Externalities in Knowledge Production: Evidence from a Randomized Field Experiment[*]

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

February 5, 2020

Abstract

Are there positive or negative externalities in knowledge production? We analyze whether current contributions to knowledge production increase or decrease the future growth of knowledge. To assess this, we use a randomized field experiment that added content to some pages in Wikipedia while leaving similar pages unchanged. We find that adding content did not have a sizable impact on long-term content growth, neither in terms of quantity or quality. However, it increased editing activity in the first two years. Our results have implications for information seeding and incentivizing contributions. They imply that additional content may inspire future contributions in the short- and medium-term, but do not generate sizable externalities in the long-term.

JEL: L17, L86, C93

Keywords: knowledge accumulation, user-generated content, Wikipedia, public goods provision, field experiment

[*] We are grateful to Yan Chen, Chris Forman, Willa Friedman, Shane Greenstein, David Hugh-Jones, Tobias Kretschmer, Giovanni Mastrobuoni, Ignacio Monzón, Juan Morales, Abhishek Nagaraj, Stefan Penczynski, Imke Reimers, Stephan Seiler, Ananya Sen, Matthias Sutter, and Michael Zhang for valuable comments. We would also like to thank the seminar participants at the University of Strasbourg, the Collegio Carlo Alberto, George Mason University, ParisTech 2019, Digital Economy Workshop 2019 (Católica Lisbon), ZEW Conference on the Economics of ICT, and the Advances with Field Experiments 2019 Conference (University of Chicago) for valuable input.

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

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

[§] Georgia Institute of Technology & University of East Anglia, michael.kummer@econ.gatech.edu.

[¶] Centre for European Economic Research (ZEW), slivko@zew.de.

Electronic copy available at: https://ssrn.com/abstract=3341630

1 Introduction

Knowledge is a key input to many economic activities and a driver of economic growth (Romer, 1990; Grossman and Helpman, 1993; Jones, 1995). An increasing share of knowledge is created in the form of user-generated content: consumer feedback systems, discussion boards, Q&A sites, open-source software, social networks, and online information repositories, such as Wikipedia. Online knowledge repositories have the potential to revolutionize how society aggregates and transmits knowledge. This potential stems from their ability to combine and aggregate the input of many individuals independent of time and location, and the generated content can be retrieved at a low cost.[1]

A key component to the success of user-generated content repositories is a sufficient flow of content contributions by users. Hence, understanding the drivers of contributions to user-generated content has been an important question in economics and management for the past two decades (Lerner and Tirole, 2003). While traditional motives of public goods provision play a vital role in user-generated content, recent literature has identified a novel driver in the form of a feedback loop—a dynamic by which small initial contributions of content inspire ever more follow-on content contributions by other users (Aaltonen and Seiler, 2016; Kane and Ransbotham, 2016; Zhu et al., 2020).

In this paper, we investigate whether there are positive or negative externalities in user-generated content production. Understanding and quantifying such externalities has important policy implications. If content generation has positive externalities on future content generation, then information seeding and paid contributions may have a high return on investment in terms of added stimulated growth (Aaltonen and Seiler, 2016).[2] On the other hand, if content generation has negative externalities, then such policies may backfire and be ineffective or even lead to worse eventual outcomes (Nagaraj, 2019).

Due to the reflection problem (Manski, 1993), externalities in content generation are difficult to identify. An externality occurs when a contribution by a user motivates other users to contribute (positive externality) or prevents further contributions (negative externality). Yet, correlation in contributions does not necessarily attest to an externality. A positive correlation may arise when two users are exposed to the same external shock,

[1] Traditional channels of personal knowledge transmission require a double-coincidence demand and supply of knowledge. The "knowledge-seeker" and the "knowledge-holder" have to meet in person or at least at the same time. The elimination of such double-coincidences has been modeled to understand the advantage of monetary over barter-economies. Kiyotaki and Wright (1989). These features give such systems a drastic competitive advantage that may affect the education sector and other traditional channels of knowledge transmission. The sector of encyclopedic knowledge is one of the most salient examples of the new technology's potential.

[2] Nagaraj (2019) describes how such policies have been used by Wikipedia (seeding articles on more than 30,000 US cities from US Census Bureau data), OpenStreetMap (US Census maps), and Reddit (fake user accounts).

Electronic copy available at: https://ssrn.com/abstract=3341630

such as a news article or a research finding. Similarly, a non-causal negative correlation may be caused by processes with periodic updates, such as elections or periodically updated statistics. To identify the causal effect, shocks to content growth and contributions must be independent over time. We estimate the causal impact of additional content on subsequent contributions using a randomized field experiment in Wikipedia. Randomization ensures that the addition of content is exogenous in terms of future content generation. As argued by the literature analyzing social interactions (including Manski (1993) in general and more specifically Aaltonen and Seiler (2016) and Zhu et al. (2020)), randomized experiments are the best way to cleanly identify causal relationships in such interactions.

The exogenous variation in our data comes from a randomized field experiment, which was conducted in 2014. The experiment added relevant content to randomly chosen Wikipedia pages while leaving similar pages unchanged. The treatment added about two paragraphs (approximately 2,000 characters) and one picture to each page in the treatment group. The pages were about mid-sized Spanish cities in different language editions of Wikipedia.[3]

For our analysis, we collect data from multiple sources. To measure the impact on the quantity of content and editing activity, we use a dataset of Wikipedia editing histories, which includes all versions of Wikipedia pages in the treatment and the control groups. To measure the effect on the quality of content, we use two approaches. First, we developed a quality rating scheme, and for each page, we obtained quality ratings by two independent raters who are fluent in the corresponding language. Second, we use text analysis to compare the content across different language editions of Wikipedia and measure similarity to the corresponding pages in the Spanish Wikipedia. Our dataset and the experimental setting allow us to analyze both short-term and long-term effects, up to four years after the experiment. Our main outcomes of interest are the quantity and quality of content. To study editing activity, we also analyze the number of unique editors, the number of edits, and the amount of content added and deleted.

We find that the additional content did not have a sizable impact on the long-term growth of the quantity and quality of content. Consequently, the pages which were improved by adding about 2,000 characters of content were still only longer by about the same amount, even four years after treatment. Similarly, while the added content increased the quality, the quality difference four years later was about the same as immediately after the treatment. We do find some short-term impact. In particular, we find that the editing activity increased in the first two years after the treatment. In this initial

[3] A comprehensive description of the experiment is provided in Hinnosaar et al. (2019), who studied the impact of this treatment on real-world outcomes.

Electronic copy available at: https://ssrn.com/abstract=3341630

stage, the treatment increased the number of Wikipedia users editing the treated pages and increased the number of edits. However, there was no sizable impact in the third and fourth year post-experiment. Moreover, even in the first years, the amount of content these users added was small and their edits were mostly limited to directly modifying the text added by the treatment.

While our study benefits from clean identification, it faces two important limitations. First, our test has low power to measure small effects on content growth. Our minimum detectable effect size is 13% for four-year growth in page length. Hence, we can reject sizable effects, but cannot exclude the possibility of small, but economically meaningful positive or even negative effects. However, our test has relatively more power to detect meaningful changes in editing activity. Second, our results might not generalize to other platforms or content in other stages of development. In our setting, the content is relatively mature but still incomplete. It is plausible that the results would be different if we had studied either new pages or almost complete pages.

Our main finding has a clear policy implication—at least in settings similar to the one studied here, investments in information seeding and incentivizing additional content contributions may temporarily increase user participation, but do not have a sizable cumulative effect on content growth. Therefore information seeding and incentivizing contributions are mainly a matter of direct cost-benefit analysis: they pay off if and only if the costs of creating the content are lower than the value of the new content. The additional costs or benefits via externalities that discourage or inspire future contributions are small.

Our paper contributes to the literature that studies externalities in user-generated content production. The closest to our work are Aaltonen and Seiler (2016), Kane and Ransbotham (2016), Nagaraj (2019), and Zhu et al. (2020). Aaltonen and Seiler (2016) and Kane and Ransbotham (2016) used detailed observational data from Wikipedia. Nagaraj (2019) used a natural experiment on OpenStreetMap, a Wikipedia-style digital map-making community, which started with better seeding information in some regions compared to others for quasi-random reasons. Zhu et al. (2020) studied a natural experiment that motivated students to contribute to Wikipedia. The papers arrived at contradicting conclusions, which warrant further investigation regarding this issue. Our paper is the first to study the question using a randomized field experiment, which allows a clean identification of the underlying externalities, especially when analyzing the long-term impact.

More generally, the paper belongs to the literature that analyzes what drives contributions of user-generated content, which represents a salient and highly relevant digital

Electronic copy available at: https://ssrn.com/abstract=3341630

public good.[4] Among the main drivers of content contributions, the studies addressed the role of personal gain (Shah, 2006), social comparison (Chen et al., 2010), group size (Zhang and Zhu, 2011), networks (Fershtman and Gandal, 2011; Ransbotham et al., 2012), spillovers (Kummer, 2014), symbolic awards (Gallus, 2017), performance feedback (Huang et al., 2018), monetary rewards vs social motives (Sun et al., 2017), contributor diversity (Ren et al., 2015), and economic conditions, such as unemployment (Kummer et al., 2019) and migration (Slivko, 2018). Social motives have been shown to affect public good provision (Goldstein et al., 2008; Lacetera and Macis, 2010; Ayres et al., 2013). Specifically, in the case of digital public goods, examples include ranking movies on MovieLens (Chen et al., 2010), editing articles on Wikipedia (Zhang and Zhu, 2011; Algan et al., 2013), and endorsing messages of Facebook users (Egebark and Ekström, 2017).[5] In our setting, social externalities are rather implicit, as individuals contributing content are not connected by any direct social ties, and the interactions with the other members of the community can occur only in the process of contributing knowledge to the Wikipedia articles. Our paper extends the literature by using variation from a randomized field experiment to measure the impact of additional content on future content generation.

The structure of the paper is as follows. In the next section, we describe the experiment and provide some background on Wikipedia editing. Section 3 describes the data. Section 4 presents the results of the impact of the treatment on the subsequent growth in content quantity and quality and on editing activity. Section 5 introduces a simple theoretical model and interprets our results in this framework. Section 6 discusses the connection between our findings and related literature. It also discusses the implications and generalizability of our findings. Section 7 concludes.

2 Experiment and background

2.1 Experiment

The field experiment added content (text and photos) to randomly chosen Wikipedia pages. The sample consisted of 240 Wikipedia pages. Specifically, it consisted of the pages of 60 Spanish cities in the French, German, Italian, and Dutch editions of Wikipedia. The cities were all medium-sized, excluding the largest like Madrid and Barcelona, and

[4] Other studies on Wikipedia have analyzed biases in Wikipedia's content (Greenstein and Zhu, 2012, 2018; Hinnosaar, 2019) and the impact of Wikipedia on market outcomes (Xu and Zhang, 2013; Hinnosaar et al., 2019) and science (Thompson and Hanley, 2018).

[5] More generally, the literature suggests strong effects of social influence on individual choices related to savings (Duflo and Saez, 2002, 2003), education (Hanushek et al., 2003; De Giorgi et al., 2010), entertainment (Salganik et al., 2006), etc.

Electronic copy available at: https://ssrn.com/abstract=3341630

also excluding the smaller cities. The Wikipedia pages in these languages were relatively short—up to 24,000 characters in each of these four languages.

Each city and each language edition of Wikipedia was treated equally. For each city, its page was assigned to the treatment group in two randomly chosen languages. In each language edition of Wikipedia, 30 randomly chosen city pages were assigned to the treatment group. Specifically, to obtain balance in the treatment and control groups, the randomization was stratified.[6] The 60 cities were divided into ten equal-sized groups. Within each group, each city was randomly assigned to one of six treatment arms. The six treatment arms were as follows: treat the city page in one of the six possible language pairs (French & German; French & Italian; French & Dutch; German & Italian; German & Dutch; Italian & Dutch). This resulted in a design where the number of pages which were treated equaled the number of those that remained in the control group.

The Wikipedia pages were treated mid-August, 2014. The treatment added about 2,000 characters of text and photos to each page in the treatment group. The added text and photos were mostly obtained from the corresponding Spanish and English language Wikipedia pages. Because all the pages were about Spanish cities, the Spanish Wikipedia typically contained more information than the other language versions. The English language version of the page, typically, was also more detailed than in the languages in the experiment. Hence, there was information available in Spanish and English pages that was missing from the other language editions of Wikipedia. The treatment translated that text and added it to the corresponding pages in the treatment group.

The treatment of the pages in Dutch Wikipedia was not successful. While in French, German, and Italian Wikipedia, the added text and photos survived well over time, all the additions to Dutch Wikipedia were deleted within 24 hours (by a single editor). Wikipedia allows anyone to edit. It also means that anyone can delete or undo the latest changes by reverting to a previous version of the page. This happened in the Dutch version of Wikipedia, where 24 hours after the treatment, all the pages looked as if they had never been treated. Therefore, we exclude Dutch pages from our main analysis and restrict attention to the 180 pages in French, German, and Italian. Robustness analysis shows that our results do not change if Dutch pages are included in the analysis.

