Wikipedia Matters[*]

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

July 14, 2019

Abstract

We document a causal impact of online user-generated information on real-world economic outcomes. In particular, we conduct a randomized field experiment to test whether additional content on Wikipedia pages about cities affects tourists' choices of overnight visits. Our treatment of adding information to Wikipedia increases overnight stays in treated cities compared to non-treated cities. The impact is largely driven by improvements to shorter and relatively incomplete pages on Wikipedia. Our findings highlight the value of content in digital public goods for informing individual choices.

JEL: C93, H41, L17, L82, L83, L86

Keywords: field experiment, user-generated content, Wikipedia, tourism industry

[*] We are grateful to Irene Bertschek, Avi Goldfarb, Shane Greenstein, Tobias Kretschmer, Michael Luca, Thomas Niebel, Marianne Saam, Greg Veramendi, Joel Waldfogel, and Michael Zhang as well as seminar audiences at the Economics of Network Industries conference (Paris), ZEW Conference on the Economics of ICT (Mannheim), Advances with Field Experiments 2017 Conference (University of Chicago), 15th Annual Media Economics Workshop (Barcelona), Conference on Digital Experimentation (MIT), and Digital Economics Conference (Toulouse) 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 & University of East Anglia, `michael.kummer@econ.gatech.edu`

[¶] ZEW, `slivko@zew.de`

1 Introduction

Asymmetric information can hinder efficient economic activity (Akerlof, 1970). In recent decades, the Internet and new media have enabled greater access to information than ever before. However, the digital divide, language barriers, Internet censorship, and technological constraints still create inequalities in the amount of accessible information.[1] 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 reference. It is the fifth most popular website in the world[2] and receives about 14 billion direct page views per month.[3,4] However, the information available across Wikipedia's 299 language editions is not the same. We analyze whether the differences in available information affect consumption choices.

We quantify the causal impact of information in Wikipedia on consumption choices by conducting a randomized field experiment. Analyzing the impact of information using observational data would have been challenging because of potential endogeneity. Popular products tend to attract more attention, and therefore, more information is available about them. While the amount of information on Wikipedia tends to be correlated with the products' popularity, the information isn't necessarily causing consumption, but may instead be its byproduct. We overcome the identification problem using randomization.

We added content to randomly chosen Wikipedia pages in randomly chosen languages. We measured the outcome using data on tourists' overnight hotel stays in Spain. The Spanish tourism sector is important in itself by accounting for almost 5% of Spain's

[1] See Aceto and Pescapé (2015), Borgman (2003), Kralisch and Mandl (2006), Mason (2017), Van Dijk (2006), and Van Deursen and Helsper (2015).

[2] Alexa Internet. `http://www.alexa.com/siteinfo/wikipedia.org`, accessed September 23, 2017.

[3] Page Views for Wikipedia. Wikimedia Statistics. `https://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm`, accessed September 23, 2017.

[4] This does not include indirect uses such as Apple's Siri or Google.

GDP.[5] It also provided a good setting for the study, since the Spanish National Statistical Institute collects information about overnight stays in Spanish hotels at the level of city, month, and tourist country of origin.

Our treatment added text and photos to the Wikipedia pages of Spanish cities in different language editions of Wikipedia. Most of the added text was translated from Spanish Wikipedia. We focused on information that was relevant to tourists, such as the city's main sights and culture. We focused our attention on cities with rather short Wikipedia pages. The randomization was done across city and language pairs. By varying the information in different language editions of Wikipedia, we can isolate the causal impact on tourists' choices.

We find that information on Wikipedia has a sizable impact on consumption choices. Our estimates show that adding about 2,000 characters (approximately two paragraphs) of text and one photo to a city's Wikipedia page increased the number of nights spent in this city by about 9% during the tourist season compared to cities in the control group.[6] The effect comes mostly from pages that were initially relatively incomplete. In particular, the treatment increases hotel stays by about 33% in cities which initially had very short pages in a particular language, while there was no effect on city-language combinations where the pages were well developed.

Using data on search activity from Google Trends and readership from Wikipedia page views, we can shed some light on the mechanism that drives our findings. The added information has no significant impact on search activity outside Wikipedia but significantly increases the articles' readership. That is, more detailed Wikipedia articles gain more attention from potential readers. The size of this effect is similar in magnitude to the effect on tourists' choices.

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

[6] Our experiment doesn't allow us to distinguish between an absolute increase in demand (market expansion) and substitution between control and treatment (business stealing). Some of the effect likely arises from rerouting tourists from other cities. The implications we highlight in this paper hold in either case.

Our results have three policy implications, which are likely to reach beyond the setting of our experiment. First, the results have implications on economic inequality and the digital divide. Language can pose barriers that hinder efficient economic activity. Language barriers have slowed innovation (Peri, 2005), decreased trade (Anderson and van Wincoop, 2004), and affected investments (Grinblatt and Keloharju, 2001). In particular, languages create a major obstacle to access to information. Large differences remain across languages in terms of information available online. Our results imply that these differences may lead to significant differences in economic behavior between 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 entire tourism industry, the impact is large. In 2015, international tourists spent 270 million nights in Spain. The same year international travel receipts equaled 51 billion euros in Spain and 116 billion in the EU.[7] While we cannot say whether online user-generated content is changing the size of expenditures or reallocating them, its impact could be in billions of euros.

Third, on the microeconomic level, our results highlight the importance of online presence. A 9% increase in consumption as a result of additional user-generated information is substantial, given that each international tourist spends about 101 euros per day while visiting Spain on average (García-Sánchez et al., 2013). The findings suggest that it is beneficial to ensure that a city, firm, or product is accurately represented online in all relevant languages.

The results of this paper pose a puzzle—why is the online presence so limited? Increasing online presence is relatively inexpensive, while our results suggest a high return on investment. The online presence puzzle differs from most of the literature examining

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

contributions to online public goods.[8] This literature finds that contributions exceed what the economic theory would suggest. While the public goods literature assumes contributions are altruistic, we concentrate on a setting where the involved parties would benefit from making more information available.

Our paper makes three methodological contributions. First, it is among the first papers to use Wikipedia as a treatment in a field experiment for studying the impact on behavior outside Wikipedia.[9] Wikipedia provides a good ground for this since anyone can freely improve it and the whole process is automatically recorded in the form of revision histories.[10] Moreover, the readership of Wikipedia articles is well-recorded in the form of page views.

Second, we use a novel dataset of real-life outcomes—overnight hotel stays. Most importantly, this dataset provides a precise measure of demand of an identical product for consumers from different countries. In Spain, hotels are legally required to record guests' country of residence. We obtained the data from the Spanish National Statistical Institute aggregated to monthly level for each city and each country of origin. For example, we know how many nights German tourists spent in a particular city in July 2015. We use the fact that German tourists are more likely to get their information from German Wikipedia and Italian tourists from Italian Wikipedia to map consumption choices back to their potential information sources.

Finally, we make a technical contribution to analyzing Wikipedia's revision histories. As our treatment adds information to Wikipedia pages, which can then be changed by other Wikipedia users, the first step in the analysis is to see how much of our additions are modified by other Wikipedia users over time. For this, we use a diff algorithm describing the shortest sequence of additions and deletions of characters to change the original text to the revised one.[11] We apply this algorithm twice. First, to quantify which parts of

[8] See Lerner and Tirole (2003), Goldstein et al. (2008), Lacetera and Macis (2010), Ayres et al. (2013), Chen et al. (2010), Zhang and Zhu (2011), and Algan et al. (2013).

[9] There is literature examining the editing behavior in Wikipedia, which we will review below.

