Discussion Paper No. 15-089

Does Wikipedia Matter? The Effect of Wikipedia on Tourist Choices

Marit Hinnosaar, Toomas Hinnosaar,
Michael Kummer, and Olga Slivko

ZEW

Zentrum für Europäische
Wirtschaftsforschung GmbH

Centre for European
Economic Research

Discussion Paper No. 15-089

Does Wikipedia Matter?
The Effect of Wikipedia on Tourist Choices

Marit Hinnosaar, Toomas Hinnosaar,
Michael Kummer, and Olga Slivko

Does Wikipedia Matter? The Effect of Wikipedia on Tourist Choices[*]

Marit Hinnosaar [†]
Collegio Carlo Alberto

Toomas Hinnosaar[‡]
Collegio Carlo Alberto

Michael Kummer [§]
Georgia Institute of Technology

Olga Slivko [¶]
Centre for European Economic Research (ZEW)

December 22, 2015

Abstract

We analyze the impact of information on Wikipedia on tourists' choices for travel destinations. Our results suggest a strong observational correlation between the amount of content on Wikipedia and tourist overnight stays. We propose a check of whether this correlation is causal. For that, we introduce randomized exogenous variation to articles' content. While our treatment is strong enough to affect content on the treated pages positively, we find no statistically significant effect of this treatment on tourist overnight stays.

JEL Classification: D29, D80, H41, J60, L17.

Keywords: Wikipedia, randomized treatment, online content, tourist visits.

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

[†] Via Real Collegio, 30; 10024 Moncalieri (Torino), Italy. marit.hinnosaar@carloalberto.org

[‡] Via Real Collegio, 30; 10024 Moncalieri (Torino), Italy. toomas.hinnosaar@carloalberto.org

[§] 221 Bobby Dodd Way, Room 222, Atlanta, GA 30332, US. michael.kummer@econ.gatech.edu.

[¶] P.O. Box 103443, D–68034 Mannheim, Germany. slivko@zew.de.

1 Introduction

Online platforms accumulating and freely providing knowledge have gained importance in all spheres of life. One of the most representative knowledge repositories is the world's largest encyclopedia Wikipedia, which has been among the top ten most popular web-sites for several years. As a result, Wikipedia has become a standard reference to check facts, and the wide dissemination of this online platform raises the question whether Wikipedia has now an impact on individual choices, and, therefore, on the economy.

In this paper, we ask whether information in Wikipedia affects travellers' choices of their preferred destinations and tourist attractions. For that purpose, we combine data on tourism from four European countries to 210 Spanish cities and over six years with variables that describe the content of the city's Wikipedia articles. Our content indicators allow us to distinguish between text (symbols, paragraphs, words) and illustrations (pictures). The four Wikipedia languages are French, German, Italian and Dutch and we use monthly overnight stays during the years 2008-2014. We match the information on tourist overnight stays in the cities (by country of origin) with the content of the Wikipedia articles for these cities such that the language in which an article is written corresponds to the country of the tourists' origin.

We perform two types of analysis. Firstly, we analyze the panel dataset that results from our data collection. This analysis has a substantial limitation. One of the core content policies of Wikipedia states that it does not publish original ideas: all material in Wikipedia must come from a reliable source. Therefore, Wikipedia aims at accurate representation of material that is available elsewhere at that point of time. Because of that, it is difficult to distinguish the impact of additional information in Wikipedia from the impact of additional information elsewhere. To tackle this issue, we introduce randomized exogenous variation in the content of 120 articles by adding some information about Spanish cities, which was present in the Spanish version of the articles but was missing in one of our languages of interest (Dutch, French, Italian, German).

We observe significant correlations between content and tourist visits. The largest effect is for the number of words in an article: 1,000 additional words about a given city in a given language are related to a 4 per cent increase in overnight stays by tourists from the corresponding country. On the contrary, in our setting for causal identification, we find no effect of an increase in content on tourist visits. This result could indicate the absence of true causal impact of content on individual choices. However, a potential explanation for the lack of statistical significance could be an insufficiently large number of treated observations.

The rest of the paper is structured as follows. Section 2 presents an overview of the literature related to our research question. Section 3 discusses the empirical approach and Section 4 describes the dataset. Section 5 conducts the empirical analyses. Section 6 discusses the obtained results, limitations and concludes.

