a given language (30 pages in each of the three languages: German, French, Italian).

Table 4: Descriptive statistics of the number of hotel nights in the control group

| | mean | sd | min | p25 | p50 | p75 | max | count |
|---|---|---|---|---|---|---|---|---|
| Nights | 23180.1 | 67229.6 | 21 | 897 | 3032 | 16790 | 543049 | 2871 |

Note: Unit of observation is a month, city, and tourist country of origin triplet. Sample includes tourists from Italy, France, Germany to the 60 cities in Spain, but only the city-country of origin pairs, which are assigned to the control group. The time period of the sample is May–October in 2010 - 2015.

Table 5: Dependent variable: Logarithm (number of hotel nights)

| | (1) | (2) | (3) |
|---|---|---|---|
| Treatment | 0.089** | 0.002 | 0.039 |
| | (0.045) | (0.038) | (0.045) |
| Treatment: Small page | | 0.332*** | |
| | | (0.100) | |
| Treatment: Large % added | | | 0.196* |
| | | | (0.099) |
| City-Language FE | Yes | Yes | Yes |
| Adj. R-squared | 0.245 | 0.248 | 0.246 |
| Observations | 5688 | 5688 | 5688 |

Note: Unit of observation is a month, city, and language (tourist country of origin) triplet. Sample includes tourists from Italy, France, and Germany to the 60 cities in Spain in May–October in 2010–2015. *Treatment* equals 1 for months after treatment for treated city-language pairs, and 0 otherwise. *Small page* equals 1 if the initial page size is below the 25th percentile, and 0 otherwise. *Large % added* equals 1 if text added to the page (as a % of the initial text in the page) is above the 75th percentile, and 0 otherwise. *Controls* include an indicator for period after treatment interacted with language fixed effects, an indicator for period after treatment interacted with city fixed effects, logarithm of number of tourists from Spain interacted with language fixed effects. Standard errors clustered by city-language pair (180 clusters).

Table 6: Robustness

| | (1)<br>Add<br>missing | (2)<br>Add<br>Dutch | (3)<br>All 12<br>months | (4)<br>12 months,<br>add missing | (5)<br>Add<br>UK | (6)<br>Share of<br>tourists |
|---|---|---|---|---|---|---|
| Treatment | 0.091** | 0.086* | 0.064 | 0.078** | 0.084* | 0.007* |
| | (0.045) | (0.047) | (0.041) | (0.039) | (0.043) | (0.004) |
| City-Language FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Log(Tourists from Spain) | No | Yes | Yes | No | No | No |
| Other controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Adj. R-squared | 0.052 | 0.212 | 0.265 | 0.002 | 0.104 | 0.026 |
| Observations | 5724 | 7584 | 9818 | 11448 | 5688 | 5688 |

Note: Repeats the regression in column (1) in table 5. In columns 1–5, dependent variable is logarithm of number of hotel nights of tourists from a given country (Germany, France, Italy). Column 1 substitutes missing observations by zeros (only for city-year pairs, when data exists for some month and tourist country of origin). Removes variables of number of tourists from Spain. Column 2 adds observations for tourists from the Netherlands, considers these all as non-treated. Column 3 adds remaining months. Column 4 adds remaining months and substitutes missing observations by zeros (only for city-year pairs, when data exists for some month & tourist country of origin), and removes variables of number of tourists from Spain. In column 5, adds logarithm of the number of tourists from UK interacted with language. In column 6, dependent variable is the number of tourists from country x divided by the number of tourists from country x plus from Spain and UK, and it removes variables of number of tourists from Spain.

Table 7: Dependent variable: Logarithm (number of hotel nights)

| | (1) | (2) |
|---|---|---|
| Log(PageViews) T-1 | 0.258*** | |
| | (0.047) | |
| GoogleTrend T-1 | | 0.010*** |
| | | (0.001) |
| City-Language FE | Yes | Yes |
| Adj. R-squared | 0.024 | 0.105 |
| Observations | 2871 | 2871 |

Note: Unit of observation is a month, city, and language (tourist country of origin) triplet. Sample is restricted to the control group. It includes tourists from Italy, France, and Germany to the 30 cities in Spain in May–October in 2010–2015. *Log(PageViews) T-1* is a one month lagged logarithm of Wikipedia page views of the article of the city in a given language. *GoogleTrend T-1* is a one month lagged Google Trend of the search term of the city from a given country. Standard errors clustered by city-language pair (90 clusters).

