Wikipedia matters

Marit Hinnosaar\textsuperscript{1,2,3} | Toomas Hinnosaar\textsuperscript{1,2} | Michael Kummer\textsuperscript{4,5,6} | Olga Slivko\textsuperscript{7}

\textsuperscript{1}University of Nottingham, Nottingham, United Kingdom
\textsuperscript{2}Collegio Carlo Alberto, Turin, Italy
\textsuperscript{3}CEPR, London, United Kingdom
\textsuperscript{4}University of East Anglia, Norwich, United Kingdom
\textsuperscript{5}Georgia Institute of Technology, Atlanta, Georgia, United States
\textsuperscript{6}ZEW, Mannheim, Germany
\textsuperscript{7}Rotterdam School of Management, Erasmus University, Rotterdam, The Netherlands

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 nontreated cities. The impact is largely driven by improvements to shorter and relatively incomplete pages on Wikipedia. Our findings highlight the value of digital public goods for informing individual choices.

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 (see Aceto & Pescapé, 2015; Borgman, 2003; Kralisch & Mandl, 2006; Mason, 2017; Van Deursen & Helsper, 2015; Van Dijk, 2006). How much does it matter for economic outcomes?

In this paper, we analyze the causal impact of online information on real-world economic outcomes. Specifically, we measure the impact of information on one of the primary economic decisions—consumption. For the source of online information, we focus on Wikipedia. It is one of the most important online sources of reference. It is the fifth most popular web site in the world\textsuperscript{1} and receives about 14 billion direct page views per month.\textsuperscript{2} 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. Because of potential endogeneity, analyzing the impact of information using observational data is challenging. Popular products tend to attract more attention, and therefore, more information is available about them. Although the amount of information on Wikipedia tends to be correlated with the products’ popularity, the information
is not necessarily causing consumption, but may instead be its byproduct. We overcome this identification problem by conducting a randomized field experiment.

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 GDP. 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 also focused our attention on cities with rather short Wikipedia pages. We randomized treatments 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 2000 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 with cities in the control group. 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.

Our results have policy implications, which are likely to reach beyond the setting of our experiment. First, 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 (Garcia-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. This result poses 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 contributions to online public goods (see Algan et al., 2013; Ayres et al., 2013; Chen et al., 2010; Goldstein et al., 2008; Lacetera & Macis, 2010; Lerner & Tirole, 2003; Zhang & Zhu, 2011). This literature finds that contributions exceed what the economic theory would suggest. Although the online public goods literature assumes contributions are altruistic, we concentrate on a setting where the involved parties would benefit from making more information available.

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. Although 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, 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 (J. E. Anderson & van Wincoop, 2004), and affected investments (Grinblatt & 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.

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. 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. Moreover, the readership of Wikipedia articles is well-recorded in the form of page views.

Second, we use a novel data set of real-life outcomes—overnight hotel stays. Most importantly, this data set 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. We apply this algorithm twice. First, to quantify which parts of 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 & 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. Among others, Chevalier and Mayzlin (2006) and Luca (2016) 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 (2018). In work concurrent and independent from ours, Thompson and Hanley (2018) 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 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 Aaltonen & Seiler, 2015; Zhang & Zhu, 2011) and biases in Wikipedia (Greenstein et al., 2016; Greenstein & Zhu, 2012, 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 web site in the world. 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.

Although 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. 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 in the EU does not speak any foreign language. It is costly to read texts in languages that the person does not speak, which means that monolingual people mostly get information from their native language Wikipedia. Supporting Information Figure A1 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 1000 most important articles in Wikipedia the median text length (relative to the corresponding page in English) varies from 5% in Latvian to 55% in French (see Supporting Information Figure A2). Supporting Information Figure A3 further illustrates that the relative depth varies by topic. For example, French, German, and Italian editions of Wikipedia relative to English have much more in-depth coverage of notable people than the vital topics in health and medicine. 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. 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 in a particular city by tourists from a particular country. We describe the sample and the experimental design below. Supporting Information Appendix C provides additional detail.

3.1 | 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 were 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).

3.2 | 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.

### 3.3 Treatment

The pages were treated in 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.