2 Literature

There is evidence that Wikipedia is becoming a standard reference source. However, according to our knowledge not much has been said about Wikipedia's impact on behavior. The popularity of Wikipedia (6th most visited website) is a clear indication that many people are interested in the content. For example, Laurent and Vickers (2009) analyze the search engine rankings and page view statistics of health-related keywords and conclude that Wikipedia is a relevant source for health information.

Since the content in Wikipedia can be generated by anyone who wishes to contribute, doubts have been raised on the reliability of the content. Adler et al. (2008) propose a system that calculates values of trust for the text in Wikipedia articles which, in a way, give an indication of text reliability. While individuals who trust the information on Wikipedia might be expected to adjust their behavior to the information they find, this is not a direct implication. Especially if Wikipedia is not consulted for choice-relevant information, but only for more general knowledge, a direct impact of the information on behavior seems unlikely.

Our paper contributes to the literature that documents the economic impact of different online platforms. The most well-known economic paper that finds an impact of Wikipedia on individual behavior was provided by Xu and Zhang (2013). It studies the impact of Wikipedia on investment decisions in the market and finds that the effect works through moderating bad news. There are studies that analyze the impact of other online sources. Several studies have addressed the impact of online reviews on demand and found positive effects (Chevalier and Mayzlin (2006); Luca (2011)). Acquisti and Fong (2013) used an experimental approach to study the impact of Facebook on discrimination in hiring. Similar to these studies, we rely on a controlled field experiment in order to evaluate the potential impact of information from Wikipedia on individual decisions.

3 Empirical Strategy

Our goal is to examine whether the content that is available on Wikipedia affects tourist choices. This relationship between tourist visits and content on Wikipedia is analyzed for Spanish cities and for visitors from four countries: France, Germany, Italy and the Netherlands. We will first document a high surface correlation between content and visits, analyzing the correlations in the cross section and focusing on one fixed point in time. This analysis is of the form:

$$(Visits)_{ij} = \mu_{ij} + \beta_0 * (WikiContent_{ij}) + \gamma * X_{ij} + \epsilon_{ij}. \tag{1}$$

We analyze this relationship in July 2008, 2010 and 2013 separately. The outcome of interest is $Visits_{ij}$, where i is the index for the Spanish city and j the visitor's country of origin. The independent variable of interest is $WikiContent_{ij}$, which is measured by different variables such as text length, pictures or the number of paragraphs. The control variables in X_{ij} are the dummies for the country of the tourist's origin.

A high correlation between content and visits almost certainly overstates the causal impact of Wikipedia on tourist choices. The two most important concerns would be (i) unobserved heterogeneity (two places might differ in what they have to offer to visitors, or they might be of different cultural or natural importance) and (ii) potential reverse causality (tourists edit Wikipedia after their visit).

To uncover the underlying relationship, we use two frameworks for the analysis. First, we use fixed effects panel regression which exploits variation within the number of tourists from a given country of origin who visit a specific town. This approach eliminates unobserved heterogeneity. Moreover we provide an analysis of the temporal pattern of content creation and tourist visits to uncover patterns of precedence. Second, we attempted a randomized controlled field experiment, in which we identified potential edits and then randomly selected which ones to actually perform.

Panel Fixed Effects Estimation:

The "contemporaneous" fixed effects regression takes the following form:

$$(Visits)_{ijt} = \mu_{ij} + \nu_t + \beta_1 * (WikiContent_{ijt}) + \epsilon_{ijt}. \tag{2}$$

The dependent variable of interest is $Visits_{ijt}$, which measures tourist visits from county j to city i in month t. We use a city-country fixed effect (μ_{ij}) and monthly dummies (ν_t, we control for year and month). Similarly, we use a lagged fixed effects regression to analyze the temporal dynamics more carefully:

$$(Visits)_{ijt} = \mu_{ij} + \nu_t + \beta_2 * (WikiContent_{ij,t-\ell}) + \epsilon_{ijt}. \tag{3}$$

The only difference between equation 3 and equation 2 is the changed focus on the lagged values of the content in Wikipedia, which is denoted by the lag-index ℓ. The new explanatory variable $WikiContent_{ij,t-\ell}$ is the available content ℓ months before the current period of observation.

Randomized Controlled Field Experiment:

Beyond the correlational analysis we attempted a controlled randomized intervention to influence the available information on Wikipedia. This design allows us to identify causal effects via a difference-in-differences approach if the scale of the experiment large enough.