[10] Editing Wikipedia requires following Wikipedia's Terms of Use and policies.

[11] For a description of the algorithm, see Myers (1986).

the page our experiment added, and second, to measure how much of our additions had survived after a few months. We find that our edits are rather persistent: about 93% of our added text still existed about four months after the treatment. This could be because information on the pages we edited was relatively scarce and (hopefully) our contributions were considered sufficiently valuable by the Wikipedia community.

Our paper contributes to media economics literature studying the impact of media on economic outcomes (for an overview see DellaVigna and La Ferrara (2016)). In particular, our paper adds to studies on 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 studies on the impact of new media and online user-generated content.[12] Among others Chevalier and Mayzlin (2006) and Luca (2011) study how product reviews affect sales. Enikolopov et al. (2018) 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 informs consumers and affects their choices. 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.

Methodologically, our paper is related to a recent study by Thompson and Hanley (2017). In work concurrent and independent from ours, Thompson and Hanley (2017) also conduct a randomized field experiment in Wikipedia. They find that Wikipedia content affects scientific articles. Their work is complementary to ours—they find that

[12] More generally, our paper relates to the literature on how ICT affects economic outcomes by changing access to information. Among other topics, this literature has studied the impact of the Internet on economic growth (Czernich et al., 2011), on labor market outcomes (Forman et al., 2012; Akerman et al., 2015), on the airline industry (Dana and Orlov, 2014; Ater and Orlov, 2015), the impact of medical records on hospital costs (Dranove et al., 2014), and the impact of e-commerce on price dispersion (Ellison and Ellison, 2009; Overby and Forman, 2014).

Wikipedia has a significant impact on knowledge production outside Wikipedia, whereas we find that the available information affects consumption choices. Taken together, the two papers establish an important insight that is difficult to document without randomized field experiments: the value of information on Wikipedia does not only derive from the entertainment value that readers obtain from consuming the content, but it is also valuable input to economic choices. As the two papers illustrate, the content affects decisions in many domains.

Our paper also relates to the emerging small branch of literature on information production on 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 et al., 2016; Greenstein and Zhu, 2018; Hinnosaar, 2019). 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 an open-access Internet encyclopedia. It is the fifth most popular website in the world.[13] It is arguably one of the most important knowledge repositories and digital public goods. Wikipedia is written by volunteers: anyone can create Wikipedia articles or edit almost any of its existing articles.

While Wikipedia exists in 299 languages, the amount of available information differs across languages. English Wikipedia is the largest, with over five million articles. Only 13 other language editions had more than a million articles in 2017.[14]

Such asymmetries are important because a significant share of the population can access information only in their mother tongue. For example, almost half of the population

[13] 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).

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

in the EU does not speak any foreign language.[15] People who speak only one language can only access the information from their native language Wikipedia. Figure A.1 shows local language Wikipedia sizes and the percentage of the population speaking more than one language. Language affects not only the topics covered but also the depth of coverage. For example, among the 1,000 most important articles in Wikipedia[16] the median text length (relative to the corresponding page in English) varies from 5% in Latvian to 55% in French (see Figure A.2). Not all topics are covered equally (see Figure A.3). Overall, the worst covered topics are in categories like philosophy and religion (12%) and health and medicine (13%).

The relevant implication for this paper is that the available information varies across the language editions of Wikipedia, both in terms of the pages that exist and in terms of the depth of information in each topic it covers. Figure 1 presents an example: it describes pages about Murcia, a large Spanish city, across the different language editions of Wikipedia. This page exists in 84 different language editions of Wikipedia.[17] The figure contrasts the 20 longest versions of the Murcia page. Not surprisingly, the page is longest in Spanish Wikipedia. In all other languages the page is at least five times shorter.

3 Experimental Design

We conducted a field experiment in which we added content (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 overnight hotel stays by the tourists from the countries where the population speaks one of the treated languages. We describe the experimental design below. Appendix C provides additional detail.

[15] About 46% of the population speaks only their mother tongue. (cf. Eurobarometer (2012)).

[16] Wikipedia keeps a list of 1,000 vital articles (`https://en.wikipedia.org/wiki/Wikipedia:Vital_articles`, accessed June 26, 2017).

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

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 from 135 Spanish cities. However, in many smaller cities, hotel data was missing for some months and some tourist countries of origin. Hence, we expected to encounter the problem of not being able to measure the effect of treatment because of missing outcome (hotel) data. We were also concerned that our fixed length treatment might not be strong enough in the case of very large cities which already had long Wikipedia pages.

Therefore, we restricted attention to a sample of cities that satisfied two criteria. First, the Wikipedia page for the city had to be relatively short—no more than 24,000 characters in each of the four languages. Second, there could be no missing hotel data for the city. Specifically, we required the data on hotel stays to exist for each month from May to October 2013 and for all four countries of origin. Sixty cities satisfied these two criteria. These restrictions gave us a sample of 240 Wikipedia pages (or city-language pairs).

Randomization We randomized across 240 Wikipedia pages (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 city pages. 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 balance in the treatment and control groups, 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 of six cities each. Within each group, we randomly assigned the city to one of six treatments. The six treatments were as follows: treat the city page in one of the six possible language pairs (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 The pages were treated mid-August, 2014. The treatment added text and photos to each page in the treatment group. The added text and photos were on topics relevant for tourists, such as the city's main sights and culture. The added text was translated mostly from the corresponding Spanish or English language Wikipedia pages and photos were from the same source.

Our goal was to *improve* the Wikipedia pages, and we deliberately avoided decreasing the quality of Wikipedia pages, for example, by deleting existing material. Our treatment followed Wikipedia's policies and added content that according to our understanding was knowledge already approved by the editors of Spanish or English Wikipedia.

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 all or part of an article, or undo the latest changes by reverting to a previous version. 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, e.g. through Apple's Siri or Google information box. Therefore, we exclude all Dutch Wikipedia articles from our main specification. However, we include Dutch articles as non-treated in our robustness analysis and the results do not change much (see Table 4).

In table 1 we show that in the German, French, and Italian Wikipedias, our added text and photos survived well. (The methodology for measuring the survival of our additions is described in Section B.) Of the added text, on average 96 percent had survived by the beginning of the month following treatment and 93 percent by the beginning of the year following treatment. We interpret this in two ways. First, the edits were sufficiently persistent to provide hope that many people had seen the information our treatment added. Strictly speaking, it is not necessary that the precise wording of our treatment survives—it is to be expected that the other Wikipedia editors improve any added contributions over time in terms of wording, references, or content. However, measuring the preserved content is more difficult than measuring the actual text. Second, we hope that our treatment additions were considered useful by fellow Wikipedia editors; otherwise, they would have either reversed the edits or further revised them.

Descriptive statistics Tables 2 and A.2, and Figure A.4 provide descriptive statistics about the balance in our treatment. Table 2 shows that there were no significant differences in the main characteristics between the treatment and control groups. Table A.1 shows descriptive characteristics of the treatment. The median treatment added about 2,000 characters of text and one photo. The treatment added relatively more to pages that were initially shorter (see Figure A.4). Table A.2 describes the initial page length by language.

Next, we provide descriptive evidence about the outcome of interest. Figure A.5 presents the histogram of the logarithm of the number of hotel nights. It shows a considerable variation in the number of hotel nights. Figure A.6 represents the percentage of missing data by calendar month. As expected, hotel visits exhibit seasonality, with slightly above ten percent 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 that tourists from country (exposed to language) j spent in city i during month t. In our main analysis, we estimate the following difference-in-differences regression:

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

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 trends; 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 3 presents the main results. Column 1 shows our estimates of the treatment effect for the entire sample. According to these estimates, the treatment increases the number of hotel nights on average by 9%. 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 our treatment increases hotel stays by about 33% in cities where the pages were initially very short in a particular language, while there was no effect on cities with longer pages. 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. As the length of the text added was about the same, the treatment was relatively larger on initially short pages (Figure A.4). The results in Column 3 confirm that the effect is larger on pages where the treatment was relatively larger.