2.2 Power analysis

We acknowledge that the sample size and power of our tests are rather small. We analyze this in appendix A, which presents power analysis for one of our main outcome variables (page length) and the main editing activity variable (number of users). As estimation

[6] For further details of the randomization, see Hinnosaar et al. (2019).

Electronic copy available at: https://ssrn.com/abstract=3341630

results in section 4 show, we would expect the power for other measures to be similar.

To provide some context, let us discuss the expected effect sizes, given the results from previous literature. The main measure studied in most previous works is the impact of added content on the number of future contributors. Aaltonen and Seiler (2016) estimate that adding 10,000 characters of content leads to 0.204 additional users per week, which corresponds to about 0.18 additional users per month for 2,000 characters as added by the treatment in our case. Zhu et al. (2020) estimates are similar: their median treatment size was 3,180 characters and estimated increase of 2 unique users per six-month period, which implies 0.21 new users per month per 2,000 added characters.[7]

The literature provides less guidance for the long-term effect size for the quantity and quality of the content. Aaltonen and Seiler (2016) use their estimates for the impact on the number of users to simulate a possible effect on the content quantity and find 45% growth in content. The only other paper to estimate this effect is Nagaraj (2019), who finds a long-term effect of negative 10%.

Figures A.1a and A.1b in appendix A describe the power analysis for page length. Figure A.1a shows that when the Dutch pages are included in the sample, as originally intended, then if the true treatment effect is 10% increase in page length over 4 years, we would reject the null hypothesis of no effect at 10%-significance level with 76% probability and at 5%-significance level with 65% probability. The minimum detectable effect size is about 12%.[8] If we exclude the Dutch pages (figure A.1b), we lose some power, but the minimum detectable effect is still around 13%. Indeed, our study is underpowered to detect small long-term effects, but there should be no difficulties detecting even half of the effect-size suggested by Aaltonen and Seiler (2016). Figures A.1c and A.1d show that our experiment has relatively more power to detect meaningful effects on editing activity. Even if we exclude the Dutch pages (figure A.1d), the minimum detectable effect size is 0.11 users, and we should certainly be able to detect the effect sizes suggested in the literature. Section 4 describes the ex-post minimum detectable effect sizes (with our realized data), which turn out to be similar to those described here (calculated based on the pre-experiment data).

[7] Kane and Ransbotham (2016) provide some evidence that the effect could be larger for less developed content. They find that in the case of less developed articles, 1%-increase in length implies 0.03–0.04 more monthly contributors. In our case, the treatment was on average 23% of the page length, which would imply 0.7–0.9 more users.

[8] The minimum detectable effect size is calculated at 5%-significance level and 80% power.

Electronic copy available at: https://ssrn.com/abstract=3341630

2.3 Background on Wikipedia and its editing

Wikipedia exists in 309 languages, and the different language editions are not identical. The differences across languages made the experimental design possible. As we show in section 3.5, Spanish Wikipedia contained much more information about Spanish cities than the pages in our sample. This imbalance allowed the treatment to translate information from Spanish Wikipedia to the target languages.

Why don't Wikipedia editors translate the content between languages themselves? First, it requires language skills. Many people are monolingual (Eurobarometer, 2012). English language skills would not be enough because the pages in English Wikipedia are also rather incomplete (as we show in section 3.5). The language skills required to translate between Spanish, French, German, and Italian are not common. Table B.1 shows that in France, Germany, and Italy, less than 10% of the population can read Spanish, and in Spain, less than 10% of the population can read the languages in our sample. Second, perhaps equally importantly, Wikipedia is written by volunteers whose motivation depends largely on how fun the editing process is (Nov, 2007). Presumably, the task of translating information is somewhat mundane compared to other possible uses of time. Third, note that for the majority of the 309 language editions of Wikipedia, automatic translation is still not good enough. Furthermore, even when automatic translation is technically possible, to use it, it would require that Wikipedians agree on which language version is the superior one.

Let us briefly describe how Wikipedia editing works on these pages. While large edits that add an entire section or paragraph definitely exist, they do not make up the majority of Wikipedia edits. Most Wikipedia edits are small, fixing typos, grammar, and formatting, rearranging text without modifying content. Figure B.1a shows that 74% of the edits to pages in our sample (before treatment) add less than 100 characters (about one sentence). 41% of edits are marked by the editor as minor edits, which means not modifying content. Only about 4% of edits add more than 1000 characters (about one paragraph).

Where do the small edits come from? We hypothesize that some of these small edits occur when editors keep track of pages that they are interested in. On Wikipedia, any registered user can sign up to be notified when a certain page has been modified. We find some evidence supporting this on the corresponding city pages in English and Spanish Wikipedia. Figure B.1b shows that almost half of these pages have more than 30 editors that get notified of all changes. Unfortunately, if there are less than 30 watchers for a page, Wikipedia does not report the exact number. Hence, we just know that only 8 pages in our sample have at least 30 persons signed up for notifications. Nevertheless,

Electronic copy available at: https://ssrn.com/abstract=3341630

the comparison with the pages in English and Spanish Wikipedia suggests that for the remaining pages, there should be some watchers as well.

3 Data

We combine multiple sources of data. To measure the impact on the quantity of content and editing activity, we use a dataset of Wikipedia editing histories. An editing history contains the full text of each revision[9] of each page starting from the creation of the page until the beginning of September 2018.

To measure the effect on the quality of content, we use two approaches. First, we developed a quality rating scheme, and for each page, we obtained quality ratings by two independent raters who are fluent in the corresponding language. They rated three revisions of each page in our sample (before treatment, after treatment, and four years after treatment) comparing it to the English Wikipedia as a benchmark. Second, we use text analysis to compare the content across different language versions of Wikipedia and measure similarity to the corresponding pages in the Spanish Wikipedia.

Our sample consists of the 180 pages in the experiment, which are the pages of 60 cities in French, German, and Italian Wikipedia. In the following subsections, we describe the construction of the dataset and variables used in the analysis.

3.1 Page length

One of our main outcome variables is the page length after the experiment. We measure page length in characters, including spaces and wiki markup commands.

Figure 1 presents average page length in the treatment and control groups. Until the experiment in August 2014, the average page length in the control and treatment groups was rather similar.[10] The experiment added significant length to the pages in the treatment group. After the experiment, the difference has been relatively stable. An exception is a sharp increase in the mean of the treatment group in August 2016. This jump comes from the efforts of a single editor who worked hard to improve one page in French Wikipedia—the page of the city of Cordoba.[11] Appendix B presents the same

[9] A revision (or an edit) is a version of a Wikipedia article saved at a specific moment by a particular user. All revisions with the corresponding metadata, including full text, user, and timestamp, are preserved by Wikipedia and publicly available.

[10] The drop in both the treatment and control groups in early 2013 comes from technical changes in Wikipedia: Addbot removed about 2,000 characters from each page with an explanation similar to "Migrating 77 interwiki links, now provided by Wikidata".

[11] By August 2016, the page of Cordoba in French Wikipedia was relatively long, with 19,426 characters (at the time 93% of the pages in our sample were shorter than that). During August 2016, this user

Electronic copy available at: https://ssrn.com/abstract=3341630

figure, first without French Cordoba (figure B.2a) and second, with the logarithm of page length (figure B.2b), both of which show no evidence of an increase in the treatment group average in 2016.

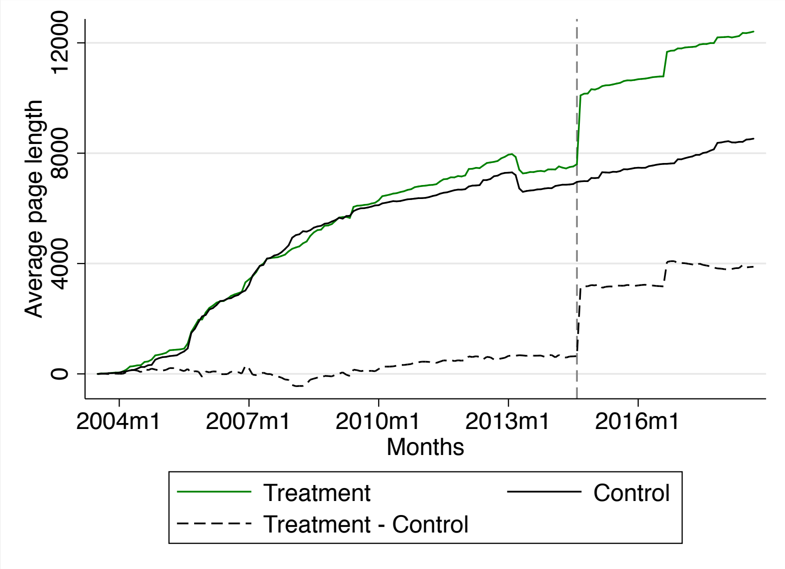


Figure 1: Average page length in the treatment and control groups

Notes: The number of observations is 90 in the control and 90 in the treatment groups. The experiment month (August 2014) is marked by the dashed vertical line.

Similar dynamics can be seen when looking at the changes separately by language (figures 2a to 2c). As expected, the placebo test with the Dutch pages (figure 2d) shows that the assignment to the treatment group had no impact. Page length is one possible output measure of knowledge production in Wikipedia. Similar dynamics as in figure 1 can also be seen in figure B.3 in appendix B, which presents alternative measures of content: images and plain text (that is, html elements removed from the parsed text).

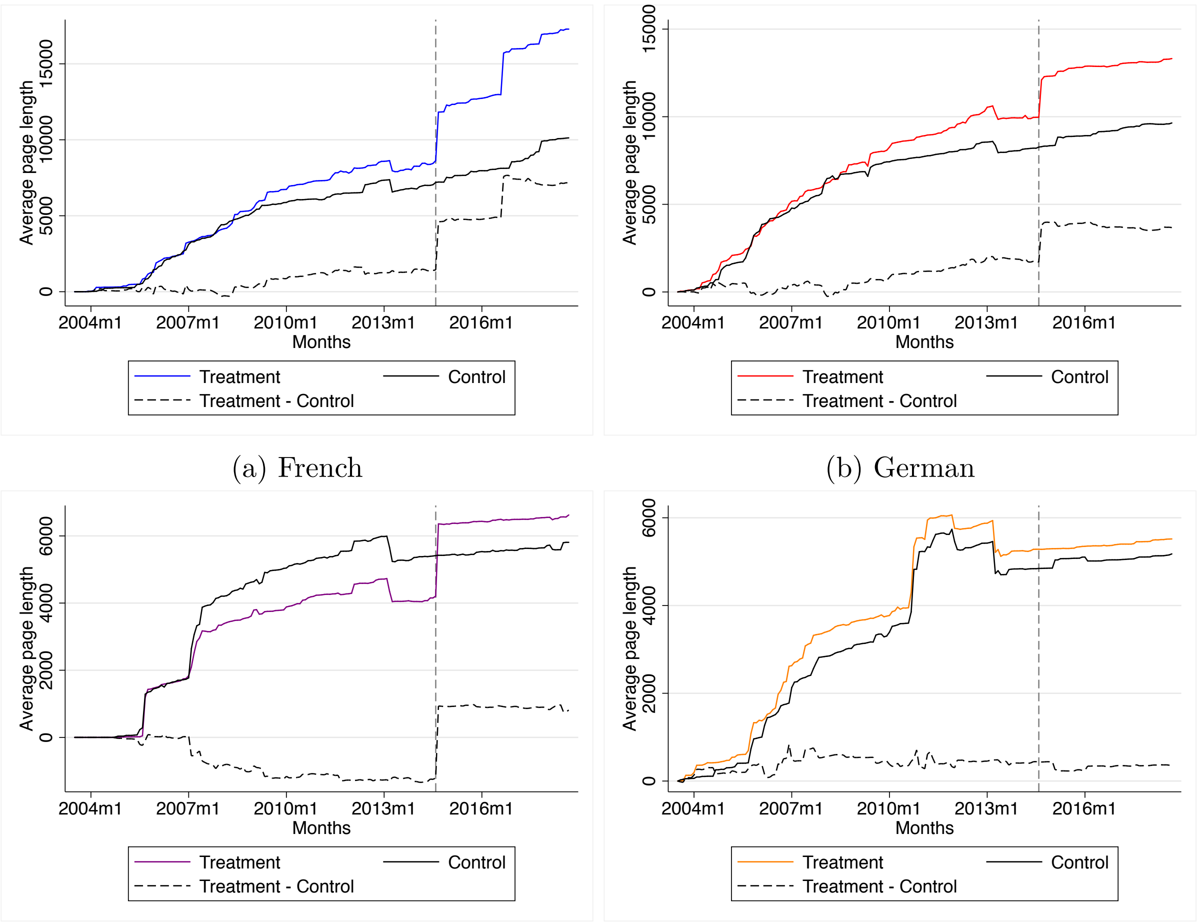
To make the treatment and the control groups comparable, we subtracted the length of text added by the treatment from the length of pages in the treatment group. Moreover, as the distribution of page lengths is relatively skewed, we use the logarithm of page length in our estimations.

3.2 Quality ratings

To assess the changes in the quality of Wikipedia articles, we hired six research assistants to rate the quality of articles. Each article in French, German, and Italian Wikipedia in

increased the page length to 100,702 characters, which is almost twice the length of the longest page at the time (57,076 characters). Our conclusions do not change if we exclude this page.