The difference-in-differences regression is:

$$(Visits)_{ijt} = \beta \ After_t + \gamma \ (After_t \times \ Affected_{ij}) + \mu_{ij} + \nu_t + \ \epsilon_{ijt}. \tag{4}$$

$After_t$ and $Affected_i$ are dummy variables. $Affected_i$ separates the city-country pairs that we treated from the untreated ones. $After_t$ equals one if the time period is after t_0, the period when we treated the chosen Wikipedia articles. The variable $Affected_i$ should not matter because of our randomization. Moreover it remains constant over time for each observation and can hence not be estimated separately from the fixed effect specific to a city-language combination. The coefficient of interest is γ for the crossterm $After_t \times \ Affected_{ij}$ of these two dummies, which measures the difference-in-differences. Note that by observing ij-pairs we can control for visits from country y to city x with visits from country z to the *same* city x.

4 Data

Our dataset is collated from different sources. Whereas the information on the amount of content has been obtained from the corresponding Wikipedia pages, the information on the tourist stays has been obtained from the records maintained by Spanish National

Statistical Institute.[1] We have panel data on the tourist visits to the observed cities and corresponding information on these cities from the Wikipedia articles in the language spoken in the country of tourist origin.

The available variables are the number of words, bytes, symbols, paragraphs and pictures contained in the Wikipedia pages of the tourist locations. Symbols (i.e. the letter count) and words directly measure content quantity, pictures inform us about the quality of the article, as illustrations are usually supposed to make the article more appealing. All these are independent variables. The outcome variable is tourist overnight stays in the hotels, where tourists are distinguished by their country of origin and the town of destination. We have selected tourists coming from four countries of origin, namely, France, Germany, Italy and the Netherlands.

Table 1 presents the main variables. The data contain information on tourist visits from Germany, France, Italy and the Netherlands to 158 Spanish cities, which sums up to 632 distinct city-language combinations. Seven years of data are available from 2008 until 2014 for each of these city-language pairs. As can be seen there is a large amount of heterogeneity between the cities, with tourist visits ranging from none to more than 680,000 in a given month. Similarly, the amount of Wikipedia content can vary between very short articles without any pictures and very long articles with 90 paragraphs and 96 pictures.

Table 1: Summary statistics of the data set.

	mean	sd	min	p10	p50	p90	max
Words	1452	1796	2	238	931	3015	19509
Bytes	9614	11900	17	1581	6081	19914	126839
Symbols	10471	12949	17	1764	6640	21931	135970
Pictures	6.8	8.1	0	1	4	15	96
Paragraphs	9.6	11	0	1	7	23	90
Nights	9492	39527	0	0	539	16573	682512
treated	.22	.41	0	0	0	1	1

NOTES: The table shows the distribution of the main variables. The unit of observations is the variable corresponding to visits/the page about city i in langugage j in month t. no. of cities = 158; no. of city-language observations = 632.

[1] These data are publicly available at www.ine.es/inebmenu/mnu_hosteleria.htm.

Table 2: Summary Statistics on Observations by Treatment Status, before Treatment.

	(1)		
	treated vs. control		
	1	2	Total
Nights	13218.5	8624.4	9658.3
	(46353.8)	(38634.8)	(40545.1)
log_nights	6.571	5.986	6.117
	(2.961)	(2.981)	(2.986)
affected	1	0	0.225
	(0)	(0)	(0.418)
Words	1017.6	1680.3	1531.2
	(717.9)	(2013.0)	(1825.6)
Bytes	6697.2	11118.6	10123.6
	(4774.5)	(13331.9)	(12094.5)
Symbols	7356.6	12141.6	11064.7
	(5274.0)	(14565.9)	(13216.2)
Pictures	5.500	7.638	7.157
	(4.532)	(8.811)	(8.098)
Paragraphs	7.288	10.67	9.908
	(7.177)	(11.79)	(11.02)
Observations	29193		

mean coefficients; sd in parentheses

NOTES: The table shows the means of the main variables by treatment status. The unit of observations is (tourists from) language j, at city i in month t. Column 1 shows treated pages and Column 2 the control group. Column 3 show the average for the entire sample. No. of obs. = 49656; no. of destinations = 158; no. of articles = 632.