Robustness In Table 4, we analyze the robustness of our main result. Columns 1–5 repeat regression in Column 1 of Table 3, so the magnitudes of the estimates are comparable. Our main result is robust to (1) different handling of missing observations,

(2) including our data from the canceled Dutch experiment, (3) including the winter months, and (4) adding different control variables.

Column 1 substitutes missing observations by zeros (only for city-year pairs where 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. Recall that half of the city pages in Dutch Wikipedia were assigned to treatment, but editing Dutch Wikipedia proved infeasible since all pages in Dutch Wikipedia were returned to their pre-treatment state within 24 hours. In column 2, we include the Dutch observations and consider them as untreated. The results are unchanged. We could estimate the same regression and add a separate indicator variable for months after treatment only for Dutch pages assigned to treatment. The estimated treatment effect remains the same.

Columns 3 and 4 add the excluded months, and Column 4 substitutes missing observations by zeros.[18] In Column 4, again, the variables that measure the number of tourists from Spain are excluded. The results are similar, but in Column 3, less statistically precise. In Column 5 we test whether our results are driven by the choice of controls. We add additional controls, namely, the logarithm of the number of tourists from the UK interacted by language. The variables that measure the number of tourists from Spain are excluded. The results are similar.

In Column 6 we analyze whether our main result is sensitive to our choice of the dependent variable. In this column, the dependent variable is the number of tourists from country j divided by the number of tourists from country j plus those from Spain and the 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.

[18] We substituted missing observations only for city-year pairs when data exists for some month and tourist country of origin.

Mechanism We analyze the mechanism by which additional information on Wikipedia changes choices. We consider three main channels. First, additional information could increase the conversion rate. That is, it could lead to a larger *share* of readers choosing the destination. Second, the information could increase the *number* of readers. Third, it could increase the underlying *interest* in the destination via indirect effects, such as word-of-mouth. We proxy the third channel using data from Google Trends. Google Trends data measures how often a particular city is searched for on Google by the population of a specific country. We can measure the combination of the first two channels using data on the page views of Wikipedia articles. Unfortunately, we don't observe whether this reflects one person reading the page many times or many people reading it once.[19] Therefore, we cannot distinguish between a higher conversion rate and a broader audience.

Table 5 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 for city i in language j during month t. In Columns 4–6, the outcome variable is the Google Trend for city i from country j during month t. The estimates in Column 1 show that the treatment increased page views by about 11 percent. Column 2 separates the effect by the length of the article (before treatment), showing that the treatment effect is larger on shorter pages. Similarly, the regression results in Column 3 show that the treatment effect is larger on pages where our treatment added a relatively larger share of text (these tended to be shorter pages). The estimates in Columns 4–6 show that our treatment did not affect Google Trends (Google Search volume). Table A.3 verifies the robustness of these estimates.

Altogether, these results show that our treatment increases article readership, and the effect is similar in magnitude to the effect on the number of hotel nights. We find no evidence that the Google Search volume increased. We conclude that the added content on Wikipedia increased demand mostly through additional readership.

[19] Wikipedia did not collect unique page views before 2015. Therefore we cannot distinguish between new and returning readers.

One possible channel for this effect is that the additional information made it easier to find relevant information about treated cities. To investigate this hypothesis, we manually collected Google search rankings for each city/language pair. As the estimation results in Table D1 Appendix D show, on average the search rankings increased by 7.75 positions, moving the average city from the third to the second page in the search results. Unfortunately, we do not know what the search ranks were before our treatment, and how long it took until the treated pages achieved a better search rank.

Our main finding suggests an online presence puzzle: since a relatively small increase in available information in Wikipedia may have a significant impact on tourist flows, one would expect that interested parties, such as hotel employees, would be willing to improve the online presence of a particular destination. There could be several reasons why this does not happen more often. First, it requires computer skills and a clear understanding of the community guidelines, as well as knowledge of foreign languages, which may not be readily available and therefore makes the process very costly. Second, the information in Wikipedia is a public good, and the low appropriability of the investment in improving its content gives incentives for free-riding.

We study this mechanism more closely in Appendix F, where we look at the correlations between page lengths, city population, number of tourists, and the number of hotel employees. While this analysis cannot show a causal relationship, the results are consistent with the free-riding hypothesis. Cities with a larger population and with more tourists tend to have longer Wikipedia pages. Controlling for other variables, cities with fewer hotel employees, and therefore lower incentive to free-ride, have longer Wikipedia pages. This effect is statistically significant for Spanish and French Wikipedia but negligible in German, Italian, and Dutch languages.[20]

[20] According to Eurobarometer (2012), 7% of Spanish residents know French as a foreign language well enough to read newspaper or magazine articles. The same percentage is much lower for German (1%), Italian (2%), and Dutch (0%).

Limitations and future research Our study faces limitations and raises questions for future research. First, our experiment was not designed to distinguish between a substitution (business stealing) and an increase in overall interest (market expansion). We would expect that our estimated treatment effect is at least partly explained by substitution from other possible tourist destinations. It appears unlikely that more information about interesting destinations leads to a significant increase in the entire tourism sector. The implications highlighted in the paper apply regardless of this ambiguity, though it would be interesting to distinguish these two effects.

We study this issue carefully in Appendix E, where we use the fact that pages in the control group in French, German, and Italian languages were exposed to a potential "business-stealing" shock, i.e. some of their neighboring cities were treated in these languages. However, pages in the Dutch control group, as well as pages in other European Union languages, were not exposed to this shock. This comparison would be enough to identify the business-stealing effect. However, our results show that we do not have enough statistical power to estimate this. We cannot reject either that the estimates are driven purely by business-stealing nor that they are driven solely by market expansion.

Second, there is a question of generalizability, as the results may be specific to the types of pages and languages used in the experiment. In our sample, the Wikipedia pages were relatively short. We would expect that additional content would have less impact when the relative improvement is small. Moreover, the presence of short Wikipedia pages partly reflected the fact that these cities were not the most popular destinations. We would expect that the impact of Wikipedia is smaller in the case of major tourist attractions. On the other hand, these places were notable enough to have Wikipedia pages and to receive regular tourist flows. It is unlikely that additional information could lead tourists to destinations without exciting attractions. In the languages included in the experiment, Wikipedia editions are still among the largest with relatively large readerships. The availability of information in local languages is probably less relevant in countries where people are used to obtaining information in English. Additionally, the countries in the experiment send large tourist flows to Spain. This means there was already a preference for Spain and left room for substitution that was discussed above. The absolute level of the treatment effect is likely to be smaller in the case of languages and countries where Spain was not a popular tourist destination.

On a more positive note regarding generilizability, the impact of Wikipedia is unlikely to be specific to the tourism industry. Instead, we would expect that the information on Wikipedia affects choices and behavior in many domains.

Another natural question for further research is whether the additional content spurs additional organic content. This question was addressed in a separate paper that uses the data from the same experiment (Hinnosaar et al., 2019). This paper shows that the added material has a relatively small effect on future content growth, suggesting that edits to user-generated content should be undertaken solely based on their value rather than possible externalities.