Table 8: Wikipedia page views and Google Trends

| | Log(Page Views) | | | Google Trends | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Treatment | 0.142*** | 0.073** | 0.078** | -0.180 | -0.415 | -0.317 |
| | (0.031) | (0.033) | (0.032) | (0.815) | (0.862) | (0.871) |
| Treatment: Small page | | 0.282*** | | | 0.892 | |
| | | (0.073) | | | (1.655) | |
| Treatment: Large % added | | | 0.249*** | | | 0.537 |
| | | | (0.070) | | | (1.634) |
| City-Language FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Adj. R-squared | 0.323 | 0.317 | 0.325 | 0.231 | 0.231 | 0.231 |
| Observations | 12709 | 12709 | 12709 | 12709 | 12709 | 12709 |

Note: In columns 1-3, dependent variable is logarithm of Wikipedia page views. In columns 4-5, dependent variable is Google Trend. Unit of observation is a month, city, and language (country) triplet. Sample includes 3 languages (countries): Italian, French, and German. Sample includes 60 cities in Spain. Time period is 2010–2015 excluding August 2014 (treatment month). *Treatment* equals 1 for months after treatment for treated city-language pairs, and 0 otherwise. *Small page* equals 1 if the initial page size is below the 25th percentile, and 0 otherwise. *Large % added* equals 1 if text added to the page (as a % of the initial text in the page) is above the 75th percentile, and 0 otherwise. *Controls* in all regressions include an indicator for period after treatment interacted with language fixed effects, an indicator for period after treatment interacted with city fixed effects. In columns 1-3, *Controls* include logarithm of page views in Spanish Wikipedia interacted with language fixed effects. In columns 4-6, *Controls* include Google Trends from Spain interacted with language fixed effects. Standard errors clustered by city-language pair (179 clusters).

Table 9: Robustness: Wikipedia page views and Google Trends

| | Page Views | | Google Trends | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Treatment | 18.676* | 0.150*** | -0.021 | -0.147 |
| | (9.665) | (0.037) | (0.029) | (0.829) |
| City-Language FE | Yes | Yes | Yes | Yes |
| Controls: Spanish-Spain | Yes | No | Yes | No |
| Controls: English-UK | No | Yes | No | Yes |
| Other controls | Yes | Yes | Yes | Yes |
| Adj. R-squared | 0.103 | 0.289 | 0.157 | 0.180 |
| Observations | 12709 | 12709 | 12709 | 12709 |

Note: The table largely repeats regressions in table 8. Dependent variable, in column 1, is Wikipedia page views, and in column 2, logarithm of Wikipedia page views. Dependent variable, in column 3, is logarithm of Google Trend, and in column 4, Google Trend. Unit of observation is a month, city, and language (country) triplet. Sample includes 3 languages (countries): Italian, French, and German. Sample includes 60 cities in Spain. Time period is 2010–2015 excluding August 2014 (treatment month). *Treatment* equals 1 for months after treatment for treated city-language pairs, and 0 otherwise. *Controls: Spanish-Spain* include either page views (column 1) in Spanish Wikipedia or logarithm of Google Trends (column 3) from Spain, all are interacted with language fixed effects. *Controls: English-UK* include either logarithm of page views in English Wikipedia (column 2) or Google Trend from UK (column 4), all are interacted with language fixed effects. *Other controls* include an indicator for period after treatment interacted with language fixed effects, an indicator for period after treatment interacted with city fixed effects. Standard errors clustered by city-language pair (179 clusters).