Electronic copy available at: https://ssrn.com/abstract=3341630



(a) French

(b) German

(c) Italian

(d) Placebo: Dutch

Figure 2: Average page length in the treatment and control groups, by language

Notes: On each figure, the number of observations is 30 in the control and 30 in the treatment groups. The experiment month (August 2014) is marked by the dashed vertical line.

our sample was evaluated by two raters, who were fluent in the respective language as well in English.

We asked the raters to evaluate the quality of three versions of 60 Wikipedia articles in our sample. Version A was the latest version before August 1st, 2014 (i.e., pre-treatment), version B the latest version before September 1st, 2014 (i.e., post-treatment), and version C the latest version before September 1st, 2018 (i.e., four years after the treatment). As a benchmark, we used articles in English-language Wikipedia (as of September 1st, 2018).[12]

Each version of each article was rated in five dimensions on a scale where 0 is the lowest possible rating and 100 means equivalence to the benchmark page from English Wikipedia (for detailed instructions, see figure B.4 in the Appendix). The five dimensions were the

[12] As we show below, the articles about Spanish cities in English-language Wikipedia are sometimes quite incomplete, so ideally we would have preferred to use Spanish Wikipedia as a comparison. Because the combination of necessary language skills is not common, it would have been prohibitively costly.

Electronic copy available at: https://ssrn.com/abstract=3341630

following. (1) *Completeness*: the article comprehensively covers all relevant aspects of the city (compared to the article in English). (2) *Well-written*: the prose is clear, concise, and spelling and grammar are correct. (3) *Illustrated*: the article includes photos that are relevant to the topic and have suitable captions (compared to the article in English). (4) *Interesting*: the article makes the city seem like an exciting place to visit (compared to the article in English). (5) *Overall*: Overall, the article is a high-quality reference source (compared to the article in English).

For example, a score of 100 in the completeness dimension means that the article covers the relevant aspects of the city as comprehensively as the corresponding page in English. It may occur if the pages cover the same topics in the same level of detail, or if they cover different topics, but the missing parts "balance out" in the eyes of the raters. A score of 50 would mean that the page is half as good and 200 that the page is twice as good as the benchmark page from English Wikipedia.

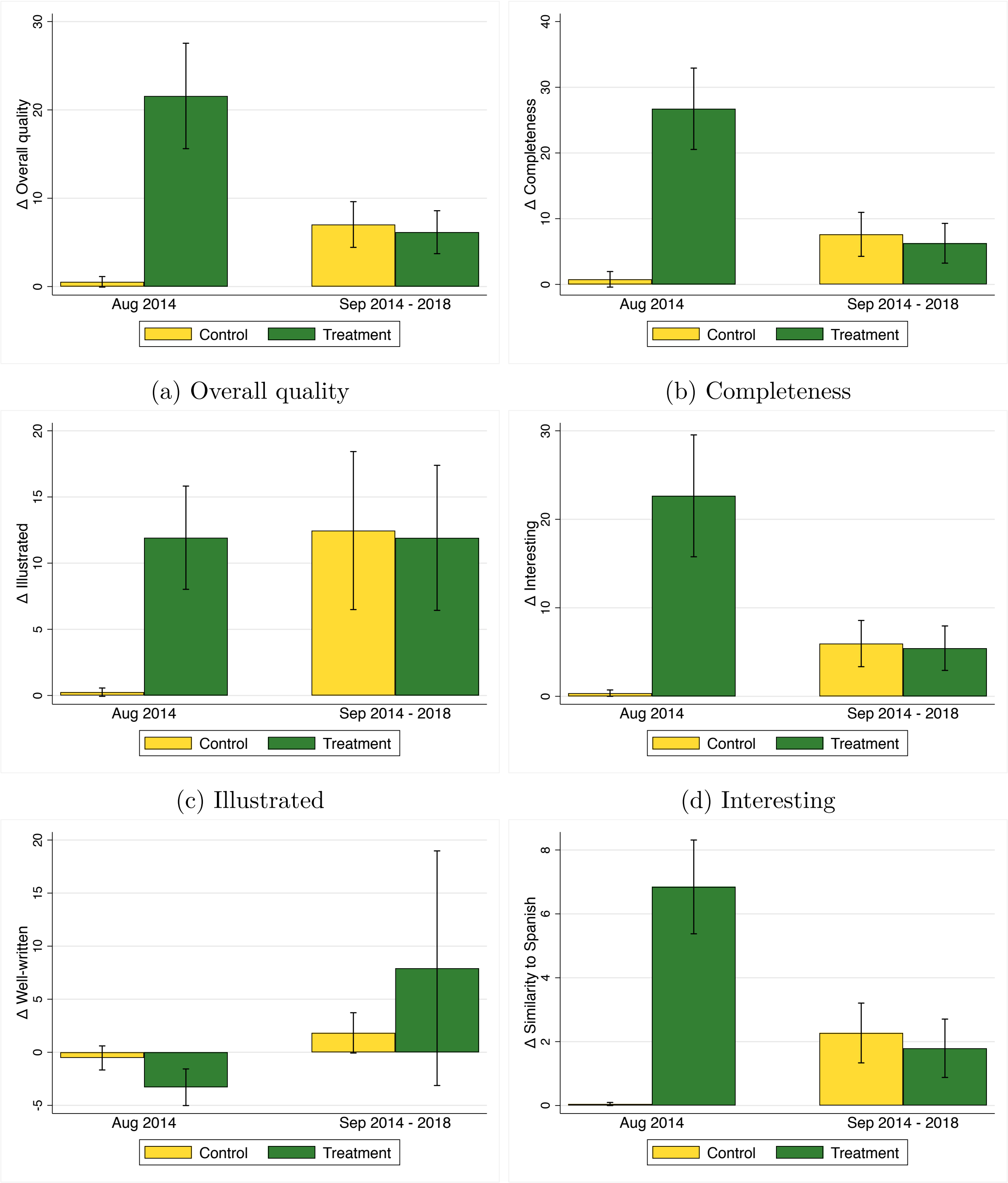
Figure 3 presents the change in quality during the treatment (August 2014) and within four years after treatment (September 2014–September 2018). It shows that in the treatment group, as expected, there was a large increase in quality during the treatment. The increase takes place in overall quality and in all measured dimensions except well-written. Indeed, during the treatment month, pages in the treatment group become slightly less well-written. This is expected because, in the experiment, the treatment text was written outside of Wikipedia and then copied to Wikipedia. During the same time in August 2014, the changes in the control group within August 2014 are negligible. Within the following four years, overall quality increases in both treatment and control groups, and interestingly these increases are much smaller than the treatment itself. However, the four-year increase in the treatment group is of a similar size as that in the control group. To further help to understand the context of the quality measures, figures B.6a–B.6b in the Appendix show that the quality measures are correlated with page length.[13]

3.3 Measures of editing activity

To better understand the process of content creation, we also study editing activity. To construct the measures of editing activity, we start with 30,601 edits (revisions) from 180 Wikipedia pages. This set includes all the edits except those generated as part of the treatment in the experiment. Following Aaltonen and Seiler (2016), we restrict the sample of edits in the following ways. First, we exclude edits by bots (about 30% of edits); these are non-human user accounts that generate automated edits. Specifically,

[13] Similar correlations between quantity and quality of content have been found previously, for example by Chen et al. (2019)

Electronic copy available at: https://ssrn.com/abstract=3341630



(a) Overall quality

(b) Completeness

(c) Illustrated

(d) Interesting

(e) Well-written

(f) Similarity to Spanish

Figure 3: Change in quality during the treatment month (August 2014) and within four years after treatment (from September 2014 until September 2018), separately for the control and treatment groups.

Notes: The number of observations is 90 in the control and 90 in the treatment groups.

Electronic copy available at: https://ssrn.com/abstract=3341630

we define bots as users whose username occurs in the list of bots (in the English, French, German, or Dutch Wikipedias) or whose username includes “bot”. Second, we exclude reverts, which are edits that restore any previous version of the same page (about 7% of remaining edits). Third, we exclude vandalism (about 0.8% of remaining edits). We use the following criteria to classify an edit as vandalism: (a) an edit that only deletes text from the previous revision, and (b) the revision immediately after vandalism reverts the article back to a past revision. Then we are left with 19,586 productive edits generated by human users.

To analyze the impact of treatment on editing activity, we construct three types of monthly measures that characterize how many people edited the pages, how many times they edited, and how much they edited. The first measure is the number of unique users editing a page per month. We define a unique user by the username for registered users and by IP address for anonymous users. The second measure is the number of edits per month. To avoid double-counting of micro-edits,[14] we first aggregate edits to the day-user-page level and then sum these up to month-page level. The third measure is edit distance—the number of characters an edit added plus the number of characters it deleted compared to the previous version of the page.[15] We aggregate the edit distance measure to monthly level. Figure B.7 in appendix B describes the average editing activity in the treatment and control groups over time.

In addition to the aggregate measures of editing activity, we separate edits that directly modify the treatment text and those that modify other parts of the page. We classify edits into these two categories using a method similar to Hinnosaar et al. (2019). For each page in the treatment group, we use the diff algorithm between the revision before and after the treatment to determine treatment text—the exact text added by the treatment. For each revision post-treatment, using the diff algorithm between the treatment text and this revision, we check whether the revision deletes any part of the treatment text. If the revision doesn’t delete anything from the treatment text, we classify the revision as one that edited other parts of the page.

[14] Many Wikipedia editors save many revisions to the same page in a short period of time, for example, generating a new revision after each sentence they write. This is partly motivated by the fact that someone else might edit the page at the same time.

[15] Edit distance is widely used in computational linguistics and computer science to measure the similarity of strings. It is a generic term that allows any weights of insert, delete, and substitution operations. Common variants put weight 1 to addition and deletions, and weight substitutions either by 1 (called Levenshtein distance) or by 2 (the measure we use). For each edit, we calculate the edit distance using PHP FineDiff class at the granularity level of a character.

Electronic copy available at: https://ssrn.com/abstract=3341630

3.4 Similarity to Spanish Wikipedia

To complement the human-rated page quality data, specifically page completeness, we also evaluate the changes in quality by computing the similarity with Spanish Wikipedia. We use the articles in the Spanish language (as of September 1st, 2018) as a benchmark as they provide the most detailed coverage of Spanish cities among all language versions on Wikipedia. Therefore, if the coverage of topics in another language becomes more similar to the corresponding article in the Spanish language, this can be interpreted as increasing completeness of the article.

As a measure of similarity, we use the Tversky index (Tversky, 1977). Formally, if the set of terms mentioned in the article of the target language is A and the set of terms mentioned in the Spanish article is B, then the Tversky index is computed as

$$S(A,B) = \frac{|A \cap B|}{|A \cap B| + \alpha |A \setminus B| + \beta |B \setminus A|}, \tag{1}$$

where $\alpha \geq 0$ and $\beta \geq 0$ are parameters. If $\alpha = \beta = 1$, the index becomes equivalent to the Jaccard similarity index, and if $\alpha = \beta = 0.5$, it simplifies to the Sørensen-Dice similarity index.[16] More generally, a larger α puts a bigger weight on A (interpreted as a variant) and β a bigger weight on B (interpreted as a prototype). As our aim is to use the Tversky index as a completeness measure (how complete is A in comparison to B), we set $\alpha = 0$ and $\beta = 1$, which simplifies the similarity measure to $S(A,B) = \frac{|A \cap B|}{|B|}$.

Before comparing the articles in the treatment languages with their Spanish counterparts, we need to extract the set of terms mentioned in each article. For this, we first translate all articles into English using Yandex Translate API.[17] Then, we remove all so-called stop words, i.e., words that do not reflect the specifics of the content.[18] We then use the Porter stemming algorithm (Porter, 1980) to remove the endings from words in English so that only root words remain. Finally, we drop all remaining strings that are shorter than three characters or contain non-alphabetic characters. This process gives us a set of terms for each article that we use in computing the Tversky index.[19]

Figure 3f presents the change in similarity during the treatment and within four years after treatment. As with quality, we see a large increase in the treatment group in

[16] Both indexes are used to measure the similarity between two documents. The Jaccard index is also known as the Intersection over Union and sometimes called the Tanimoto similarity. Another similarity measure used in earlier works in economics (Thompson and Hanley, 2018; Chen et al., 2019) is cosine similarity. It would be preferable if we want to capture the similarity of two pages not only in terms of content covered but the language and tone. As we aim to capture completeness of the page compared to a benchmark, we found the Tversky index most appropriate for the task.

[17] For more information, see `https://tech.yandex.com/translate/`.

[18] For example, pronouns ("it", "their") and prepositions ("on", "before"). The full list is in figure B.5.

[19] Terms such as "america", "archipelago", "area", "arona"; about 250–500 terms for each article.

Electronic copy available at: https://ssrn.com/abstract=3341630

similarity during treatment and much smaller increases within the following four years in both treatment and control groups. Figures B.6c–B.6d in the appendix show that similarity to Spanish is correlated with page length and also with completeness (compared to English).