Comparison with other results from the literature. Our estimates suggest that improved information in Wikipedia could lead to a 9% change in tourist choices of their destination. Other researchers in other settings have found various effect sizes. The only experimental work studying the effect of Wikipedia outside of Wikipedia is the concurrent and independent work by Thompson and Hanley (2017), who commissioned Wikipedia articles in Chemistry and Econometrics, but only uploaded a subsample of 88 randomly chosen articles, leaving the remaining articles as a control observation. They then compared text similarity to related scientific publications. They find that the Wikipedia articles influenced the language in the associated articles significantly: 1 in 830 words was influenced by the language in Wikipedia. While it is difficult to compare the magnitude of this effect with our results, their findings show that scientists use Wikipedia as a reference in their writing. Moreover, as there are spillover effects from scientific writing to other economic behaviors, these effects may be sizable.

The literature has also found sizable effects of media on consumption and financial behavior. Bursztyn and Cantoni (2015) studied the impact of TV advertising using differential access of Eastern Germans to Western German TV before reunification. According to their estimates, an exposure of one more minute of advertising per day resulted in 1.5% increase in consumption of advertised categories after the reunification. Berg and Zia (2017) used a sample of 1,000 individuals in South Africa and encouraged half of them to watch a soap opera that had a subplot on debt-management. Treated individuals were almost twice as likely to borrow from formal channels and less likely to engage in gambling.

Finally, there is significant evidence that consumer reviews have a large impact on consumer behavior. Luca (2011) shows that an additional star in Yelp ratings increases sales by about 5% for independent restaurants (and finds no impacts on chains). Similarly, Anderson and Magruder (2012) find that an additional half-star in Yelp ratings causes restaurants to sell out 49% more often.

5 Discussion

We found a significant causal impact of user-generated content on Wikipedia on real-life choices. The estimated effect suggests that a well-targeted two-paragraph improvement of Wikipedia may lead to a 9% increase in tourists' overnight visits. The median monthly number of hotel nights spent by tourists from the three effectively treated countries to the cities in the control group was about 3,000 (during the six months from May to October). This implies an increase of about 270 nights per month. Even if there were no tourists in the remaining six months, this implies about 1,600 additional hotel nights per year.

What are the implications for the local economy? According to recent estimates (García-Sánchez et al., 2013), each international tourist visiting Spain spends about 101 euros per day on average. Back-of-the-envelope calculations suggest that improving a city's Wikipedia page can lead to approximately 160,000 euros of additional revenue per year. This implies a considerable impact on local hotels and the overall local tourist industry.

Our results highlight the importance of online presence. Ensuring that a city, firm, or product is accurately represented in online information sources of all relevant languages is relatively cheap, i.e. almost free or a few hundred dollars in mainly one-time costs. In comparison, the 9%-increase in demand is rather large, suggesting a high return to investment.

Finally, the amount of information available in different languages varies significantly. Our results imply that this may lead to large differences in economic decisions and economic outcomes as well. This opens up a more general discussion about economic inequality and the digital divide across cultural and ethnic groups.

References

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

Aceto, G. and A. Pescapé (2015): "Internet Censorship Detection: A Survey," *Computer Networks*, 83, 381–421.

Akerlof, G. A. (1970): "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, 84, 488–500.

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

Algan, Y., Y. Benkler, M. Fuster Morell, and J. Hergueux (2013): "Cooperation in a Peer Production Economy Experimental Evidence from Wikipedia." *manuscript*.

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

Anderson, M. and J. Magruder (2012): "Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database," *Economic Journal*, 122, 957–989.

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

Ayres, I., S. Raseman, and A. Shih (2013): "Evidence from two large field experiments that peer comparison feedback can reduce residential energy usage," *Journal of Law, Economics, and Organization*, 29, 992–1022.

Berg, G. and B. Zia (2017): "Harnessing Emotional Connections to Improve Financial Decisions: Evaluating the Impact of Financial Education in Mainstream Media," *Journal of the European Economic Association*, 15, 1025–1055.

Borgman, C. L. (2003): *From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World*, MIT Press.

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

Chen, Y., F. M. Harper, J. Konstan, and S. X. Li (2010): "Social Comparisons and Contributions to Online Communities: A Field Experiment on Movielens," *American Economic Review*, 100, 1358–98.

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

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

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

DellaVigna, S. and E. La Ferrara (2016): “Economic and Social Impacts of the Media,” in *Handbook of Media Economics*, ed. by S. Anderson, D. Stromberg, and J. Waldfogel, Amsterdam: Elsevier.

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, 239–270.

Ellison, G. and S. F. Ellison (2009): “Search, Obfuscation, and Price Elasticities on the Internet,” *Econometrica*, 77, 427–452.

Enikolopov, R., M. Petrova, and K. Sonin (2018): “Social Media and Corruption,” *American Economic Journal: Applied Economics*, 10, 150–74.

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, 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, 613–630.

Goldstein, N. J., R. B. Cialdini, and V. Griskevicius (2008): “A room with a viewpoint: Using social norms to motivate environmental conservation in hotels,” *Journal of Consumer Research*, 35, 472–482.

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, 343–348.

——— (2018): “Do Experts or Crowd-based Models Produce More Bias? Evidence from Encyclopedia Britannica and Wikipedia,” *MIS Quarterly*, 42, 945–959.

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

Hinnosaar, M. (2019): “Gender Inequality in New Media: Evidence from Wikipedia,” *Journal of Economic Behavior & Organization*, 163, 262–276.

Hinnosaar, M., T. Hinnosaar, M. Kummer, and O. Slivko (2019): “Externalities in Knowledge Production: Evidence from a Randomized Field Experiment,” *manuscript*.

Kralisch, A. and T. Mandl (2006): “Barriers to Information Access Across Languages on the Internet: Network and Language Effects,” in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS’06)*, IEEE, vol. 3, 54b–54b.

Lacetera, N. and M. Macis (2010): “Social Image Concerns and Prosocial Behavior: Field Evidence from a Nonlinear Incentive Scheme,” *Journal of Economic Behavior & Organization*, 76, 225–237.

Lerner, J. and J. Tirole (2003): “Some Simple Economics of Open Source,” *Journal of Industrial Economics*, 50, 197–234.

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

Mason, R. O. (2017): “Four Ethical Issues of the Information Age,” in *Computer Ethics*, Routledge, 41–48.

Myers, E. W. (1986): “AnO(ND) Difference Algorithm and its Variations,” *Algorithmica*, 1, 251–266.

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

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

Thompson, N. and D. Hanley (2017): "Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial," *manuscript*.

Van Deursen, A. J. and E. J. Helsper (2015): "The Third-Level Digital Divide: Who Benefits Most from being Online?" in *Communication and Information Technologies Annual*, Emerald Group Publishing Limited, 29–52.

Van Dijk, J. A. (2006): "Digital Divide Research, Achievements and Shortcomings," *Poetics*, 34, 221–235.