### 3.4 Survival of added material

Although 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 h. 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, for example, 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 nontreated 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 Supporting Information Appendix B) Of the added text, on average 96% had survived by the beginning of the month following treatment and 93% 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

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Survival over time of text and photos which we added to Wikipedia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>French</td>
</tr>
<tr>
<td>% text survived: 24 h</td>
<td>100.0</td>
</tr>
<tr>
<td>% text survived: next month</td>
<td>98.7</td>
</tr>
<tr>
<td>% text survived: next year</td>
<td>95.1</td>
</tr>
<tr>
<td>% photos survived: 24 h</td>
<td>100.0</td>
</tr>
<tr>
<td>% photos survived: next month</td>
<td>100.0</td>
</tr>
<tr>
<td>% photos survived: next year</td>
<td>100.0</td>
</tr>
<tr>
<td>Number of observations</td>
<td>30</td>
</tr>
</tbody>
</table>

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 h, by the beginning of the next calendar month after treatment, by the beginning of the next calendar year after treatment.
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.

### 3.5 Descriptive statistics

Table 2 and Supporting Information Table A2 and Figure A4 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. Supporting Information Table A1 shows descriptive characteristics of the 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 Supporting Information Figure A4). Supporting Information Table A2 describes the initial page length by language.

Next, we describe the outcome of interest. Supporting Information Figure A5 presents the histogram of the logarithm of the number of hotel nights. It shows a considerable variation in the number of hotel nights. Supporting Information Figure A6 represents the percentage of missing data by calendar month. As expected, hotel visits exhibit seasonality, with slightly above 10% missing data from May to October and up to 40% missing data in December and January.

### 4 RESULTS

#### 4.1 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:

$$\log (\text{Nights}_{jit}) = \alpha + \beta \text{Treatment}_{jit} + \gamma X_{jit} + \text{CityLanguageFE}_{ij} + \varepsilon_{jit}. \quad (1)$$

The variable of interest $\text{Treatment}$ 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 $\text{CityLanguageFE}_{ij}$ and time varying control variables, $X_{jit}$. 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.\textsuperscript{16} The interaction with language fixed effects captures the possibility that the correlation of preferences of tourists from different countries with those of preferences of Spanish visitors may be different.\textsuperscript{17} 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.

#### TABLE 2 Covariate balance table

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Coefficient</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log(Sum of tourists in 2013)</td>
<td>$-0.002$</td>
<td>.958</td>
</tr>
<tr>
<td>Log(Number of tourists)</td>
<td>$-0.012$</td>
<td>.527</td>
</tr>
<tr>
<td>Tourist data missing</td>
<td>$0.045$</td>
<td>.556</td>
</tr>
<tr>
<td>Log(Initial text length)</td>
<td>$-0.000$</td>
<td>.994</td>
</tr>
</tbody>
</table>

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: $\text{TreatmentGroup}_{i} = \text{Constant} + \beta \text{Variable}_{i} + \varepsilon_{i}$, 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 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 with the initial text length. As the length of the text added was about the same, the treatment was relatively larger on initially short pages (Supporting Information Figure A4). The results in column 3 confirm that the effect is larger on pages where the treatment was relatively larger.

### 4.2 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 with the initial text length. As the length of the text added was about the same, the treatment was relatively larger on initially short pages (Supporting Information Figure A4). The results in column 3 confirm that the effect is larger on pages where the treatment was relatively larger.

### 4.3 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 pretreatment state within 24 h. In column 2, we include the Dutch observations and consider them as untreated. The results are unchanged. If we would estimate the same regression and add a separate indicator variable for months after treatment only for Dutch pages assigned to treatment, then the estimated treatment effect would remain the same.

Columns 3 and 4 add the excluded months, and column 4 substitutes missing observations by zeros. In column 4, again, the variables that measure the number of tourists from Spain are excluded. The results are similar, but in column...
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 \( j \) plus those from Spain and UK. Again, the variables that measure the number of tourists from Spain are excluded. Although the results are not comparable in magnitude, the treatment effect is positive and statistically significant.