4.1 Experimental Data

The main idea of our controlled randomized intervention is to selectively translate content about sightseeing and local cuisine that is available in Spanish or English to one of the other language versions of Wikipedia where this information is not yet present. The size of the content added to every treated article is about two or three paragraphs, which is substantial compared to the existing content (we provide a check for that below). For this treatment, we picked 60 Spanish towns out of available 135, for which we observe visitors (and Wikipedia articles) from four countries of origin (France, Italy, Germany, Netherlands). Among these 60 towns randomly selected for the treatment, we assigned to each town a couple of languages out of, overall, six available couples. For the treatment of Wikipedia articles covering selected Spanish towns in assigned languages, in August 2014 we translated and added pieces of text that were already available in English and/or Spanish versions of the article, i.e. before late vacation choices for fall.

The available data on visitors for 135 towns from four countries of origin and Wikipedia articles on four languages for the year 2014 provide us with four observations for each city in each month. This fact enables us with two strategies for obtaining valid control observations for any treated observation. First from tourists to the same city but from a different country, where we did not treat the Wikipedia page or, second, from tourist visits from the same country but to other Spanish cities, whose Wikipedia page we did not modify.

Table 2 contrasts the treated and the control group before the treatment. The first column shows the treated observations and the second column shows the control group. It can be seen that the treated observations on average had a slightly higher number of visitors from the four countries we considered than the control. At the same time their Wikipedia pages were shorter on average. The mean number of monthly visits from country observed is 13,218 for the treated and 8,600 for the untreated observations. At the same time an average treated Wikipedia page had 7.29 paragraphs, while an untreated page would on average have more then 10.7 paragraphs. These differences are not statistically significant, but the point estimates for the mean differ considerably. The high standard deviation (and resulting skewness) in the dependent variable points to the necessity of using logs, which has the additional advantage of allowing us to interpret the resulting coefficients in per cent.

4.2 Analysis of the Cross Sectional Variation

Table 5 shows the results of the cross-sectional OLS regressions between the content measures of a Wikipedia article about a city in that language and the corresponding tourist visits. Column (1) contains the results for the measure of words in the article, (2) for the number of paragraphs and (3) for the number of pictures in the article. This relationship is shown for three years, July 2008 (Cols. 1-3), 2010 (Cols. 4-6) and 2013 (Cols. 7-9). There is no time dimension since we only kept the month of July in the given year, and the reference group are French tourists. Our table shows that most visitors are German, then French and Italian, and the fewest tourists were from the Netherlands. This ordering is generally valid, except in August, when the large majority of French employees go on vacation. Moreover we observe a considerable dynamic in the composition of the tourists over the years.

The correlational regressions in Table 5 show a strong and remarkably robust positive statistical correlation between the content that is available in the Wikipedia of a language and the number of tourists that visit the city. While this correlation does not afford a causal interpretation, it clearly shows that online content is related to tourist visits. According to these raw correlation estimates we see that 1000 additional words coincide with 44 (!) per cent more visitors. Similarly, a city with a paragraph less would have 8 per cent fewer visitors and an article with one additional picture indicates 9 per cent more visitors to that city.

5 Results

From the analysis of cross-sectional variation we notice high correlations between article content and tourist visits of Spanish cities which tend to decrease in more recent years compared to earlier years. Now, we go further and estimate regressions relaxing restrictive assumptions of OLS regressions and making use of panel structure of our data. In this section we first present fixed-effects panel regressions and then move on to the result of our randomized field experiment.

5.1 Correlational Analysis

We firstly analyze the relationship between overnight stays and different indicators of the amount and quality of content on Wikipedia articles for Spanish cities. These indicators are the number of words, text symbols, paragraphs and illustrations. Our first set of OLS regressions with fixed effects is presented in Table 3. The dependent variable in these specifications is the logarithm of total monthly overnight stays in a given city by tourists proceeding from a given country. Dummies for time indicators are included into each regression.

We can observe a strongly significant relationship between the changes in tourist overnight stays and content size measured both in words (column 1) and number of text symbols (column 2) in an article. 1,000 additional words about a given city in a given language are related to an increase of approximately 4 per cent in tourist overnight stays by tourists from the country where this language is predominantly spoken. For the number of symbols in an article, this effect is lower in magnitude, as one would expect, about 0.5 per cent. While the indicators of content quantity show significance, our "quality" measures, the structure of the page (# of paragraphs) and illustrations (# of pictures), do not seem to have any effect on tourist visits.