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

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

Figures

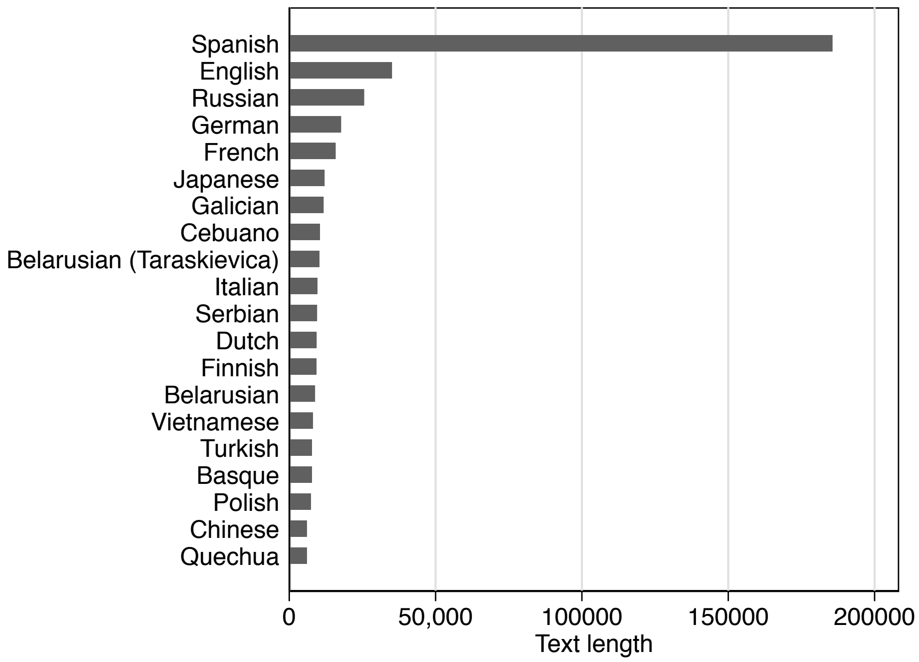

Figure 1: Length of a city page by Wikipedia language edition

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

Tables

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

	French	German	Italian	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: Covariate balance table

	Coefficient	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: 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 (in parentheses) are clustered by city-language pair (180 clusters). *** Indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

Table 4: Robustness

	(1)	(2)	(3)	(4)	(5)	(6)
	Add	Add	All 12	12 months,	Add	Share of
	missing	Dutch	months	add missing	UK	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 3. 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. Standard errors (in parentheses) are clustered by city-language pair. *** Indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

Table 5: Wikipedia page views and Google Trends

	Log(Page Views)			Google Trends		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.116***	0.070**	0.069**	-0.180	-0.415	-0.317
	(0.030)	(0.033)	(0.032)	(0.815)	(0.862)	(0.871)
Treatment: Small page		0.219***			0.892	
		(0.073)			(1.655)	
Treatment: Large % added			0.183***			0.537
			(0.069)			(1.634)
City-Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Adj. R-squared	0.581	0.566	0.582	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 (in parentheses) are clustered by city-language pair (179 clusters). *** Indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

Online Appendices

A Additional Tables and Figures

Table A.1: 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 A.2: 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.

Table A.3: Robustness: Wikipedia page views and Google Trends

	Page Views		Google Trends	
	(1)	(2)	(3)	(4)
	Add English	Share of views	Add UK	Share of trend
Treatment	0.153***	0.011***	-0.147	0.000
	(0.047)	(0.004)	(0.829)	(0.005)
City-Language FE	Yes	Yes	Yes	Yes
Controls: English-UK	Yes	No	Yes	No
Other controls	Yes	Yes	Yes	Yes
Adj. R-squared	0.379	0.101	0.180	0.009
Observations	12709	12575	12709	12709

Note: The table largely repeats regressions in table 5. Dependent variable, in column 1, is logarithm of Wikipedia page views, and in column 2, the number of page views of the article in language x divided by the sum of the number of page views of English, Spanish, and language x. Dependent variable, in column 3, is Google Trend, and in column 4, Google Trend from country x divided by the sum of Google trends from UK, Spain, and country x. 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: English-UK* include either logarithm of page views in English Wikipedia (column 1) or Google Trend from UK (column 3), 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 (in parentheses) are clustered by city-language pair (179 clusters). *** Indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

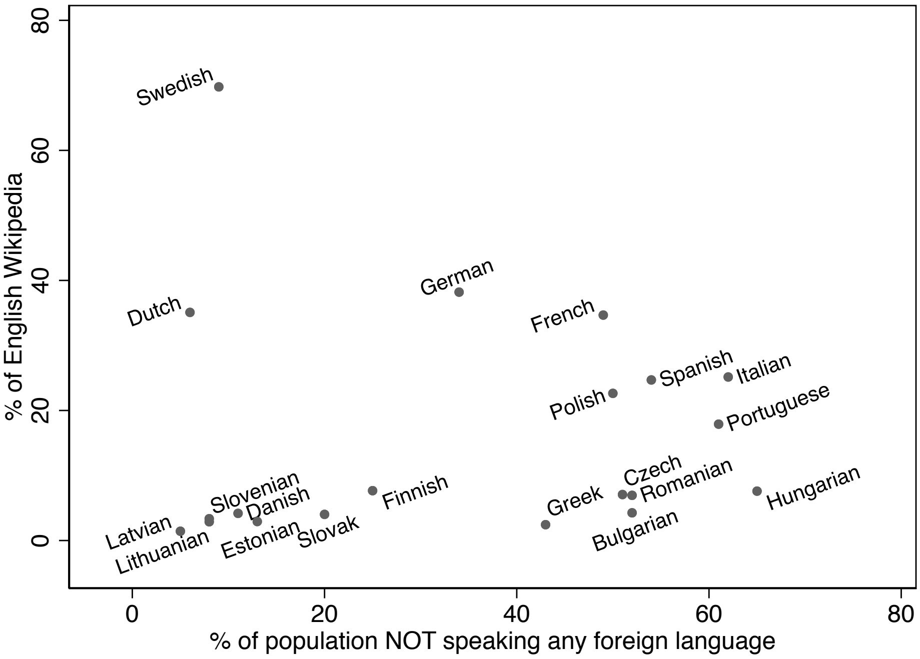

Figure A.1: Size of Wikipedia and percentage of population not speaking any foreign language

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).

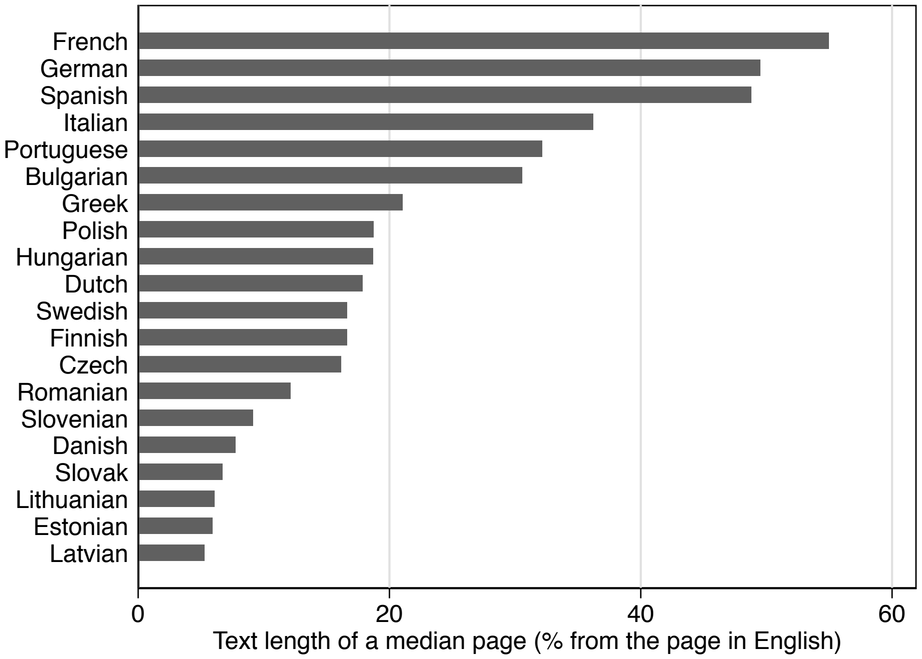

Figure A.2: Median article length by language

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 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.