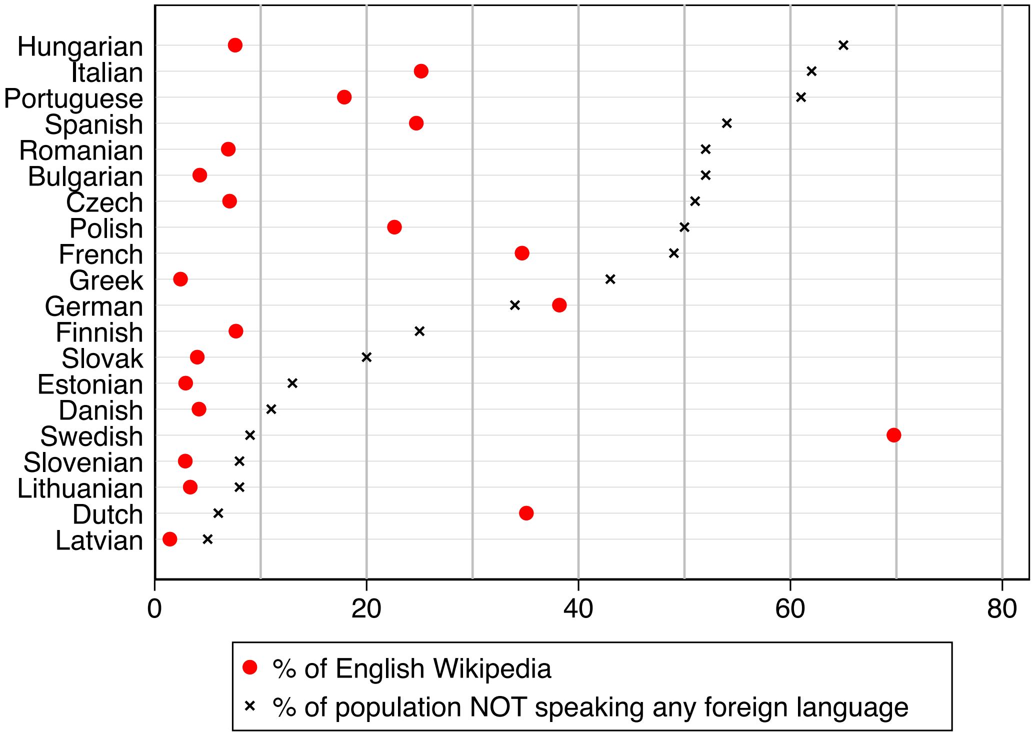

Figure 1: Size of Wikipedia and percentage of population not speaking any foreign language

Note: The size is measured by the number of articles in the local language Wikipedia as a percentage to the number of articles in English language Wikipedia. Data source for language skills is Eurobarometer (2012).

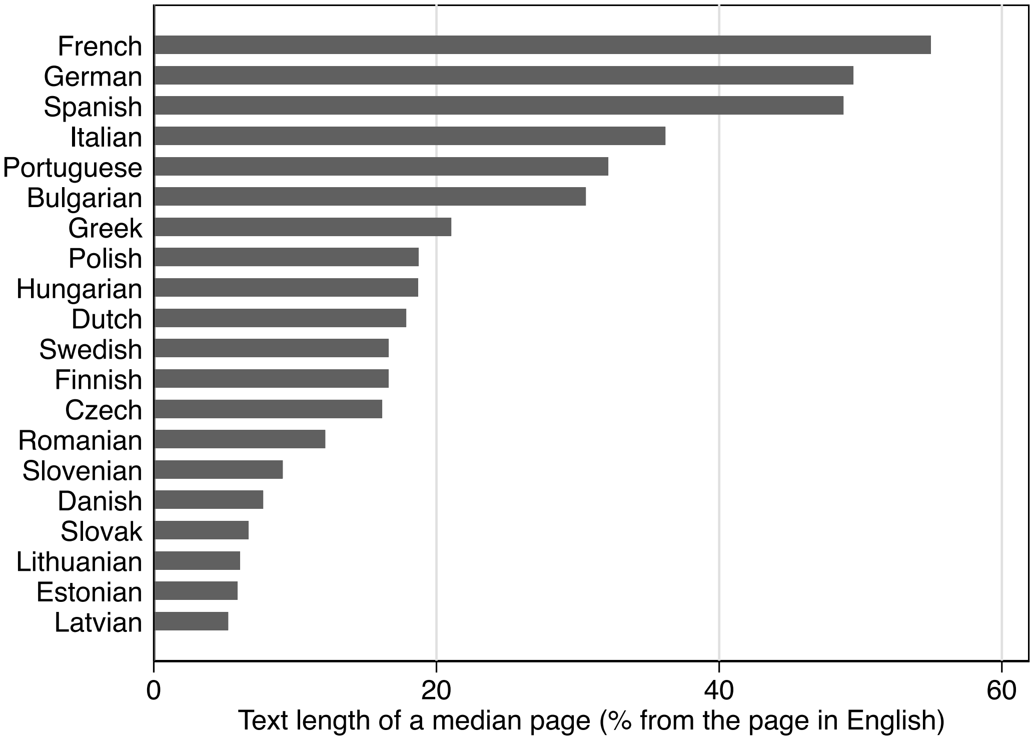

Figure 2: Median article length by language

Note: The sample includes pages in the list of 1000 vital articles chosen by Wikipedia community. For each page, the relative text length is calculated as the percentage of of the length of text in the local language Wikipedia compared to that of the English language Wikipedia edition. The graph presents the median of the relative text lengths by language.