Using contemporaneous values in the panel data analysis comes with the drawback that, as one could expect, tourists to book holiday trips in advance. Therefore, we also examine whether the lagged values of content indicators have a stronger relationship with

Table 3: Wikipedia content and hotel overnight stays in Spanish cities.

	Log Overnight stays			
	(1)	(2)	(3)	(4)
Words, tsds	0.041**			
	(0.017)			
Symbols, tsds		0.005**		
		(0.002)		
Paragraphs			-0.003	
			(0.005)	
Pictures				-0.003
				(0.004)
Year dummies	Yes	Yes	Yes	Yes
Month dummies	Yes	Yes	Yes	Yes
Mean dep. Variable	6.2	6.2	6.2	6.2
Observations	31394	31394	31394	31394
Number of Pages	531	531	531	531
R^2	0.273	0.273	0.273	0.273

Notes: The table shows the results of the reduced form regressions estimating the relationship between article content on Wikipedia and tourist hotel overnight stays. The four columns show the coefficients for mean content indicators. Fixed effects, year and month dummies are included into all regressions. Column (1) shows the number of words (in thousands), (2) the number of symbols, (3) the number of paragraphs and (4) the number of pictures in each article. Fixed Effects Panel-Regressions with heteroscedasticity robust standard errors. The unit of observations is the number of total monthly hotel overnight stays in the city by tourists proceeding from a country where the main spoken language corresponds to the language of the Wikipedia article. Standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1.

Table 4: The lagged amount of Wikipedia content and hotel overnight stays in Spanish cities.

	Log Overnight stays			
	(1)	(2)	(3)	(4)
Words, tsds	0.041** (0.017)			
Words in t-1, tsds		0.038** (0.017)		
Words in t-2, tsds			0.037** (0.018)	
Words in t-3, tsds				0.038** (0.017)
Year dummies	Yes	Yes	Yes	Yes
Month dummies	Yes	Yes	Yes	Yes
Mean dep. Variable	6.2	6.2	6.2	6.2
Observations	31394	31238	31053	30857
Number of Pages	531	531	531	531
R^2	0.273	0.275	0.277	0.279

NOTES: The table shows the results of the reduced form regressions estimating the relationship between article content on Wikipedia and tourist hotel overnight stays. The four columns show the coefficients for mean content indicators. Fixed effects, year and month dummies are included into all regressions. Column (1) shows the number of words (in thousands) in each Wikipedia article, columns (2)-(4) the corresponding lags of the number of words for periods t-1, t-2, t-3. Fixed Effects Panel-Regressions with heteroscedasticity robust standard errors. The unit of observations is the number of total monthly hotel overnight stays in the city by tourists proceeding from a country where the main spoken language corresponds to the language of the Wikipedia article. Standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1.

Table 5: Wikipedia content and touristic visits - Crossectional OLS-Regressions.

	2008			2010			2013		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
German visitors	0.357	0.315	0.258	0.564***	0.467***	0.451**	0.507*	0.421	0.432
	(0.318)	(0.313)	(0.323)	(0.168)	(0.172)	(0.176)	(0.271)	(0.262)	(0.274)
Italian visitors	-0.180	0.015	-0.058	-0.059	-0.012	-0.060	-0.424*	-0.271	-0.427*
	(0.315)	(0.310)	(0.301)	(0.153)	(0.166)	(0.141)	(0.235)	(0.243)	(0.230)
Dutch visitors	-0.657*	-0.479	-0.831**	-0.333**	-0.242	-0.488***	-0.343	-0.108	-0.277
	(0.339)	(0.342)	(0.336)	(0.153)	(0.176)	(0.152)	(0.247)	(0.253)	(0.249)
Words, tsds	0.443***			0.325***			0.297***		
	(0.105)			(0.069)			(0.045)		
Paragraphs		0.081***			0.058***			0.063***	
		(0.011)			(0.013)			(0.009)	
Pictures			0.095***			0.070***			0.068***
			(0.013)			(0.018)			(0.009)
Constant	6.578***	6.402***	6.573***	6.681***	6.587***	6.729***	6.883***	6.648***	6.895***
	(0.258)	(0.251)	(0.234)	(0.228)	(0.259)	(0.255)	(0.182)	(0.203)	(0.177)
Mean dep. Variable	7.03	7.03	7.03	7.22	7.22	7.22	7.30	7.30	7.30
Observations	312	312	312	341	341	341	428	428	428
Adj. R^2	0.111	0.142	0.121	0.111	0.129	0.091	0.104	0.140	0.096