3.5 Summary statistics

Comparison of pages in treatment and control groups. Table 1 presents the comparison of pre-treatment page length, quality, and editing activity in the treatment group versus the control group. The table shows that there were no significant differences between the two groups before the treatment. However, as we saw earlier, the treatment and control groups are not identical either. Figures B.8–B.9 in the appendix present the kernel density estimates of full distributions separately for the treatment and control groups for all the variables in table 1. We conclude that while there are differences between the groups, by and large, the randomization was rather successful.

Table 1: Comparison of pre-treatment characteristics in the treatment group versus the control group.

	Control group mean	Treatment group mean	t-test p-value	Wilcoxon test p-value	Obs.
	(1)	(2)	(3)	(4)	(5)
Log. length before treatment	8.586	8.611	0.842	0.655	180
Quality rating before treatment	72.700	70.917	0.810	0.602	180
Quality: complete before treatment	71.656	71.706	0.996	0.758	180
Quality: interesing before treatment	75.800	72.917	0.777	0.561	180
Quality: well-written before treatment	92.867	93.122	0.858	0.388	180
Quality: illustrated before treatment	72.372	73.772	0.854	0.441	180
Similarity to Spanish before treatment	17.488	17.251	0.885	0.820	180
Aver. # of users before treatment	0.378	0.333	0.321	0.285	180
Aver. # of edits before treatment	0.396	0.351	0.341	0.364	180
Aver. edit dist. before treatment	79.013	76.786	0.882	0.738	180
Aver. capped edit dist. before treatment	42.511	38.238	0.470	0.583	180

Notes: Column 1 and 2 present the means of pre-treatment values of variables, separately for the control and the treatment group. Column 3 presents the p-value of the t-test for whether the difference between the control and treatment groups is significantly different from zero. Column 4 presents the p-value of the corresponding Wilcoxon rank-sum test. Column 5 presents the number of observations used in each test.

Are the pages in our sample already rather complete? To answer the question, ideally, we would like to use Wikipedia's own quality ratings. Unfortunately, the pages in

Electronic copy available at: https://ssrn.com/abstract=3341630

our sample have not been rated. Each language edition of Wikipedia has its own quality rating system, but only in the English Wikipedia is the system widely used. Therefore, we assess the completeness of the pages in two steps. First, we determine the quality of the corresponding city pages in the English Wikipedia. Then we compare the pages in our sample to the English pages.

Figure 4 presents the distribution of the quality ratings in English Wikipedia of the 60 city pages that correspond to the pages in our sample. These city pages have been assigned the lowest possible grades (Stub, Start, and class C) or have not been rated at all (9 cities).[20] The highest rated pages, class C, according to the Wikipedia rating scale, still miss important content. Hence, we conclude that the 60 city pages are all low-quality articles in the English Wikipedia.[21]

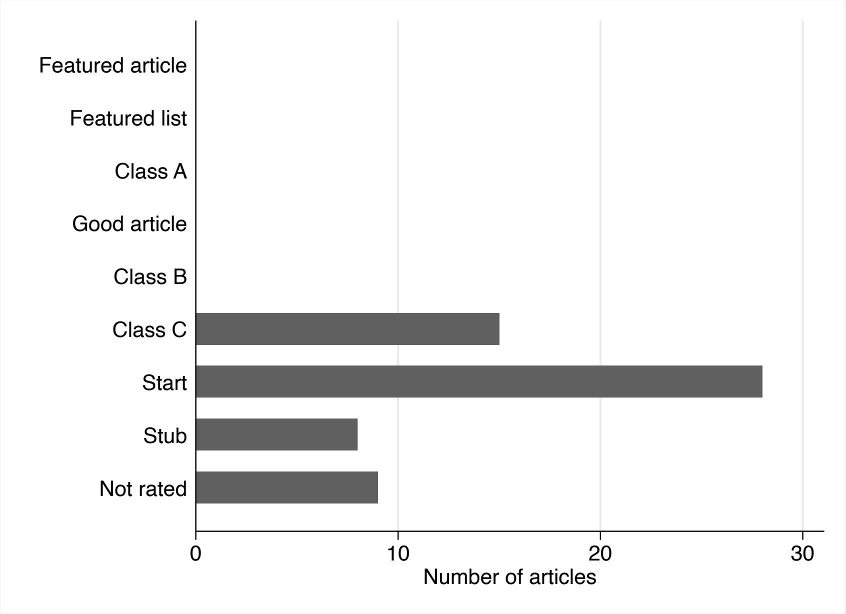


Figure 4: Quality of the 60 city pages (corresponding to our sample) in the English Wikipedia

Having seen that all the English pages still have room for improvement, how do the pages in the sample compare to the English pages? Table 2 compares our pages to the pages in English in 3 dimensions: relative length, (calculated) similarity, and (human

[20] The three lowest grades in the Wikipedia content assessment are: Stub—a very basic description of the topic; Start—developing but still quite incomplete; Class C—substantial but is still missing important content or contains much irrelevant material. Source: `https://en.wikipedia.org/wiki/Wikipedia:Content_assessment`.

[21] Note that these articles, on average, are also not very important according to the English Wikipedia article importance scheme. The scheme uses ratings from Low, Mid, High, to Top. Only 7 of the pages are rated as High importance, 15 as Mid, and 9 as Low importance, 29 of the articles have not been assigned an importance rating, which probably also implies that those articles are not highly important.

Electronic copy available at: https://ssrn.com/abstract=3341630

rated) completeness. Panel A column 1 shows that after treatment the median page in the treatment group is about 75% of the relative length, about 33% in terms of similarity, and almost 100% in terms of completeness compared to the corresponding page in English. Recall here that we saw above that the pages in English were far from complete. Panel A, column 2 shows that, as expected, after treatment the median page in the control group is relatively shorter and of relatively lower quality. But before the treatment (panel B), the median pages in the treatment and control groups are similar.

Table 2: Completeness of the median page compared to English and Spanish Wikipedia

	Compared to English		Compared to Spanish	
	Treatment	Control	Treatment	Control
	(1)	(2)	(3)	(4)
	Panel A: After treatment (2014 September)			
Relative length	75.6	53.4	50.2	35.2
Similarity	33.1	27.8	24.1	17.5
Completeness	98.4	72.4	.	.
	Panel B: Before treatment (2014 August)			
Relative length	57.9	53.2	37.7	35.1
Similarity	26.2	27.7	17.3	17.5
Completeness	71.7	71.7	.	.

Notes: Each cell presents the median from 90 pages either in the treatment group (columns 1 and 3) or control group (columns 2 and 4).

Columns 3–4 in table 2 present the comparison with the pages in the Spanish Wikipedia. Compared to Spanish, the pages in the sample are even shorter and less similar. After the treatment, the median page in the treatment group is still only 50% of the length of the corresponding city page in Spanish. While it is not clear that all that material should be included in the French, German, and Italian Wikipedia, at least we can say that there is additional material that was important enough to be included in the Spanish Wikipedia. Also note that the treatment added only material on topics relevant for tourists, such as the city's main sights and culture. Hence, while the treatment made the pages more complete on these topics, on other topics that are typically covered on each city page, such as demographics, economy, education, and government, there is probably still room for improvement.

4 Results

We start by estimating the impact of the treatment on growth after treatment, both in length and quality. We are looking at growth during four years after treatment. After that,

Electronic copy available at: https://ssrn.com/abstract=3341630

we go into more details in two ways. First, we study the effects on different dimensions of quality. Then we analyze the short- and long-term effects using a difference-in-differences estimator.

To estimate the effect of the treatment on growth after treatment, we compare the growth of pages (indexed by i) in the treatment and control group controlling for city and language fixed effects:

$$\Delta y_i = \beta_0 + \beta_1 TreatmentGroup_i + LanguageFE_i + CityFE_i + \varepsilon_i \tag{2}$$

The coefficient of interest is β_1 on $TreatmentGroup_i$, which is an indicator variable that takes value one if the page was assigned to the treatment group and zero if it was assigned to the control group. The outcome variable $\Delta y_i = y_{2018September} - y_{2014September}$ measures the change in outcome from September 2014 to September 2018. Specifically, the outcome variables are the change in the logarithm of page length, change in the overall quality rating, and the change in the similarity to the corresponding Spanish Wikipedia article.[22]

Table 3 presents the estimates of the effect of treatment on subsequent growth over four years post-treatment. All the point estimates are rather small, for example, 7% of the standard deviation for length and 3% for quality.[23] But the estimates are imprecise. The 95%-confidence interval for length is from -7% to +12%, and the ex-post minimum detectable effect size is 13%. Similarly, the 95%-confidence interval for the growth in quality is from -4 to +3 points and the ex-post minimum detectable effect size is 5 points. Note that these bounds are rather large compared to the average four year growth, but small, only about one fourth, compared to the treatment itself. The estimates for the similarity index are analogous to the quality rating. Hence, while we cannot precisely measure small effects, we can rule out large effects on long-term growth.[24]

Table 4 presents the results of the treatment on the growth in different dimensions of quality. In all four dimensions of quality, measuring how complete, interesting, illustrated, and well-written the page is, the point estimate of the impact is rather small, and the estimates are not statistically significant. The largest point estimate is for well-written,

[22] To make the pages comparable, we subtract the length of text added by the treatment from the length of pages in the treatment group after treatment (both in 2014 and in 2018). We do that because the outcome variable measures the percentage change, and hence, without subtracting the length of the treatment text, the treatment group would have a higher base when calculating the percentage, then the same increase in characters would give a smaller percentage for the treatment group.

[23] For expositional clarity, we interpret the coefficient in column 1 as measuring a percentage change in length. This is a logarithmic approximation that performs well when changes are small which is the case in our sample.

[24] Table B.2 in the appendix shows that the results are robust to alternative control variables. Table B.3 shows that the results are also robust when including the Dutch pages either in the control group or estimating intention-to-treat.

Electronic copy available at: https://ssrn.com/abstract=3341630

Table 3: The long-term effect of treatment on the growth in page length and quality. Dependent variable: $y_{2018Sep} - y_{2014Sep}$.

	Change in page length or quality ($y_{2018Sep} - y_{2014Sep}$)		
	ΔLog. page length	ΔQuality rating	ΔSimilarity to Spanish
	(1)	(2)	(3)
Treatment group	0.026	-0.375	-0.590
	(0.048)	(1.777)	(0.624)
Language FE	Yes	Yes	Yes
City FE	Yes	Yes	Yes
Mean dep. var.	0.190	6.589	2.032
SD dep. var.	0.353	11.958	4.409
Adj. R-squared	0.259	0.116	0.199
Observations	180	180	180

Notes: Each column presents estimates from a separate cross-section regression of 180 Wikipedia pages. The dependent variable $\Delta y_i = y_{2018September} - y_{2014September}$ is the change in logarithm of page length (column 1), change in the overall quality rating (column 2), and the change in the similarity to the corresponding Spanish Wikipedia article (column 3). All regressions include language fixed effects and city fixed effects. Standard errors are reported in parentheses.

which is consistent with what we would expect, given that the treatment slightly reduced the quality in this dimension.

Table 4: The long-term effect of treatment on the growth in different dimensions of page quality. Dependent variable: $y_{2018Sep} - y_{2014Sep}$.

	Change in page quality ($y_{2018Sep} - y_{2014Sep}$)			
	ΔComplete	ΔInteresting	ΔIllustrated	ΔWell-written
	(1)	(2)	(3)	(4)
Treatment group	-0.967	-0.150	-2.037	4.658
	(2.254)	(1.851)	(3.752)	(5.980)
Language FE	Yes	Yes	Yes	Yes
City FE	Yes	Yes	Yes	Yes
Mean dep. var.	6.939	5.700	12.186	4.875
SD dep. var.	15.212	12.201	27.280	37.879
Adj. R-squared	0.122	0.080	0.243	0.003
Observations	180	180	180	180

Notes: Each column presents estimates from a separate cross-section regression of 180 Wikipedia pages. The dependent variable $\Delta y_i = y_{2018September} - y_{2014September}$ is the change in the following dimensions of page quality: complete (column 1), interesting (column 2), illustrated (column 3), and well-written (column 4). All regressions include language fixed effects and city fixed effects. Standard errors are reported in parentheses.

Next, we analyze both the short- and long-term effects. We estimate the following

Electronic copy available at: https://ssrn.com/abstract=3341630

difference-in-differences regression using page-level monthly panel data:

$$y_{it} = \sum_{s} \beta_s \cdot \mathbf{1}[Year_t = s] \cdot TreatmentGroup_i + MonthFE_t + LanguageCityFE_i + \varepsilon_{it} \quad (3)$$

where the sum over s is taken over the years $\{-3, -2, -1, 1, 2, 3, 4\}$ and $Year_t$ measures years since the experiment. The year of the experiment is the baseline, which is why $s = 0$ is excluded. The regression includes fixed effects for each page i (language and city pair) and for each time period t. The coefficients of interest are the β-s on $TreatmentGroup_i$ and year dummy interactions. All the year and treatment group interactions, including for pre-treatment years, are presented graphically in figure B.10 in the appendix. Figure B.10 shows that there is no evidence of differential pre-treatment trends.

Table 5 presents the estimates of short- and long-term effects from regression (3). In column 1, the outcome variable is the logarithm of page length minus treatment text.[25] The estimates of the short- and long-term impacts of treatment on page length are all rather small but imprecise. The estimate of the long-term (four-year) effect on page length is similar to that in table 3.