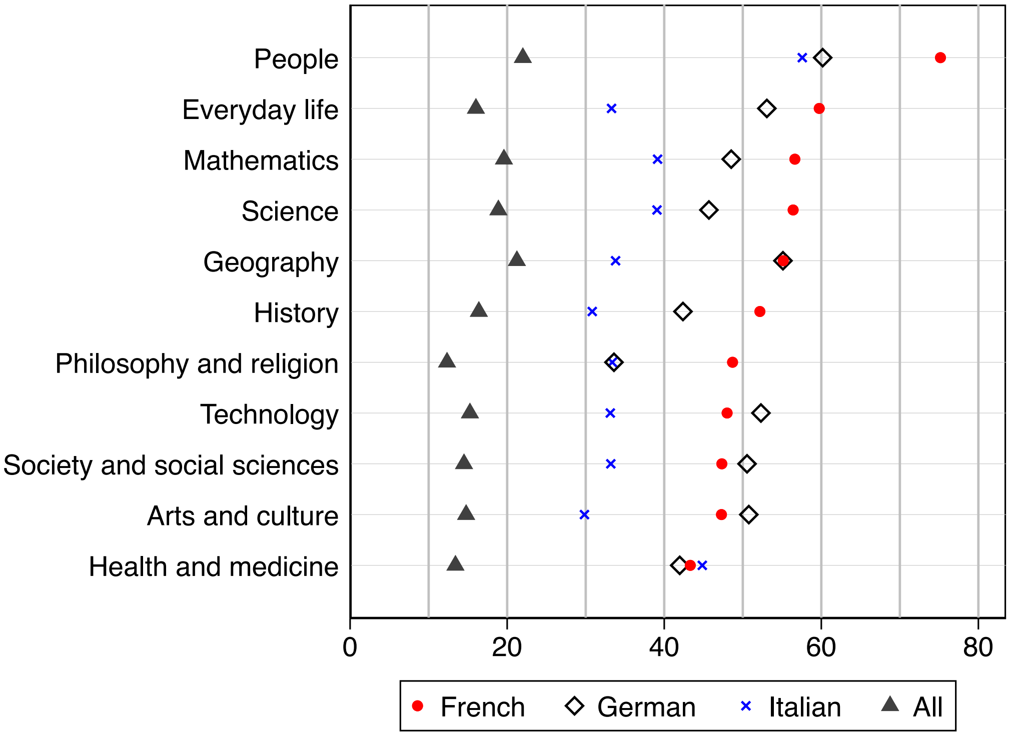

Figure 3: Median article length by topic

Note: The sample includes pages in the list of 1000 vital articles chosen by Wikipedia community. For each page, the relative text length is calculated as the percentage of of the length of text in the local language Wikipedia compared to that of the English language Wikipedia edition. The graph presents the median of the relative text lengths by article category. For each category, it presents the overall median and median by language (French, German, Italian).