Notes: The table shows the results of the crosssectional relationship between Wikipedia Content and Tourist Visits. The first three columns show the relationship for July 2008, the second three Columns (4)-(6) for July 2010 and the last 3 columns (7)-(9) for July 2013. Column (1) shows words, (2) shows the the number of paragraphs and (3) the number of pictures in the articles. OLS-Regressions with heteroscedasticity robust standard errors. The unit of observations is the outcome for tourists from country j in city i (no time dimension). Standard errors in parentheses: *** $p<0.01$, ** $p<0.05$, * $p<0.1$;

tourist choices. The results are presented in Table 4. They show that the contemporaneous correlations are the strongest, but those with the lagged amount of words are similarly strong (about 3.8 per cent). This suggests that there could be some potential for a causal impact of the tourist-relevant information in Wikipedia articles on tourist decisions.

5.2 Controlled Randomization

We now turn to the results from the difference-in-differences based on our controlled randomized field experiment. Table 6 shows that our treatment affected the content on Wikipedia. We verify the effect of our treatment on a monthly basis. If there were multiple user edits of the same article in that month, we have multiple values for the outcome of interest (e.g. number of words of an article). To make sure, that any specific user edit did not have extraordinary impact, we evaluate the mean and the median of the article characteristics in any given month. If the article did not change during any given months, these values are the same. Group 1 (Columns 1-3) show mean outcomes and Columns 4-6 the median. The first two columns of each group presents the length of an article (in bytes (1), and number of paragraphs (2)). The third column shows the number of pictures. The coefficient "affected_after" quantifies how we affected the content in the treated articles. Treatment added 1200 bytes and 0.83 additional paragraphs on average. Sometimes more content was added and sometimes our content was rejected. Our treatment had no significant effect on pictures.

Table 7 shows the effect of our treatment on Wikipedia on overnight stays. The first two columns show overnight stays as dependent variable and the second two columns (3)-(4) show the log of overnight stays. Column (5) shows a dummy that takes the value of 1 if there were any overnight stays and 0 otherwise.

The effect of our treatment is measured by the coefficient $After_t \times Affected_{ij}$. While the coefficient has a positive tendency, it is never statistically significant. If anything we can say that there seems to be a weak positive relationship, but, based on our design, we cannot reject the hypothesis that Wikipedia content has no causal effect on tourist visits.

Robustness:

Across all specifications (also the ones not reported), the overall positive tendency for the autumn months after treatment prevails, but effects are negative at times. The results also remain essentially the same when expanding the comparison group and using

Table 6: Wikipedia Content and our Treatment.

	Mean Content			Median Content		
	(1)	(2)	(3)	(4)	(5)	(6)
	Symbols (bytes)	Paragraphs	Pictures	Symbols (bytes)	Paragraphs	Pictures
After	1232.565***	0.963***	1.106***	1196.263***	0.939***	1.085***
	(247.114)	(0.173)	(0.215)	(247.499)	(0.173)	(0.216)
After x Affected	1237.261***	0.832***	-0.302	1234.272***	0.835***	-0.299
	(407.736)	(0.285)	(0.354)	(408.370)	(0.285)	(0.357)
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
Month dummies	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. Variable	12145.3	11.1	7.7	12158.1	11.1	7.7
Observations	3528	3528	3528	3528	3528	3528
Number of Pages	527	527	527	527	527	527
R^2	0.029	0.080	0.037	0.028	0.079	0.037

NOTES: The table shows the results of the reduced form regressions estimating the effect of treatment on Wikipedia outcomes. The first three columns show mean outcomes and the second three Columns (4)-(6) the median. Column (1) shows bytes, (2) shows the the number of paragraphs and (3) the number of pictures in the articles. Fixed effects panel-regressions with heteroscedasticity robust standard errors. The unit of observations is the outcome of a page i on Wikipedia of language j in month t. Standard errors in parentheses: *** $p<0.01$, ** $p<0.05$, * $p<0.1$;

Table 7: Wikipedia Treatment and Overnight Stays.