In columns 2 and 3, the outcome variables are the number of users (people editing the page) and edits per month. The treatment increased the number of users and the number of edits during the first two years after the experiment. Specifically, the treatment increased the monthly number of users editing the page by about 0.11 users (column 2) and the monthly number of edits by 0.12–0.13 edits (column 3). However, these increases are only short-lived. In the third and fourth year, for both measures, the effect of treatment is insignificant, and the coefficients are small in magnitude. In the fourth year, we can reject an effect of the same size found in the first years (0.11 users).[26]

What do these editors and edits do in the first two years post-treatment if it has surprisingly little effect on page length? A natural explanation could be that the additional edits simply polish the text added by the treatment. To study this, we re-calculated the number of edits per month while excluding edits that directly edited the text added by the treatment. The estimates using this outcome variable are presented in column 4. The results show that when excluding the edits that directly affect the text added by the treatment, then the treatment effect is much smaller. We conclude that most of the short-term increase in editing comes from editing the content added by the treatment.

In column 5, the outcome variable is the number of characters added plus deleted. The treatment had no statistically significant effect on the measure. Coefficients vary in sign

[25] To make the pages comparable, we subtract the length of text added by the treatment from the length of pages in the treatment group after treatment. Hence, the estimates should be interpreted as the effect of treatment on page length after removing the mechanical increase created by the treatment.

[26] For the long-term (four-year) effect, the ex-post minimum detectable effect size is 0.11 users.

Electronic copy available at: https://ssrn.com/abstract=3341630

Table 5: The short- and long-term effects of treatment on subsequent page length and editing activity.

	Log. length excluding treatment	# users	# edits	# edits excluding treatment	Edit distance	Capped edit distance
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment × year 1	0.045	0.112***	0.118**	0.006	-13.609	12.970
	(0.032)	(0.043)	(0.049)	(0.045)	(40.176)	(9.250)
Treatment × year 2	0.051	0.109**	0.128**	0.067	107.260	17.967*
	(0.039)	(0.047)	(0.056)	(0.055)	(101.633)	(10.414)
Treatment × year 3	0.049	0.008	0.001	-0.046	-37.161	-6.849
	(0.047)	(0.043)	(0.049)	(0.049)	(30.389)	(9.412)
Treatment × year 4	0.020	0.012	0.013	-0.041	2.626	-3.498
	(0.056)	(0.041)	(0.046)	(0.044)	(74.799)	(8.922)
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Treatment ×{ year-1,-2,-3 }	Yes	Yes	Yes	Yes	Yes	Yes
Language-City FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. var.	8.651	0.329	0.347	0.329	89.491	36.062
SD dep. var.	0.813	0.694	0.759	0.740	1120.108	131.787
Observations	17100	17100	17100	17100	17100	17100

Notes: A unit of observation is a page-month pair. The dependent variable is the logarithm of page length minus treatment text (column 1) or a monthly number of users (columns 2), edits (columns 3), edits not editing treatment text (column 4), edit distance (column 5), or capped edit distance (column 6). *Treatment × year 1* is an indicator variable that takes value one during the first year post-treatment if the page belongs to the treatment group and zero otherwise; and similarly for the other years. All regressions include page fixed effects and month fixed effects. All the year and treatment group interactions, including for pre-treatment years, are presented graphically in figure B.10 in the appendix. The sample is a balanced sample from September 2010 to August 2018, excluding the treatment month of August 2014. Standard errors, reported in parentheses, are clustered by page (180 pages). *** Indicates significance at the 1 percent level, ** at a 5 percent level, * at a 10 percent level.

and, with the exception of the second year, are rather small in magnitude. In the second year, the (statistically insignificant) coefficient estimate implies the treatment effect of about 100 characters per month.

Because the distribution of the number of characters added plus deleted has a long tail, in column 6, we use an alternative measure calculated from individual capped edits. The individual edits are capped from above at the 90th percentile. The 90th percentile equals about 500 characters and is about 10 times larger than the median edit. In this way, the capped edit distance measure gives smaller weight to long edits. Estimates in column 6 show that in the second year after the experiment, the treatment increases the capped change in characters. The magnitude of the effect is small—an increase of about 18 characters per month. Provided that the average word length across languages in the experiment is about 10.4 characters, our treatment increased the edit distance by about

Electronic copy available at: https://ssrn.com/abstract=3341630

two words.[27] The small magnitude of the short-term increase is in line with the findings from columns 3–4, which showed that most of the increase in editing comes from editing the content added by the treatment. In later years, the point estimates of the treatment effect are even smaller, and the estimates are statistically insignificant.

Finally, we analyze whether the effect of the treatment is heterogeneous, varying by subgroups of the pages. Tables B.4 and B.5 in the appendix re-estimate the regressions in table 3, allowing the effect of treatment to vary by the quality, completeness, relative length (compared to the pages in Spanish Wikipedia), and the age of the page. While all the estimates are statistically insignificant, the pattern in the point estimates seems to suggest that the treatment effect was larger on lower quality and less complete pages.

Multiple hypothesis testing. We run many tests, and with a large number of tests, we could easily get false positives, i.e., simply by chance, some estimates could turn out to be statistically significant. That is, the probability of incorrectly rejecting at least one null hypothesis is greater than the probability of incorrectly rejecting each individual hypothesis test. To address the concern, we adjust for multiple hypothesis testing by adjusting for the family-wise error rate (the probability of incorrectly rejecting at least one null hypothesis belonging to the same family of hypothesis).

We prefer not to assign all the tested hypotheses to the same family because we view that some of our outcomes are more important than others. Specifically, we are mainly interested in long-term outcomes and on the impact on length and quality, not on the inputs like various measures of editing. Therefore, to adjust the family-wise error rate, we group our hypotheses into the following families: (1) long-term effects on length and quality, (2) short-term effects on editing behavior, (3) long-term effects on editing behavior. Specifically, we use Westfall and Young (1993) multiple hypothesis p-value adjustment as implemented by Jones et al. (2019), employing 10,000 bootstrap draws.

Tables B.6–B.8 in the appendix present the adjusted p-values. As expected, the adjusted p-values are much larger, and therefore our conclusions regarding the long-term effects are unchanged. Using the stricter p-values, the short-term effects on editing activity become statistically insignificant. Therefore, a conservative approach would be to interpret the findings of the short-term effects simply as suggestive. On the other hand, the stricter p-values might be viewed as too conservative, considering that the literature oftentimes does not correct for multiple hypothesis testing, and the number of our tests has increased in part to show robustness. For this reason, in our preferred estimates, we use unadjusted p-values.

[27] Source: `http://www.ravi.io/language-word-lengths`.

Electronic copy available at: https://ssrn.com/abstract=3341630

5 Theoretical framework

To provide a framework for interpreting our results and unify findings from the related literature, we introduce a simple theoretical model of the private provision of public goods. As we showed in the previous section, we find no sizable long-term effects from added content on content growth. The previous literature has found evidence for both large positive effects as well as a negative effect. Our model aims to provide a theoretical framework to discuss how externalities would impact the cumulative growth of output. We then interpret our results in this framework. In the next section, we also compare and discuss the differences in related literature in the context of the model.

5.1 Model

The model focuses on provision of a single public good (a Wikipedia page) by a sequence of agents (editors). The initial state (value) of the public good is $X_0 \geq 0$. Time is discrete and in each period $t \in \mathbb{N}$, one agent t arrives, gets an i.i.d. draw of parameters $(\alpha_t, \beta_t, \gamma_t)$, observes the current state X_{t-1}, chooses a contribution $x_t \geq 0$, which increases the state to $X_t = X_{t-1} + x_t$, and then leaves the model.[28] Agent t gets payoff

$$u_t(x_t, X_{t-1}) = v_t(X_t) + w_t(x_t) - c_t(x_t, X_{t-1}), \tag{4}$$

where $v_t(X_t) = \alpha_t X_t = \alpha(X_{t-1} + x_t)$ is the value of the public good, $w_t(x_t) = \beta_t x_t$ is the private benefit of the agent's own contribution (for example warm-glow), $c_t(x_t, X_{t-1}) = \gamma_t x_t \cdot \left(\frac{x_t}{2} - \mu(X_{t-1})\right)$ is the cost of contribution x_t, and $\mu(X_{t-1})$ is the externality.[29]

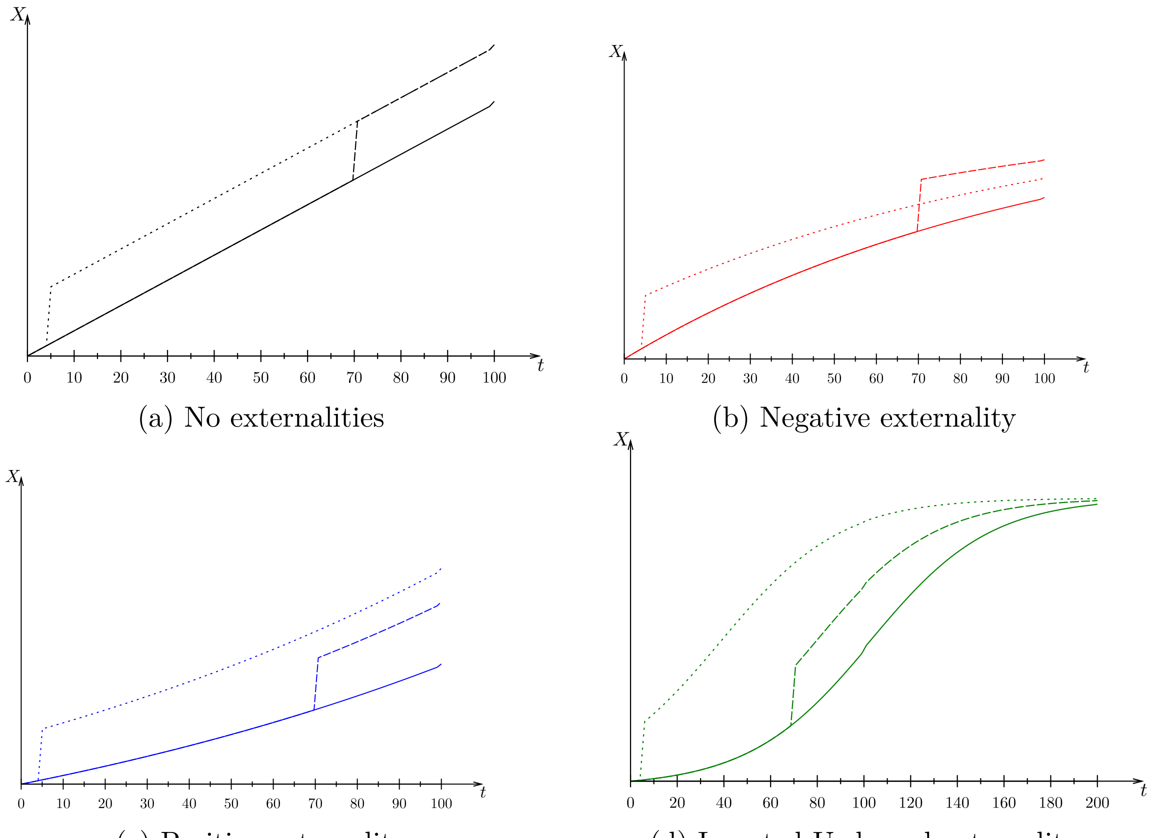
Let us first consider a benchmark with no externalities, i.e., $\mu(X_{t-1}) = 0$.[30] Then the optimal contribution of agent t that maximizes (4) is $x_t^* = \max\left\{0, \frac{\alpha_t+\beta_t}{\gamma_t}\right\}$. Note that in this case, the contributions are independent of the current state and therefore the expected growth rate of contributions is a constant that equals $\mathbb{E}\max\left\{0, \frac{\alpha_t+\beta_t}{\gamma_t}\right\}$. This implies that an exogenous contribution Δ (such as provided by the experiment) has the same effect at any period (parallel shift). This is illustrated by figure 5a.

[28] Extending the model to stochastic arrivals would not change the qualitative results.

[29] For simplicity we assume only cost externalities. A significant simplification of the analysis is that the benefit from the public good only depends on the current state. If agents' benefit would depend on the expected eventual state of the public good, they would have to take into account how their contributions affect the future contributions. This would be a much more complicated sequential game, which under some conditions can be solved using the inverted best-response approach introduced in Hinnosaar (2018). Qualitative implications would remain the same without these simplifications.

[30] Without externalities, this model is similar to the model in Chen et al. (2019), with two differences: the payoffs in their model depend on social impact (i.e., number of viewers) and participation is endogenous. A crucial difference in our model is the inclusion of externalities, which we discuss below.

Electronic copy available at: https://ssrn.com/abstract=3341630



(a) No externalities

(b) Negative externality

(c) Positive externality

(d) Inverted-U-shaped externality

Figure 5: The expected growth of the state under different assumptions about the externality function $\mu(X_{t-1})$

Notes: Time (or number of contributors) on the horizontal axis and state of the public good on the vertical axis.