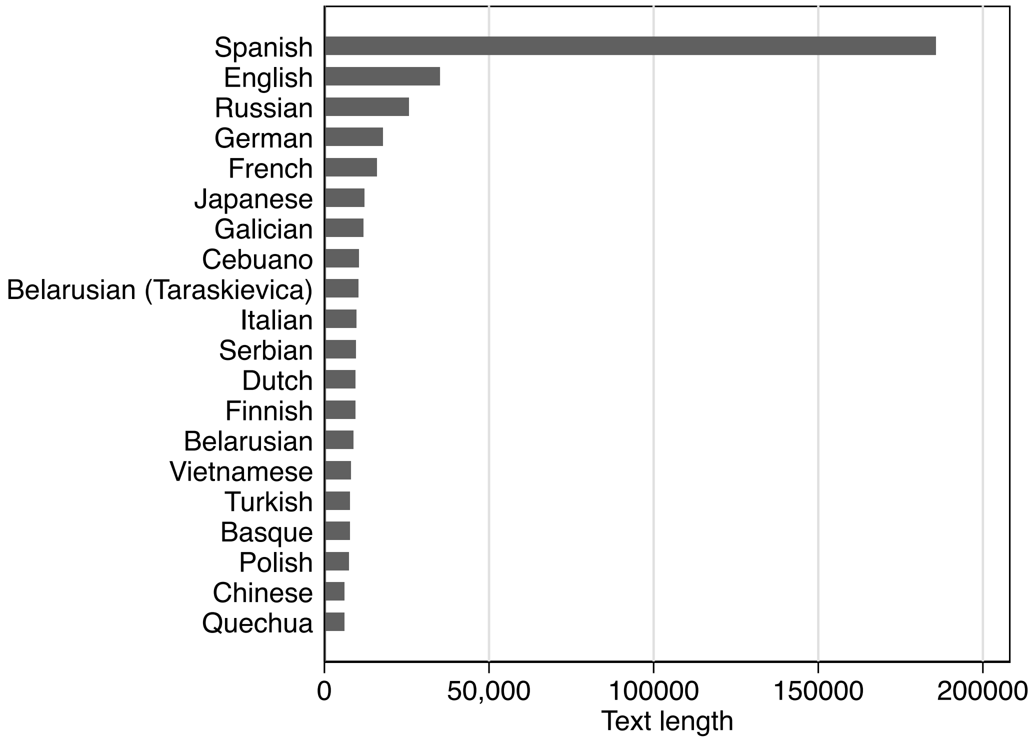

Figure 4: Length of a city page by Wikipedia language edition

Note: The page of the Spanish city exists in 84 Wikipedia language editions. Graph includes 20 languages in which the page is the longest.

| Revision | Text | Difference | Length |
|---|---|---|---|
| Pre-treatment | abc | | 3 |
| Post-treatment | adce | diff(abc,adce)=a~~b~~dce | Added 2 (de) |
| Survived | acef | diff(de,acef)=ace~~f~~ | Survived 1 |

Figure 5: Illustration how we used diff algorithm to quanitify the additions by treatment and the survival of the additions.

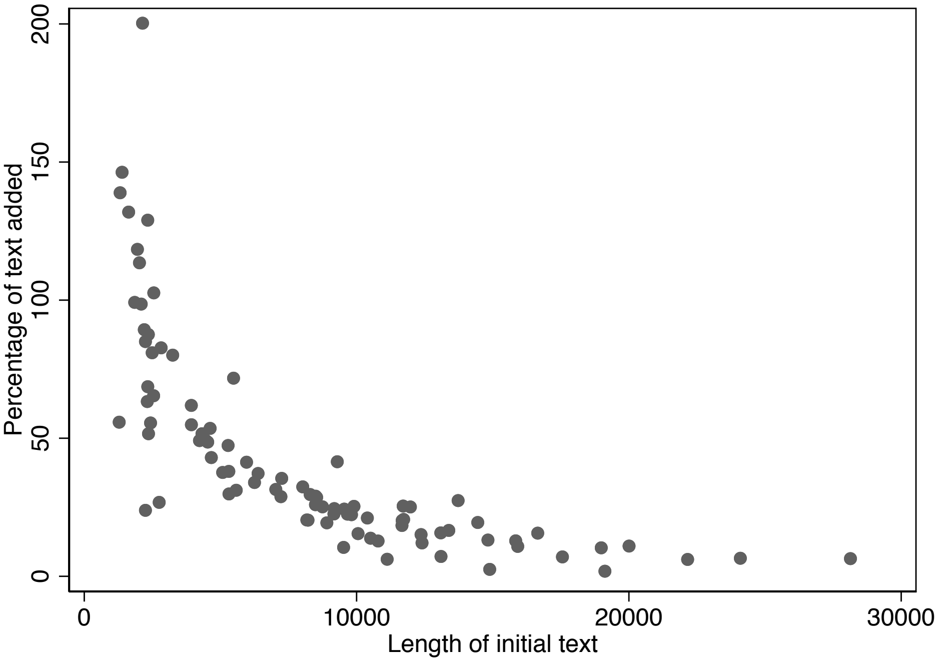

Figure 6: Length of text added (as % of initial text) vs length of initial text

Note: Unit of observation is a Wikipedia page in a given language (30 pages in each of the three languages: German, French, Italian). Sample includes treated pages.