	Overnight stays		Log(overnight stays)		Dummy
	(1)	(2)	(3)	(4)	(5)
	Raw	Controls	AllCities	VisitedCities	AnyNights
After	621.663*	534.944	0.140**	0.083***	0.005
	(375.047)	(427.146)	(0.058)	(0.026)	(0.010)
After x Affected	998.709	1000.051	0.030	0.038	0.006
	(793.882)	(783.687)	(0.107)	(0.046)	(0.018)
Year dummies	No	Yes	Yes	Yes	Yes
Month dummies	No	Yes	Yes	Yes	Yes
Mean dep. Variable	13104.34	13104.34	6.78	7.34	0.92
Observations	3896	3896	3896	3599	3896
Number of Cities	540	540	540	533	540
R^2	0.002	0.029	0.086	0.129	0.030

Notes: The table shows the results of the reduced form regressions estimating the effect of treatment on tourist outcomes. The first two columns show overnight stays and the second two Columns (3)-(4) show the log of overnight stays. Column (5) shows a dummy that is 1 if overnight stays were larger than 0. Fixed effects panel-regressions with heteroscedasticity robust standard errors. The unit of observations are the visitors from country j of a Spanish city i in month t. Standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1;

the years 2012-2015 in the analysis. However, some coefficient estimates become very small negative and insignificant.

The estimations continue to be very similar if November is included, but they are different with December. The reason for that may be that December is a special and very different month for tourism.

6 Conclusions

In this paper, we analyze the relationship of content availability on Wikipedia and choices of tourism destinations. Specifically, we ask whether tourists from four European countries are more likely to visit a (Spanish) tourist destination, if they can find more information about the place in their native language on Wikipedia. To study this question we combine the data on monthly tourist visits to Spanish cities by country of origin and the content that was available on the corresponding language version of Wikipedia.

We analyze the data in descriptive cross-sections, as a panel and in a randomized controlled field experiment. We document a strong correlation between available content and tourist visits in a cross-sectional analysis. We then show that there is a strong correlation between the length of the article and the number of visitors, when accounting for unobserved heterogeneity in a fixed effects panel estimation. This analysis reveals however, that the effect is much smaller than in the cross section and even insignificant for the number of paragraphs and pictures.

To address the concern of reverse causality we run a randomized controlled field experiment where we randomly choose 120 from 540 potential treatments of content about a city in one of four European languages and estimate a difference-in-differences. The resulting estimates are positive but insignificant, so that we can not reject the null hypothesis that the content on Wikipedia does not affect choices of tourist destinations.

Very broadly speaking, there is wide anecdotal evidence that people consult Wikipedia to find information. It is highly plausible that the availability or absence of information influences what we know and which choices we make. This is clearly important question, because providing information is not very costly, but might provide great value if it could be shown that many users base their choices on online information that is provided by volunteers. Quantifying the intensive and extensive margin of users who change their choices or decide where to visit is, hence, of high importance. We provide evidence that this correlation is very strong and also persists when controlling for unobserved heterogeneity. However, as our research shows, the precise impact is hard to measure in a strictly causal design, because of the endogeneity of content generation itself.

While the correlations in the cross-section and panel analysis are strong and robust, we could not show a significant effect when using a strategy based on randomized treatments. At the same time the estimated causal coefficients have a consistent tendency of being positive. The question that we cannot answer at this point is whether the lack of statistical significance is simply due to low statistical power of our experiment or whether it means that Wikipedia's impact on tourist choices is only limited. Further experimental research that expands the scope of our analysis would be desirable. While the cost of such treatments can be elevated, using the suggested research design to study other areas of information acquisition, such as medicine or school choices could be fruitful directions.

References

Acquisti, Alessandro and Christina M Fong, "An Experiment in Hiring Discrimination via Online Social Networks," *Available at SSRN 2031979*, 2013.

Adler, B Thomas, Krishnendu Chatterjee, Luca De Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman, "Assigning Trust to Wikipedia Content," in "Proceedings of the 4th International Symposium on Wikis" ACM 2008, p. 26.

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

Laurent, Michaël R and Tim J Vickers, "Seeking Health Information Online: Does Wikipedia Matter?," *Journal of the American Medical Informatics Association*, 2009, *16* (4), 471–479.

Luca, Michael, "Reviews, Reputation, and Revenue: The Case of yelp.com," *Com (September 16, 2011). Harvard Business School NOM Unit Working Paper*, 2011, (12-016).

Xu, Sean Xin and Xiaoquan Michael Zhang, "Impact of Wikipedia on Market Information Environment: Evidence on Management Disclosure and Investor Reaction," *MIS Quarterly*, 2013, *37* (4), 1043–1068.