The externality function $\mu(X_{t-1})$ enters the cost function so that the marginal cost is

$$\frac{\partial c_t}{\partial x_t} = \gamma_t \left(x_t - \mu(X_{t-1}) \right). \tag{5}$$

The marginal cost depends linearly on contribution x_t and additively on externality. Therefore, if the externality function $\mu(X_{t-1})$ is constant, the state X_{t-1} does not affect marginal cost. If the externality function $\mu(X_{t-1})$ is decreasing, there is a negative externality, i.e., the marginal cost increases with the state. For example, this may occur with free-riding: the better the current state of the public good, the fewer reasons there are to contribute more. On the other hand, if the externality function $\mu(X_{t-1})$ is increasing, there is a positive externality, i.e., the marginal cost decreases with the state. For example, earlier contributions may give later contributors ideas (inspire) on how to contribute. Perhaps the most realistic case is an inverted-U-shaped $\mu(X_{t-1})$ function. Initially, when the state is low, it is quite costly to contribute, when the state increases,

Electronic copy available at: https://ssrn.com/abstract=3341630

it becomes easier, but eventually, as the state becomes very high, it becomes again more and more costly to find something to contribute.

The optimal contribution of agent t depends on parameters and the externality:

$$x_t^*(X_{t-1}) = \max\left\{0, \frac{\alpha_t + \beta_t}{\gamma_t} + \mu(X_{t-1})\right\} \tag{6}$$

When the externality is negative, the optimal contribution x_t^* becomes smaller as the state X_{t-1} increases (it becomes more costly to contribute). Therefore, the equilibrium growth rate of the state decreases over time. An exogenous addition Δ (for example, an experimental treatment) has a long-term effect smaller than Δ. Figure 5b illustrates this case. On the other hand, if the externality is positive, it becomes easier to contribute. Then, the equilibrium contributions and the growth rate of the state are increasing over time. In this case, an exogenous addition Δ has a long-term effect larger than Δ. See figure 5c for an illustration.

Finally, when the externality function is inverted-U-shaped, then the growth rate is initially increasing as contributions become easier. However, over time, as the state converges to its upper limit, the contributions become more costly, and therefore the growth slows down. An early treatment Δ leads to fast growth by reducing the costs, whereas a later treatment Δ (of the same amount) has a reduced impact as it slows down the growth. Figure 5d illustrates this case.

5.2 Interpretation in relation to theoretical framework

In this subsection, we connect Wikipedia editing with the theoretical framework. The first question to ask is which channels lead to either positive or negative externalities? We can highlight several channels leading to positive externalities. *Attention*: some Wikipedia editors sign up as watchers for a page and get notifications for each time the page is edited. The reason they do it is to maintain the quality of the page, and therefore some of these notifications must lead to future edits. Moreover, contributions increase the number of page views, mostly by increasing visibility in search rankings (Hinnosaar et al., 2019). Some of these additional eyeballs are likely to convert to contributions (Kane and Ransbotham, 2016; Zhu et al., 2020). *Learning and inspiration*: existing content provides contributors new information about a topic and gives them ideas for contributions (Olivera et al., 2008). *Social motives*: additional content signals potential interest in the community (Zhang and Zhu, 2011; Chen et al., 2019). All these channels would lead to positive externalities, i.e., more contributions by earlier editors either make it easier to contribute or raise their benefits from contributions. Without loss of generality,

Electronic copy available at: https://ssrn.com/abstract=3341630

this can be modeled as a reduction of the marginal cost of contributions, as we did in the previous section. These channels likely have a declining impact, i.e., during the early stages of the page development, we would expect the impacts to be larger than for mature pages, which already have a large amount of content.

On the other hand, there are also channels that lead to negative externalities: *Crowding out*: if users add the content that adds most value at the least possible cost ("low hanging fruits"), then new content raises the cost of future contributions. *Freeriding*: new content tells potential editors that someone is already taking care of the edits of this particular page, and they may choose to point their attention elsewhere. *Increasing complexity*: a Wikipedia page should be a coherent reference source; the more content there is, the more possibilities there are for combinations on how the material could be organized and more parts that need to have the same style and structure. All these channels lead to negative externalities, i.e., more contributions by earlier editors either make it more difficult to contribute or reduce their benefits from contributions. Without loss of generality, this can be modeled as an increasing marginal cost of contribution. It is likely that these channels for negative externalities become more prevalent as content matures and converges to completeness.

Combining these observations, in different settings, the total effect may lean either towards positive or negative externalities, depending on which channels are more active. It is likely that for relatively new and incomplete pages, channels inducing positive externalities are more prevalent. As the page matures, channels for positive externalities become less important, and channels for negative externalities more important. This would imply an inverted-U-shaped externality that we saw in the previous section.

The remaining question is how large are these externalities, which is the empirical question that our analysis addresses. Our empirical results show that a treatment Δ has approximately the same effect Δ on the outcomes four years later. This finding is consistent with two possibilities: either there are no externalities (figure 5a), or there are both positive and negative externalities in the same magnitude (figure 5d). Both cases are plausible, as the treatment contributed to relatively mature pages that were still far from being complete. The results about treatment heterogeneity (tables B.4 and B.5) provide some evidence in favor of the second possibility as they seem to indicate that the treatment had a more positive effect on lower quality and less complete pages.

Electronic copy available at: https://ssrn.com/abstract=3341630

6 Discussion

6.1 Comparison with related literature

Kane and Ransbotham (2016), Aaltonen and Seiler (2016), Zhu et al. (2020), and Nagaraj (2019) study the same question and reach different conclusions. Aaltonen and Seiler (2016) and Kane and Ransbotham (2016) use detailed observational data from Wikipedia, and Zhu et al. (2020) a natural experiment on Wikipedia, and they all found positive externalities—contributions bring more contributions in the future. On the other hand, Nagaraj (2019) uses a natural experiment in a Wikipedia-style mapping service and finds a negative impact on the long-term quality of output (about ten percent higher error rate). Our empirical results and the theoretical model enable us to bridge the gap between their opposing conclusions.

Aaltonen and Seiler (2016) and Kane and Ransbotham (2016) study the impact of added content on the short-term editing activity. The estimates from Aaltonen and Seiler (2016) imply that additional content of 2,000 characters leads to about 0.18 additional monthly users. Zhu et al. (2020) estimate the impact over six months and find an effect of similar magnitude, about 0.21 additional monthly contributors per 2,000 added characters. Our estimates for the first two years after the experiment are slightly smaller, about 0.11 additional users per month. Therefore our findings largely confirm the findings from related literature. The small difference in magnitude might be explained by the fact that the other papers are measuring more immediate effects, while we measure the impact over a year or two.

However, we find that there is no sizable long-term impact on editing activity or content growth. This finding is related to the results in Kane and Ransbotham (2016), who find that although additional content may bring additional contributions to pages that are relatively incomplete, this effect may disappear once the pages become complete. Our results on treatment heterogeneity provide some evidence towards this, although the effects are not statistically significant.

As we discussed in the previous section, such differences in editing activity are consistent with a simple explanation that in earlier stages of the content life-cycle, the channels leading to positive externalities are more active. In contrast, the channels leading to negative externalities become more dominant in the long-term. In particular, as some editors have signed up as watchers who make sure that added content is up to the quality standards (see section 2.3), we would expect that immediately after content is added there is increased editing activity focusing on the added content. Our analysis of the short-term

Electronic copy available at: https://ssrn.com/abstract=3341630

editing activity supports this conjecture.[31]

On the other hand, Nagaraj (2019) studies the impact of better-quality seeding data on quality outcomes ten years later. He finds a negative effect of about 10%, which is a large difference from the conclusion of Aaltonen and Seiler (2016), whose simulations implied that the positive effect on short-term editing activity could lead to about 45% better output. Our results about the impact of added content on the long-term growth in content quantity and quality provide some support in favor of the first number. While we do not have enough power to measure small positive and small negative effects, we can reject sizable long-term impacts.

The finding from Nagaraj (2019) is partially consistent with the implications of our theoretical framework. Assuming that a time horizon of ten years is sufficient for the content to be close to completeness, we would expect that early differences in content would disappear in the long-term. In other words, in the long term, we would expect the negative externality to dominate. Explaining the negative impact on outcomes requires an explanation beyond our model and analysis. The key mechanism proposed by Nagaraj (2019) to explain the negative externality is the "ownership effect", which plays a less prominent role in Wikipedia. Nagaraj (2019) suggests that contributors who added particular bridges or streets on the user-generated map may feel more responsible for keeping these objects updated over time. Therefore, the treatment of adding more seeding information may backfire by not allowing the ownership of objects to arise naturally. All other papers mentioned here, including our work, focus on textual content in Wikipedia, where ownership is less clear, and we would thus expect the negative effect of adding content to be less prominent.

6.2 Implications

Many user-generated content platforms use managerial interventions that aim at motivating users to contribute new content. Examples include seeding the platform with initial content, compensating users for their contributions, or running campaigns to help to get the process started. Whether such policies should be used depends on whether the added content inspires an upward spiral of more user-generated content or whether it discourages future contributions. This choice is a critical managerial decision, because firm wikis, archives, or Q&A forums all depend on sufficient provision of information. On the other hand, such interventions are costly and require committing resources that cannot be in-

[31] Note that there are other, more subtle differences in the research environments that may affect the outcomes. For example, in our setting, the added content comes from an outside source. Instead, in the settings of Kane and Ransbotham (2016), Aaltonen and Seiler (2016), and Zhu et al. (2020), the added content is created by the community.

Electronic copy available at: https://ssrn.com/abstract=3341630

vested elsewhere. Hence, it is important to understand not only the direction but also the magnitude of any possible externality that added content might have on follow-on contributions.

Our findings have a clear policy implication. The additional content temporarily stimulates the participation of users in the areas where the content was added. This participation, however, has no sizable cumulative effect on growth. Therefore, decisions regarding investments in information seeding and incentivizing contributions should primarily be based on direct cost-benefit analysis: they pay off if and only if the costs of creating the content are lower than the value of the new content to the users and eventually to the platform. The additional costs or benefits via externalities that discourage or inspire future contributions are small.

6.3 Generalizability

As we already discussed, many channels may lead to positive and negative externalities. Which channels are more active depends on a particular setting, time of treatment, and time of measurement of the outputs. We already saw how the differences in timing within the page lifecycle and time horizon might lead to different conclusions.

A word of caution is in order here. The results from Wikipedia might not generalize to other user-generated content platforms. As an example, a relevant difference among the platforms is the magnitude of the contributor's personal benefit. In Wikipedia, the personal benefit from contributing is likely to be smaller than in open-maps or open-source software. For example, a user of open-maps could directly benefit from correcting a mistake on a map, while an error in Wikipedia is unlikely to have any personal consequences. Another example is the "ownership effect" proposed by Nagaraj (2019), which seems to be a key driver in the Wikipedia-style mapping service, where each object (street, building, etc.) explicitly states who has edited this object. As Wikipedia does not display who wrote each part of the page, the ownership assignment is less clear. Therefore, we would expect this channel for negative externality to be less active.

Moreover, our results also might not generalize to settings in other stages in content development. Our theoretical model highlighted that the externalities might depend on the existing amount of available content, additional content in early stages being more beneficial than later. Our empirical results about the heterogeneity provide some suggestive evidence that this might be the case. However, the experiment was not designed to analyze this heterogeneity and is underpowered to do that. We would encourage future research that aims at uncovering how the existing content (or more generally the lifecycle of the content development) affects the externalities in content creation.

Electronic copy available at: https://ssrn.com/abstract=3341630

7 Conclusions

In this paper, we show that the addition of content has no sizable impact on the subsequent long-term growth of content. We identify the causal effect using exogenous variation from a randomized field experiment in Wikipedia. We find that the treatment, which added content to randomly chosen Wikipedia pages, increased subsequent content generation in the first two years but had no sizable impact on the long-term growth of content, both in terms of its quantity and quality. Specifically, we find evidence of temporary increases in user participation, in particular, increases in the number of edits and editors in the first two years after the treatment. However, the amount of content these users added was small, and most of their edits modified the content added by the treatment.

Our findings have a clear policy implication—in settings like the one studied here, information seeding and motivating content creation is not necessarily enough to generate a sizable increase in future content generation. However, these policies are also not counterproductive as they can stimulate a small number of additional edits, and the discouragement effect on future contributions is also small. Therefore, it is mostly a matter of direct cost-benefit analysis whether such policies pay off.

Our results may not generalize to settings where other channels leading to positive or negative externalities are more prominent. For example, it is possible that in the early stages of content development, information seeding may be beneficial. On the other hand, it is also possible that in situations where individual contributions are well-identified, the seeding policies may backfire.

References

Aaltonen, A. and S. Seiler (2016): "Cumulative Growth in User-Generated Content Production: Evidence from Wikipedia," *Management Science*, 62, 2054–2069.

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

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.

Chen, Y., R. Farzan, R. E. Kraut, I. YeckehZaare, and A. F. Zhang (2019): "Motivating Contributions to Public Information Goods : A Personalized Field Experiment on Wikipedia," *manuscript.*

Electronic copy available at: https://ssrn.com/abstract=3341630

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.

De Giorgi, G., M. Pellizzari, and S. Redaelli (2010): “Identification of Social Interactions Through Partially Overlapping Peer Groups,” *American Economic Journal: Applied Economics*, 2, 241–75.