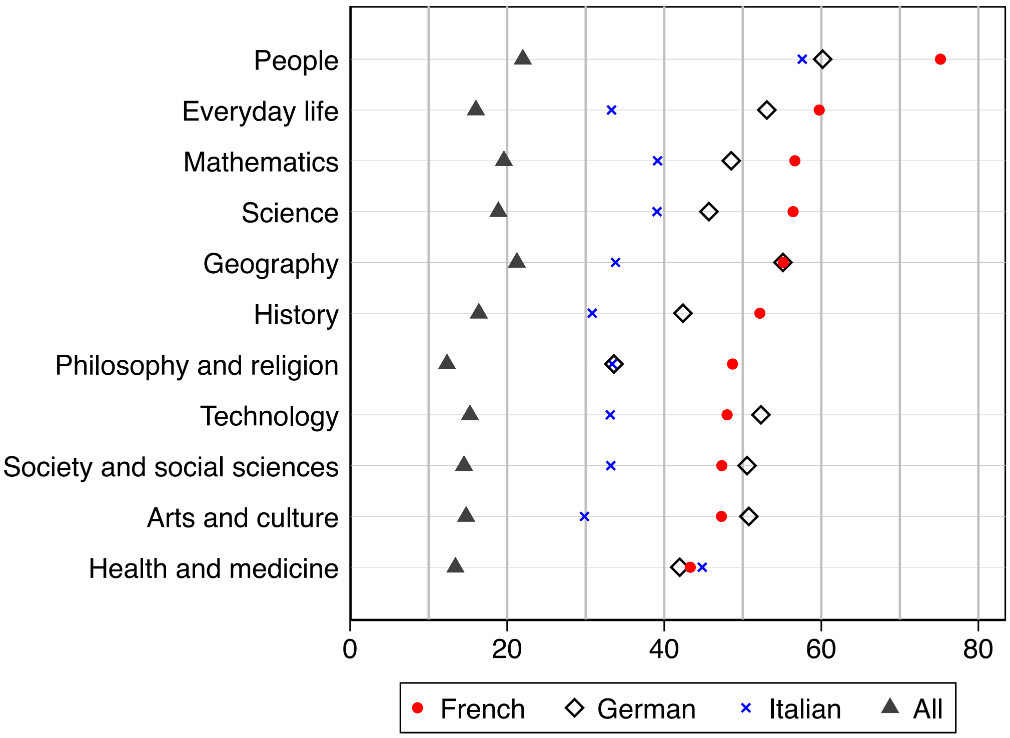

Figure A.3: Median article length by topic

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 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).

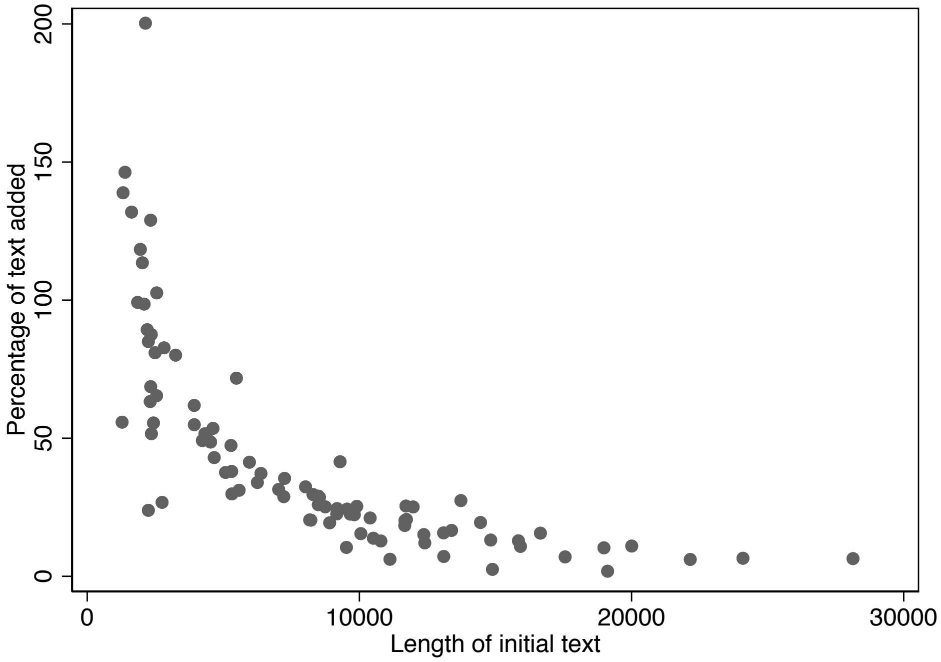

Figure A.4: Length of text added (as % of initial text) vs length of initial text

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.

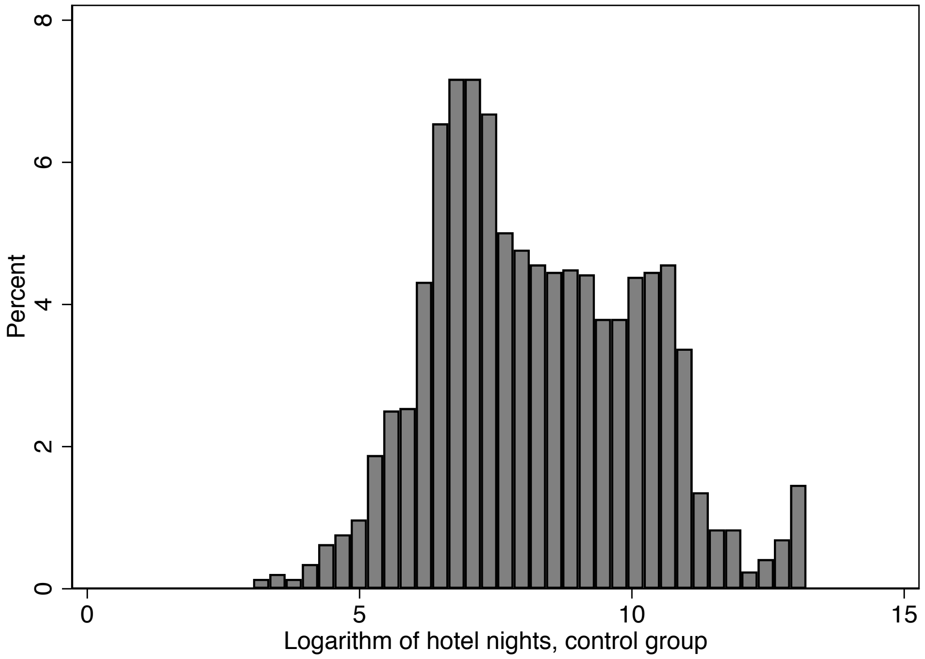

Figure A.5: Logarithm of number of hotel nights in the control group

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.