### 4.4 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 cannot observe whether this reflects one person reading the page many times or many people reading it once. 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%. 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). Supporting Information Table A3 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.
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 Supporting Information Table D1 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 (i.e., relying on others to do the work).

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. Although 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 than cities with more hotel employees. This effect is statistically significant for Spanish and French Wikipedia but negligible in German, Italian, and Dutch languages.

<table>
<thead>
<tr>
<th>TABLE 5</th>
<th>Wikipedia page views and Google Trends</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Log(Page Views)</td>
</tr>
<tr>
<td></td>
<td>(1) (2) (3)</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.116*** 0.070** 0.069***</td>
</tr>
<tr>
<td></td>
<td>(0.030) (0.033) (0.032)</td>
</tr>
<tr>
<td>Treatment: Small page</td>
<td>0.219***</td>
</tr>
<tr>
<td></td>
<td>(0.073)</td>
</tr>
<tr>
<td>Treatment: Large % added</td>
<td>0.183***</td>
</tr>
<tr>
<td></td>
<td>(0.069)</td>
</tr>
<tr>
<td>City-Language FE</td>
<td>Yes</td>
</tr>
<tr>
<td>Adj. $R^2$</td>
<td>0.581</td>
</tr>
<tr>
<td>Observations</td>
<td>12,709</td>
</tr>
</tbody>
</table>

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 and 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). *** and ** indicate significance at 1% and 5% percent level.

4.5 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 Supporting Information 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, that is, 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 generalizability, 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., 2020). 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.

4.6 | 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 (2018), 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. Although 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, to the extent that there are spillovers 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 1000 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 (2016) shows that an additional star in Yelp ratings increases sales by about 5% for independent restaurants (and finds no impacts on chains). Similarly, M. 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 3000 (during the 6 months from May to October). A treatment effect of 9% implies an increase of about 270 nights per month. Even if there were no tourists in the remaining 6 months, this implies about 1600 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, that is, 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.

ACKNOWLEDGMENTS
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.

ENDNOTES
3This does not include indirect uses such as Apple’s Siri or Google.
5Our experiment does not allow us to distinguish between an absolute increase in demand (market expansion) and substitution between control and treatment (business stealing). Some of the effects likely arises from rerouting tourists from other cities. The implications we highlight in this paper hold in either case.
6There is literature examining the editing behavior in Wikipedia, which we will review below.
7Editing Wikipedia requires following Wikipedia’s Terms of Use and policies.
8For a description of the algorithm, see Myers (1986).
9More 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 (Akerman et al., 2015; Forman et al., 2012), on the airline industry (Ater & Orlov, 2015; Dana & Orlov, 2014), the impact of medical records on hospital costs (Dranove et al., 2014), and the impact of e-commerce on price dispersion (Ellison & Ellison, 2009; Overby & Forman, 2014).
10Only 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).
12About 46% of the population speaks only their mother tongue. (cf. Eurobarometer, 2012).
14Wikipedia data on Murcia was accessed on June 20, 2017.
15An alternative would be to ignore the fact that all Dutch treatments were removed at the same time and to estimate an intention to treat effect. Then the estimated effect would be slightly smaller and less precise.
16A possible concern with this control, as well as the rest of our analysis, is that there could be general equilibrium effects. For example, if the treatment increased the number of tourists from France, this may either raise prices or reduce the capacity for Spanish and other
visitors. If there is a positive treatment effect and general equilibrium effects are sizeable, our estimates may be biased. Our analysis is conducted under the assumption that such general equilibrium effects are negligible.

17 If instead we are worried that the correlation of preferences of Spanish tourists and foreigners varies across cities, then we could add interactions with city fixed effects. The estimated treatment effect remains similar.

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

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

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%).

ORCID
Marit Hinnosaar https://orcid.org/0000-0002-0936-0104
Toomas Hinnosaar http://orcid.org/0000-0002-5256-2481
Michael Kummer https://orcid.org/0000-0002-9664-5895
Olga Slivko https://orcid.org/0000-0001-6062-2870

REFERENCES


**SUPPORTING INFORMATION**

Additional Supporting Information may be found online in the supporting information tab for this article.