Duflo, E. and E. Saez (2002): “Participation and Investment Decisions in a Retirement Plan: The Influence of Colleagues’ Choices,” *Journal of Public Economics*, 85, 121–148.

——— (2003): “The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence From a Randomized Experiment,” *Quarterly Journal of Economics*, 118, 815–842.

Egebark, J. and M. Ekström (2017): “Liking What Others "Like": Using Facebook to Identify Determinants of Conformity,” *Experimental Economics*, 1–22.

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

Fershtman, C. and N. Gandal (2011): “Direct and Indirect Knowledge Spillovers: the "Social Network" of Open-Source Projects,” *RAND Journal of Economics*, 42, 70–91.

Gallus, J. (2017): “Fostering Public Good Contributions with Symbolic Awards: A Large-Scale Natural Field Experiment at Wikipedia,” *Management Science*, 63, 3999–4015.

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

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

Grossman, G. M. and E. Helpman (1993): *Innovation and Growth in the Global Economy*, MIT press.

Electronic copy available at: https://ssrn.com/abstract=3341630

Hanushek, E. A., J. F. Kain, J. M. Markman, and S. G. Rivkin (2003): “Does Peer Ability Affect Student Achievement?” *Journal of Applied Econometrics*, 18, 527–544.

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): “Wikipedia Matters,” *Journal of Economics & Management Strategy*, forthcoming.

Hinnosaar, T. (2018): “Optimal Sequential Contests,” *manuscript.*

Huang, N., G. Burtch, B. Gu, Y. Hong, C. Liang, K. Wang, D. Fu, and B. Yang (2018): “Motivating User Generated Content with Performance Feedback: Evidence from Randomized Field Experiments,” *Management Science*, forthcoming.

Jones, C. I. (1995): “R&D-Based Models of Economic Growth,” *Journal of Political Economy*, 103, 759–784.

Jones, D., D. Molitor, and J. Reif (2019): “What Do Workplace Wellness Programs Do? Evidence From the Illinois Workplace Wellness Study,” *Quarterly Journal of Economics*, 134, 1747–1791.

Kane, G. C. and S. Ransbotham (2016): “Content as Community Regulator: The Recursive Relationship Between Consumption and Contribution in Open Collaboration Communities,” *Organization Science*, 27, 1258–1274.

Kiyotaki, N. and R. Wright (1989): “On Money as a Medium of Exchange,” *Journal of Political Economy*, 97, 927–954.

Kummer, M. (2014): “Spillovers in networks of user generated content: Pseudo-experimental evidence on Wikipedia,” *manuscript.*

Kummer, M., O. Slivko, and X. Zhang (2019): “Unemployment and Digital Public Goods Contribution,” *Information Systems Research*, forthcoming.

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.

Electronic copy available at: https://ssrn.com/abstract=3341630

Manski, C. F. (1993): “Identification of Endogenous Social Effects: The Reflection Problem,” *Review of Economic Studies*, 60, 531–542.

Nagaraj, A. (2019): “Information Seeding and Knowledge Production in Online Communities: Evidence from OpenStreetMap,” *manuscript.*

Nov, O. (2007): “What Motivates Wikipedians?” *Communications of the ACM*, 50, 60–64.

Olivera, F., P. S. Goodman, and S. S.-L. Tan (2008): “Contribution Behaviors in Distributed Environments,” *MIS Quarterly*, 32, 23–42.

Porter, M. F. (1980): “An Algorithm for Suffix Stripping,” *Program*, 14, 130–137.

Ransbotham, S., G. C. Kane, and N. H. Lurie (2012): “Network Characteristics and the Value of Collaborative User-Generated Content,” *Marketing Science*, 31, 387–405.

Ren, Y., J. Chen, and J. Riedl (2015): “The Impact and Evolution of Group Diversity in Online Open Collaboration,” *Management Science*, 62, 1668–1686.

Romer, P. M. (1990): “Endogenous Technological Change,” *Journal of Political Economy*, 98, S71–S102.

Salganik, M. J., P. S. Dodds, and D. J. Watts (2006): “Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market,” *Science*, 311, 854–856.

Shah, S. K. (2006): “Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development,” *Management Science*, 52, 1000–1014.

Slivko, O. (2018): “Online "Brain Gain": Do Immigrants Return Knowledge Home?” *manuscript.*

Sun, Y., X. Dong, and S. McIntyre (2017): “Motivation of User-Generated Content: Social Connectedness Moderates the Effects of Monetary Rewards,” *Marketing Science*, 36, 329–337.

Thompson, N. and D. Hanley (2018): “Science Is Shaped by Wikipedia: Evidence From a Randomized Control Trial,” *manuscript.*

Tversky, A. (1977): “Features of Similarity,” *Psychological Review*, 84, 327–352.

Electronic copy available at: https://ssrn.com/abstract=3341630

Westfall, P. H. and S. S. Young (1993): *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, vol. 279, John Wiley & Sons.

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.

Zhu, K., D. Walker, and L. Muchnik (2020): "Content Growth and Attention Contagion in Information Networks: A Natural Experiment on Wikipedia," *Information Systems Research*, forthcoming.

Electronic copy available at: https://ssrn.com/abstract=3341630

A Online Appendix: Power analysis

We conduct the power analysis for one of our main outcome variable, page length, and the main editing activity variable, number of users, using data from years 2010–2014 before the experiment and Monte Carlo simulations with 10,000 samples. We draw random samples (with replacement) of size 60 of cities. That gives us the samples of pages of size either 240 (with Dutch pages) or 180 (without Dutch pages). We randomize according to the randomization protocol described in section 2.1 and estimate the following regressions.

In the case of page length, we compare the growth of pages (indexed by i) in the treatment and control group controlling for language fixed effects:

$$\Delta logLength_i = \beta_0 + \beta_1 TreatmentGroup_i + LanguageFE_i + \varepsilon_i \tag{7}$$

The outcome variable is the change in the logarithm of page length from 2010 to 2014 (that is growth during four years before treatment).

In the case of the number of users, we compare the average number of uses in the treatment and control group, controlling for past number of users and language fixed effects:

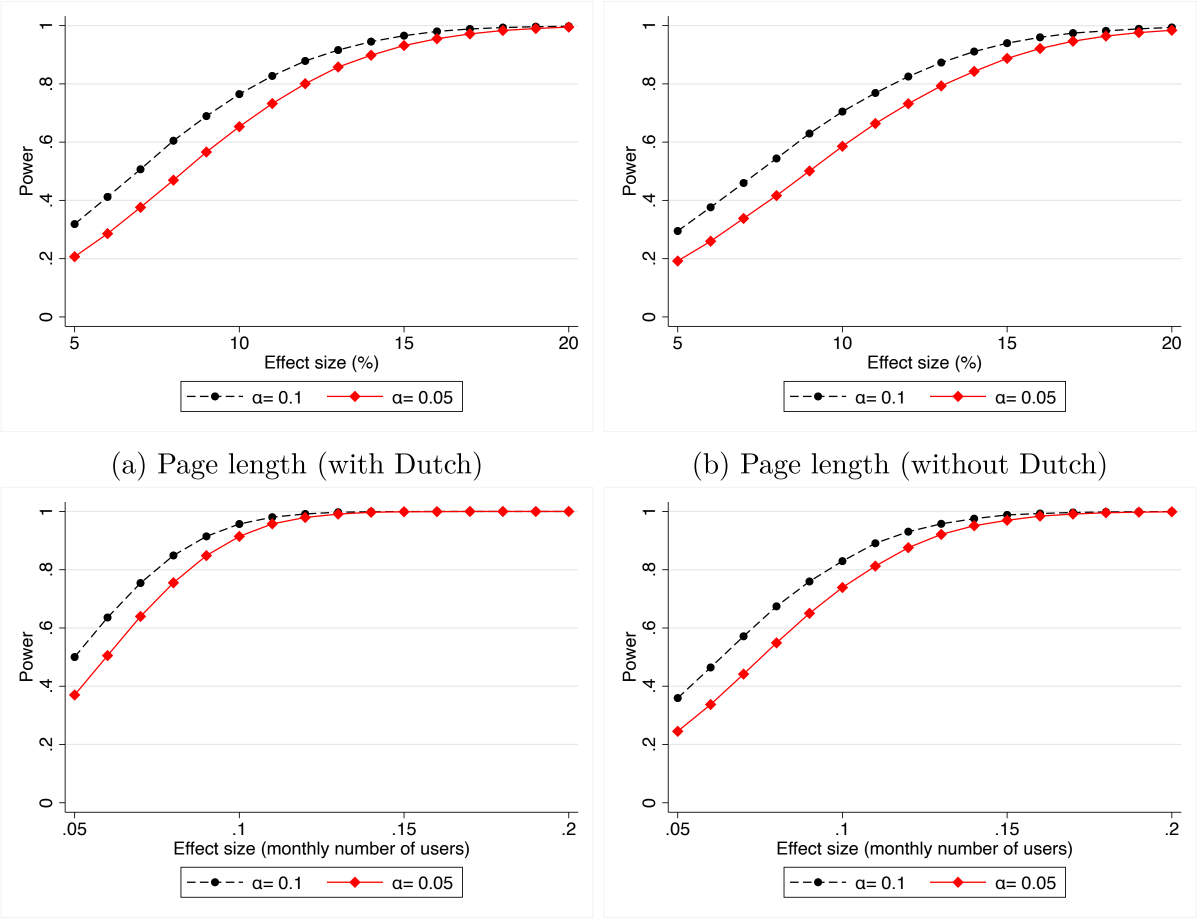
$$Users_{2014,i} = \beta_0 + \beta_1 TreatmentGroup_i + \beta_2 Users_{2011,i} + LanguageFE_i + \varepsilon_i \tag{8}$$

The outcome variable $Users_{2014,i}$ is the yearly average of the number of monthly users from August 2013 to July 2014, and we calculate $Users_{2011,i}$ analogously. Note that since this cross-section regression includes the lagged outcome variable $Users_{2011,i}$, we expect the estimates to be similar to a difference-in-differences estimator.

Figure A.1 presents the relationship between power and the true effect size at 5% and 10%-significance level. For both page length and the number of users, it shows power with and without the Dutch pages, hence with the sample size of either 240 or 180 pages. Figure A.1a shows that when the Netherlands is included in the sample, as originally intended, then if the true treatment effect is 10% increase in page length over four years, we would reject the null hypothesis of no effect at 10%-significance level with 76% probability and at 5%-significance level with 65% probability. The minimum detectable effect size is about 12%. If we exclude the Netherlands (figure A.1b), we lose some power, but the minimum detectable effect is still around 13%. Figure A.1d shows that even if we exclude the Netherlands, the minimum detectable effect size is 0.11 users, and we should certainly be able to detect the effect sizes suggested in the literature.

To summarize, our study is underpowered to detect small long-term effects on page length, but we can detect even half of the effect-size suggested by Aaltonen and Seiler

Electronic copy available at: https://ssrn.com/abstract=3341630



(a) Page length (with Dutch)

(b) Page length (without Dutch)

(c) Number of users (with Dutch)

(d) Number of users (without Dutch)

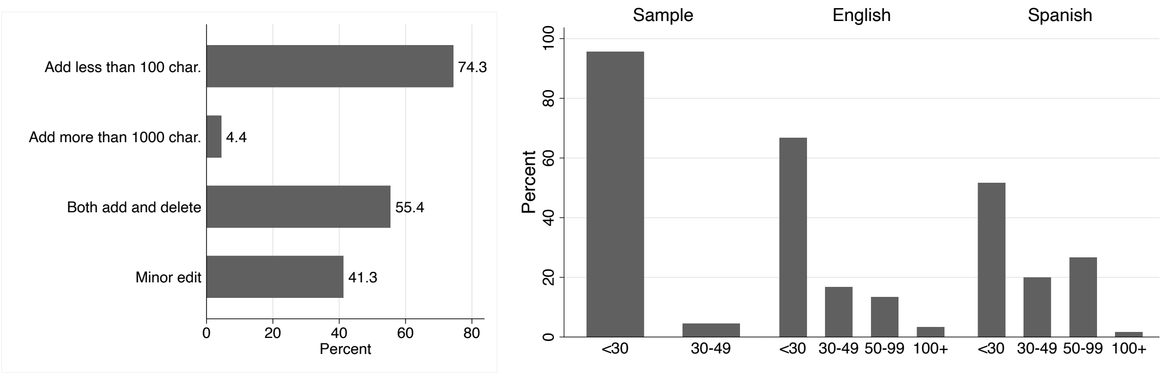
Figure A.1: Power analysis for the effect on the page length and the number of users

Notes: Calculated using data from years 2010–2014 before the experiment and Monte Carlo simulations with 10,000 samples.

(2016). On the number of users, our experiment has relatively more power, being able to detect the effect sizes suggested in the literature.