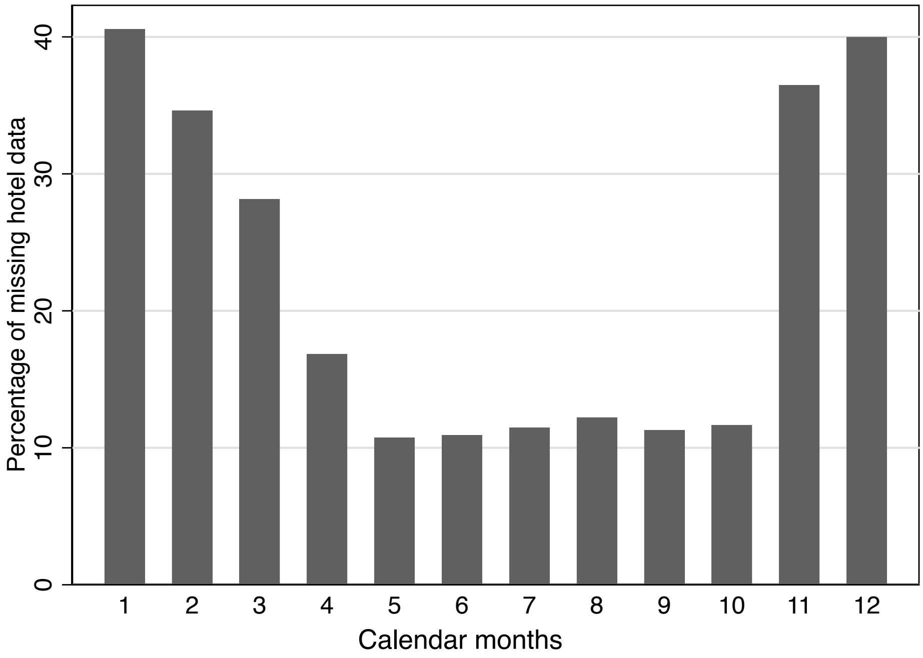

Figure A.6: Percentage of missing hotel data, over 12 calendar months (January–December)

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.

B Measuring Our Treatment and Its Survival

We applied a 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.[21] See figure B1 for illustration.

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 B1: Illustration how we used diff algorithm to quanitify the additions by treatment and the survival of the additions.

[21] It is a 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.

C Implementation of the Experiment

Selection of pages and randomization: We chose the Spanish cities according to the availability of data at the level of the city, month, and tourist country of origin from INE (Spanish National Institute of Statistics). The data set included monthly information on nights spent in hotels of 135 Spanish cities, where the city is included in the sample if it has at least three non-empty hotels. The overnight hotel stays in this data set separate tourists by the country of origin, which was crucial for our experimental design. The dataset reported the number of visitors from the United Kingdom, Japan, United States, Germany, Belgium, France, Italy, Netherlands, Portugal; grouping the rest of the nationalities under the category “Other”. Out of this set of countries, we picked four for our treatment: German, French, Italian, and Dutch. These countries all have a language that is strongly associated with the population of the country. Moreover, at least one member of our research team could verify the quality of texts in these languages.

Out of the 135 cities, we chose those with the pages shorter than 24,000 symbols in each language of treatment and available data (no missing values) for the period May to October 2013. Due to these criteria, the largest cities, including the capitals of provinces, for example, Madrid and Barcelona, as well as the smallest cities, with many missing values in the Summer 2013 period, were not included in the experimental setting. On the sample of remaining cities, we implemented the stratified randomization. We ordered the cities by the number of tourists’ overnight stays and divided into groups of 6 cities, such that every city in the group receives one of the 6 treatments: 1) treat Wikipedia pages of the city in German & French, 2) in German & Italian, 3) in German & Dutch, 4) French & Italian, 5) French & Dutch, 6) Italian & Dutch. Overall, we had 10 groups of cities. 60 cities were treated, each in 2 languages, which resulted in a sample of 120 treated city-language pairs and 120 controls.

Texts for treatment: To create texts for treatment of Wikipedia articles, we proceeded in three steps. In the first step, our research assistants with knowledge of English and Spanish languages compared city pages in Spanish and English Wikipedia to find information that would be valuable for visitors of the city, especially in the following sections:

- Introduction
- Climate
- Main sights / Landmarks and Culture (including Festivals) / Culture and Sights / Tourism
- Museums, Churches, Castles, Other Points of Interest
- Transport / Traffic
- Airport, Train, Road, Boats

The information in Wikipedia city pages follows a standard structure and the pages of the Spanish cities are much more complete in Spanish than in other languages. For example, if we compare the Spanish and English articles on Benalmadena, we see many similarities in the composition of some sections: history, geography, sights.[22] However, the details provided about sights are very short, almost bullet point, in the English version, while being detailed in the Spanish version. Moreover, the Spanish version contains not only the sights but also cultural events related to the city, which could also be relevant to English-speaking readers. Wikipedia pages in other languages, such as French or Italian, contain even fewer details than the English page.

Overall, research assistants were asked to find 3–4 paragraphs of information missing, which is about A4 page size. To simplify the translation to treatment languages, they translated the parts of this information that were available only in Spanish to English.

As the second step, we hired translators for each of the four treatment languages (French, German, Italian, and Dutch). We gave them the information in English collected

[22] See `https://es.wikipedia.org/wiki/Benalm%C3%A1dena` for the Spanish page and `https://en.wikipedia.org/wiki/Benalm%C3%A1dena` for the English page in Wikipedia.

in the first step as well the information from the Wikipedia target page. Their task was to translate parts from the material in English that were missing in the target language so that the missing information could be added to Wikipedia pages in the corresponding treatment languages. Further requirements were as follows:

- The text can contain only correct information. The added text must be polished and easily readable.
- The text should not repeat already existing information.
- The translation does not have to be a word-by-word translation. The names of the places and monuments etc. should be correct.
- The whole text did not need to look like a fully coherent story but could be a collection of paragraphs, where each paragraph provides information about some topic.

As a result of this procedure, for each Wikipedia article covering a Spanish city in a determined language of treatment, we had some new information to be uploaded to the article. The final step was to upload the resulting text to Wikipedia pages, carefully merging it with the existing text.

Uploading texts: For uploading the new information to the corresponding city-language articles on Wikipedia, we created 16 user accounts. Each account was randomly assigned to a set of cities for uploading the content. The newly created content, as well as two photos from the Spanish version of the article, were then uploaded to Wikipedia in mid-August, 2014.

Community reaction: In the first day after the treatment, we checked the articles, and in French and Italian language, no problems occurred. In German Wikipedia, an administrator contacted us and asked about the purpose of the content uploads and whether we have any commercial interest. We replied, providing information about our institution and the research group, and about the research project. This response was sufficient. In Dutch Wikipedia, a single editor reverted all our edits within 24 hours. We discussed the issue with this editor, but the editor was not willing to accept our contributions and therefore we decided to stop the uploads on the Dutch Wikipedia and accept the reversions.

D Quantifying Search Costs

In this appendix, we study the impact of the treatment on Google Search Rank of a particular city. This data was collected in June 2019 by submitting searches in the format "Turismo di Elche" (for each city in Italian, German, and French languages) and recording the position of the first result from Wikipedia related to a particular location. For example, for this specific search, the first result appeared in position 14 (second page, position number four).

Table D1 shows that search ranks affiliated with the pages in the treated group are on average about 7–9 positions higher than in the control group (which on average appear at position 22.4). We interpret the increased visibility in search results as a reduction of search costs.

Table D1: Dependent variable: Google Search Rank

	(1)	(2)
Treatment	-9.200***	-7.725**
	(3.472)	(3.647)
City FE	No	Yes
Language FE	No	Yes
Mean dependent variable	22.433	22.433
Adj. R-squared	0.033	0.051
Observations	180	180

Note: The unit of observation is a city and language pair. The dependent variable is Google Search Rank for searches in the format "Turismo di Elche" in Italian, French, and German. Treatment equals one for treated city-language pairs and zeroes otherwise. Standard errors (in parentheses), *** Indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

E Business Stealing vs. Market Expansion

A natural question to ask is how much of our results come from business stealing and how much from market expansion. For example, our results are consistent with 4.5% business stealing, i.e. rerouting tourists from the cities in the control group to the cities the treatment group. But they are also consistent with 9% market expansion, i.e. increasing the overall number of tourists to the medium-sized Spanish cities. Perhaps more plausibly, it could be a combination of these two effects.