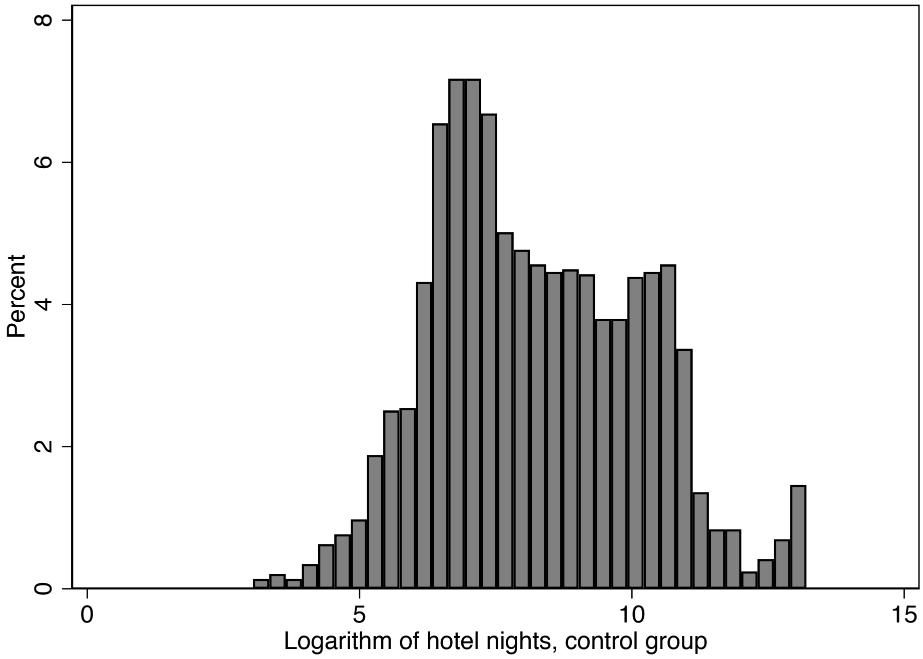

Figure 7: Logarithm of number of hotel nights in the control group

Note: Unit of observation is a month, city, and tourist country of origin triplet. Sample includes tourists from Italy, France, Germany to the 60 cities in Spain, but only the city-country of origin pairs, which were assigned to the control group. The time period of the sample is May–October in 2010 - 2015.

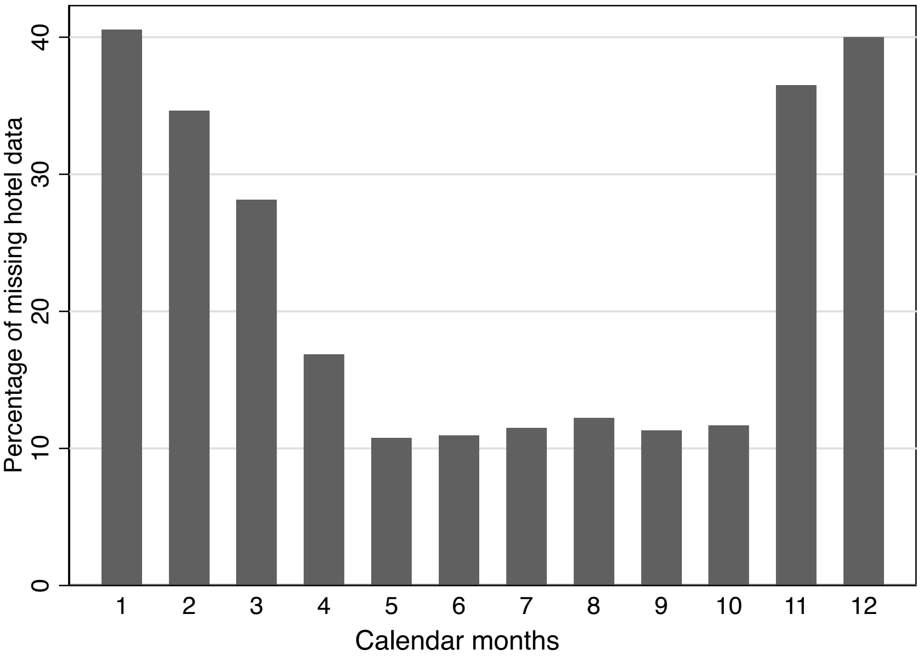

Figure 8: Percentage of missing hotel data, over 12 calendar months (January–December)

Note: Unit of observation is a month, city, and tourist country of origin triplet. Sample includes tourists from Italy, France, Germany to the 60 cities in Spain, but only city-country of origin pairs, which were assigned to the control group. The time period of the sample is 2010 - 2015.

# A Appendix: Additional tables and figures

Table A1: Wikipedia page length before treatment, by language

| | Initial text length | | | |
|---|---|---|---|---|
| | p25 | p50 | p75 | count |
| France | 2435 | 8336 | 13101 | 30 |
| Germany | 5483 | 9420 | 13387 | 30 |
| Italy | 2354 | 4974 | 8534 | 30 |
| Total | 2824 | 8098 | 11675 | 90 |

Note: Unit of observation is a city page in a given language Wikipedia. Sample includes pages in the treatment group.