Our experiment was not designed to decompose these two effects. However, we can study this issue by using the fact that the pages in the control group in French, German, and Italian languages were exposed to a potential "business-stealing" shock, i.e. some of their neighboring cities were treated in these languages. The pages in the Dutch control group (where the experiment was canceled), as well as pages in other European Union languages, were not exposed to such shock. By comparing these two groups, we could identify the business-stealing effect.[23]

To estimate it, we run the following regression, which is similar to our main difference-in-difference specification:

$$\begin{aligned}\log(Nights_{ijt}) = {} & \beta_0 + \beta_1 \cdot FranceGermanyItaly_{ij} \cdot PostTreatment_t \\ & + \beta_2 \cdot PostTreatment_t + \beta_3 \cdot log(TouristsFromSpain_{ijt}) \\ & + \beta_4 \cdot log(TouristsToBarcelonaMadrid_{ijt}) + CityCountryFE_{ij} + \varepsilon_{ijt}\end{aligned} \tag{2}$$

The sample consists of two parts: first, we take all observations in our control group, which are untreated city-language pairs in French, German, and Italian, for which the variable $FranceGermanyItaly_{ij}$ equals one; and second, control group in Dutch language (Column 1) or all other European Union countries (Column 2), for which $FranceGermanyItaly_{ij}$ equals zero. The coefficient of interest is β_1, which should capture the business-stealing

[23] We are grateful to the anonymous referee for this suggestion.

effect, i.e. by how much the number of tourists in our control group is lower than we would predict by looking at tourist flows from other similar countries. To control for the city-specific time trends, we include the logarithm of the number of tourists from Spain going to the same city. To control for the country-of-origin-specific time trends, we include logarithm of the number of tourists from the same country going to Barcelona and Madrid (which are cities outside our sample). For the other European countries, due to privacy-related data limitations, we have data only for the UK, Portugal, Belgium, and then the remaining European Union countries altogether. We exclude Belgium because it was partly treated by the treatment of French pages.

Table E1 reports our estimation results from this regression. Although the point estimates would imply a sizable business-stealing effect of 2.6–5%, there is not enough data to precisely estimate the coefficients. In particular, we are unable to reject either of the two extremes: that our main result is purely driven by business stealing or that it is solely driven by market expansion (i.e. reject business stealing).

Table E1: Measuring business stealing. Dependent variable: Logarithm (number of hotel nights)

	Control groups of France, Germany, Italy, Netherlands	Added other European Union countries
	(1)	(2)
France, Germany, Italy × Post treatment	-0.026	-0.050
	(0.076)	(0.048)
Post treatment	0.090	0.127***
	(0.063)	(0.024)
Log(Tourists from Spain)	0.363***	0.398***
	(0.038)	(0.034)
Log(Tourists to Barcelona, Madrid)	0.802***	0.682***
	(0.074)	(0.065)
City-Country FE	Yes	Yes
Adj. R-squared	0.191	0.150
City-country pairs	120	300
Observations	3794	9293

Note: The unit of observation is a month, city, and tourist country of origin triplet. The sample includes tourists to the 60 cities in Spain in May–October in 2010–2015. In column 1, the sample includes tourists from the control groups in France, Germany, Italy, and the Netherlands. Column 2 adds to the sample tourists from the UK, Portugal, and the remaining EU countries (except Belgium) going to the 60 cities. *PostTreatment* equals 1 for the time periods post-experiment. *Log(TouristsFromSpain)* is the logarithm of the number of tourists from Spain going to the same city. *Log(TouristsToBarcelonaMadrid)* is the logarithm of the number of tourists from the same country going to Barcelona, Madrid. All regressions include fixed effects for each city and country of origin pair. Standard errors (in parentheses) are clustered by city-country pair. *** Indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

F Descriptive Analysis of Page Lengths

In this appendix, we look at the descriptive statistics of page lengths of Wikipedia city pages in different languages. We check their correlations with explanatory variables such as the population of the city, the number of tourists, and also the number of hotel employees. This analysis allows us to address one possible explanation for the online presence puzzle—interested parties, such as hotel managers, could consider the costs of increasing digital presence and the appropriability of the returns on this investment. As the information on Wikipedia is a public good, there is an incentive to free-ride and wait for someone else to contribute. For a fixed population size and the number of tourists, the free-riding hypothesis would suggest that cities with fewer hotel employees should have longer Wikipedia pages. To address this question, we run the following regression.

$$\begin{aligned} \log(PageLength_{ij}) = \beta_0 + \beta_1 \log(HotelEmployees_i) + \beta_2 \log(Nights_{ij}) \\ + \beta_3 \log(Population_i) + \varepsilon_{ij} \end{aligned} \tag{3}$$

The variable $PageLength_{ij}$ is the length of the Wikipedia page of city i in language j, $HotelEmployees_i$ is the total number of hotel employees in city i, $Nights_{ij}$ is the number of overnight stays by visitors from country j in city i, and $Population_i$ is the population of city i. We use values from August 2013 for all variables. The coefficient of interest is β_1, which would be negative under the free-riding hypothesis and zero otherwise.

Table F1 presents the regression estimates. While the analysis does not prove a causal relationship, it is consistent with natural explanations. Cities with a larger population have longer Wikipedia pages, but this effect is statistically significant only for Spanish Wikipedia. Cities with more visitors from a particular country have longer pages in their language, but again the effect is significant only in the Spanish language.

Importantly, controlling for the population and the number of tourists, the cities with fewer hotel employees, have longer Wikipedia pages. This effect is statistically significant only in the Spanish and French languages, which could be explained by language skills, but also by the fact that due to data limitations for the number of hotel employees we have a relatively small number of observations. Altogether, this provides suggestive evidence in favor of the free-riding hypothesis.

Table F1: Dependent variable: Logarithm (Wikipedia page length)

	(1)	(2)	(3)	(4)	(5)
	Pages in	Pages in	Pages in	Pages in	Pages in
	German	Italian	French	Dutch	Spanish
Log # Hotel Employees	-0.088	-0.309	-0.690**	-0.305	-0.829***
	(0.206)	(0.194)	(0.279)	(0.269)	(0.145)
Log Tourist Overnight Stays	0.034	-0.019	0.236	0.168	
	(0.107)	(0.156)	(0.196)	(0.150)	
Log Spanish Overnight Stays					0.651***
					(0.173)
Log Population	0.059	0.104	0.127	0.074	0.250***
	(0.077)	(0.066)	(0.085)	(0.064)	(0.076)
Observations	35	32	28	25	38
Adj. R-squared	-.071	.29	.27	-.049	.6

Note: The unit of observation is a city page in a given language Wikipedia in August 2013, one year before the treatment. Each column samples Wikipedia pages in German, Italian, French, Dutch, and Spanish, correspondingly, and shows whether the page length is correlated to the hotel industry measures to account for potential strategic considerations of the hotel managers in contributing to the online presence of the Spanish cities. The dependent variable is the *logarithm of the number of symbols* that a Wikipedia page contains. In columns (1)–(4) this measure of page length includes only the clean texts, while in column (5) the length of Spanish the pages contains Wikipedia markup, due to data collection. *Log Tourist Overnight Stays* measures the number of tourists coming to the city from Germany, Italy, France, Netherlands, and Spanish residents, correspondingly. *Log # Hotel Employees* measures the total hotel employees in each city. *Log Population* controls for the size of the city measured by the number of inhabitants in 2012. Standard errors (in parentheses), *** Indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.