Electronic copy available at: https://ssrn.com/abstract=3341630

B Online Appendix: Additional figures and tables

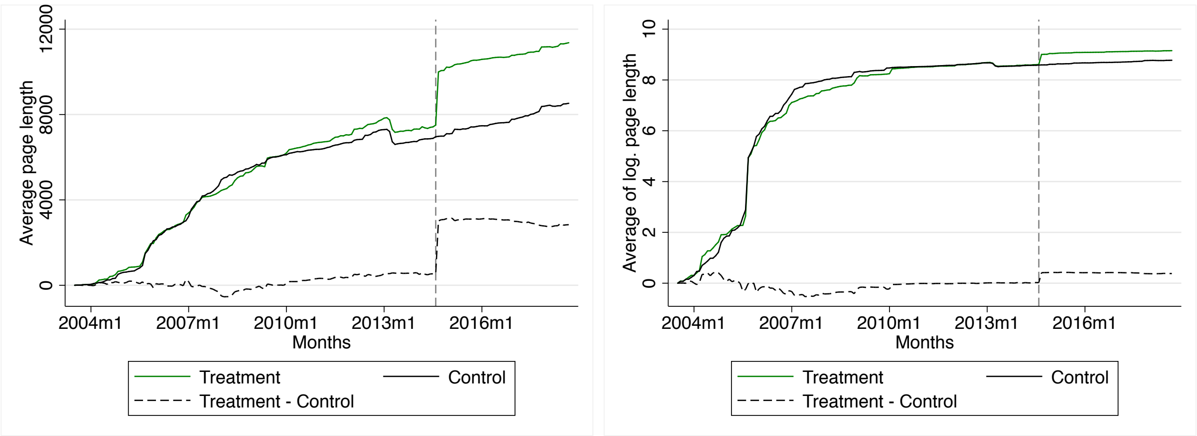


(a) Edits by type (as a percentage) (b) Distribution of pages by the number of watchers

Online Appendix Figure B.1: Edit types and editors watching watching each page

Notes: Figure B.1a presents the edits by type as a percentage out of 100 calculated using the edits in the sample of 180 pages pre-treatment. Figure B.1b presents the distribution of the number of watchers in the sample, in the English Wikipedia, and in the Spanish Wikipedia as measured in January 2020.

Electronic copy available at: https://ssrn.com/abstract=3341630

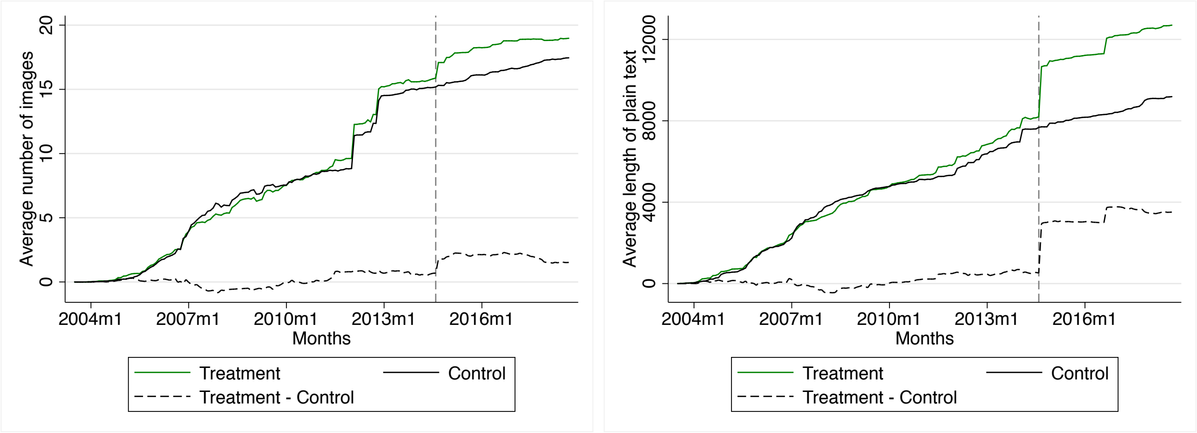


(a) Page length without Cordoba in French

(b) Logarithm of page length

Online Appendix Figure B.2: Robustness: Page length

Notes: The number of observations used to calculate the average is 90 in the control group and 89 (Figure B.2a) or 90 (Figure B.2b) in the treatment group. The experiment month (August 2014) is marked by dashed vertical line.



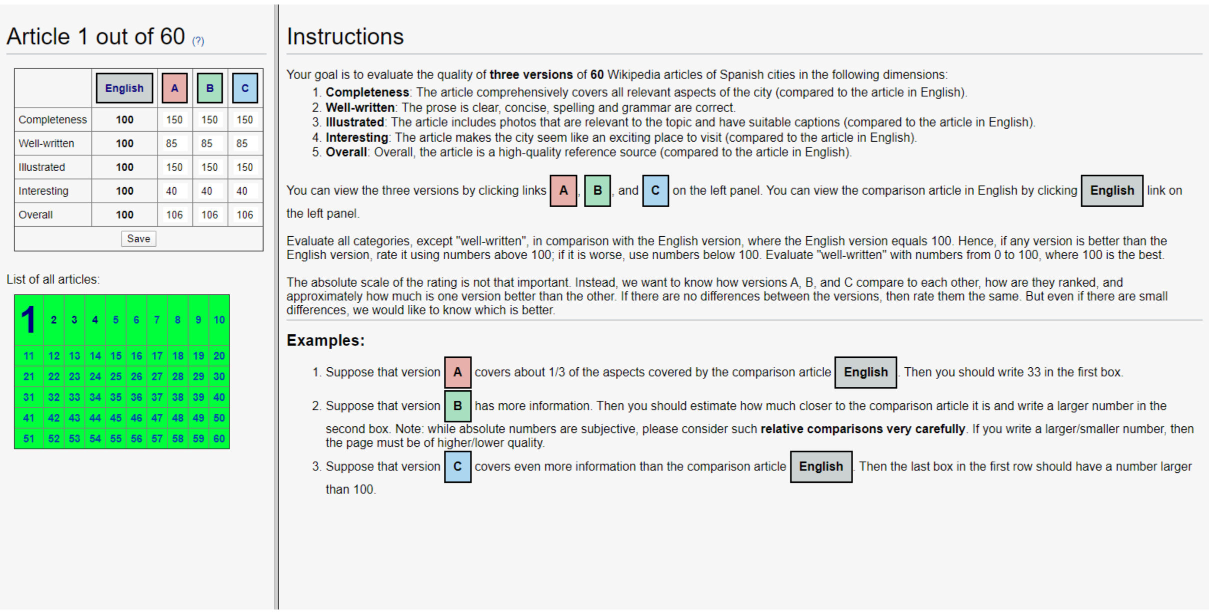
(a) Images

(b) Plain text

Online Appendix Figure B.3: Other output measures

Notes: The number of observations is 90 in the control and 90 in the treatment groups. The experiment month (August 2014) is marked by dashed vertical line. *Plain text* is obtained by removing html elements from the parsed text.

Electronic copy available at: https://ssrn.com/abstract=3341630



Article 1 out of 60 (?)

	English	A	B	C
Completeness	**100**	150	150	150
Well-written	**100**	85	85	85
Illustrated	**100**	150	150	150
Interesting	**100**	40	40	40
Overall	**100**	106	106	106

Save

List of all articles:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60

Instructions

Your goal is to evaluate the quality of **three versions** of **60** Wikipedia articles of Spanish cities in the following dimensions:

1. **Completeness**: The article comprehensively covers all relevant aspects of the city (compared to the article in English).
2. **Well-written**: The prose is clear, concise, spelling and grammar are correct.
3. **Illustrated**: The article includes photos that are relevant to the topic and have suitable captions (compared to the article in English).
4. **Interesting**: The article makes the city seem like an exciting place to visit (compared to the article in English).
5. **Overall**: Overall, the article is a high-quality reference source (compared to the article in English).

You can view the three versions by clicking links **A**, **B**, and **C** on the left panel. You can view the comparison article in English by clicking **English** link on the left panel.

Evaluate all categories, except "well-written", in comparison with the English version, where the English version equals 100. Hence, if any version is better than the English version, rate it using numbers above 100; if it is worse, use numbers below 100. Evaluate "well-written" with numbers from 0 to 100, where 100 is the best.

The absolute scale of the rating is not that important. Instead, we want to know how versions A, B, and C compare to each other, how are they ranked, and approximately how much is one version better than the other. If there are no differences between the versions, then rate them the same. But even if there are small differences, we would like to know which is better.

Examples:

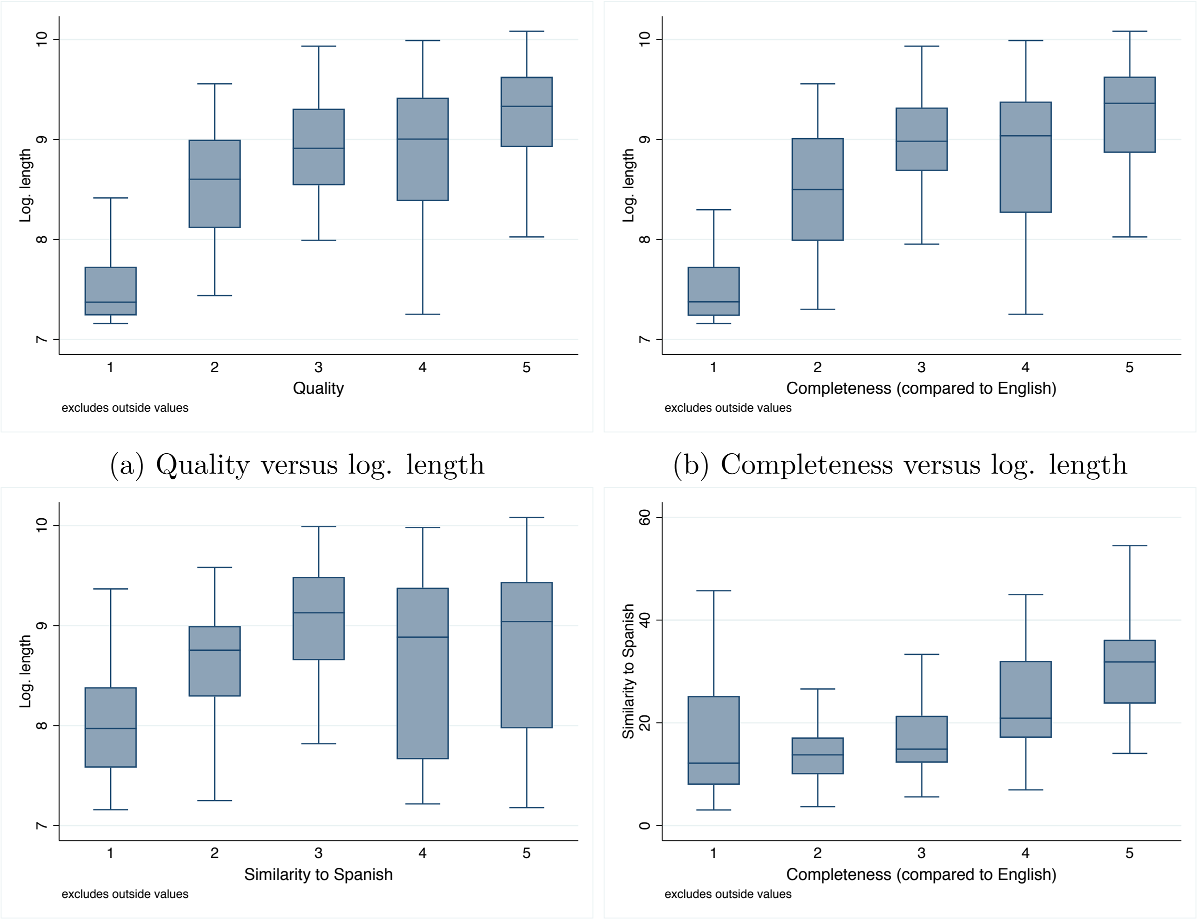
1. Suppose that version **A** covers about 1/3 of the aspects covered by the comparison article **English**. Then you should write 33 in the first box.
2. Suppose that version **B** has more information. Then you should estimate how much closer to the comparison article it is and write a larger number in the second box. Note: while absolute numbers are subjective, please consider such **relative comparisons very carefully**. If you write a larger/smaller number, then the page must be of higher/lower quality.
3. Suppose that version **C** covers even more information than the comparison article **English**. Then the last box in the first row should have a number larger than 100.

Online Appendix Figure B.4: A screenshot with instructions given to research assistants who rated the quality of articles as described in section 3.2

'ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out', 'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'its', 'yours', 'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', 'the', 'themselves', 'until', 'below', 'are', 'we', 'these', 'your', 'his', 'through', 'don', 'nor', 'me', 'were', 'her', 'more', 'himself', 'this', 'down', 'should', 'our', 'their', 'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all', 'no', 'when', 'at', 'any', 'before', 'them', 'same', 'and', 'been', 'have', 'in', 'will', 'on', 'does', 'yourselves', 'then', 'that', 'because', 'what', 'over', 'why', 'so', 'can', 'did', 'not', 'now', 'under', 'he', 'you', 'herself', 'has', 'just', 'where', 'too', 'only', 'myself', 'which', 'those', 'i', 'after', 'few', 'whom', 't', 'being', 'if', 'theirs', 'my', 'against', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', 'than'

Online Appendix Figure B.5: List of stop words used to clean the text for computing the Tversky similarity measure in section 3.4

Electronic copy available at: https://ssrn.com/abstract=3341630



(a) Quality versus log. length

(b) Completeness versus log. length

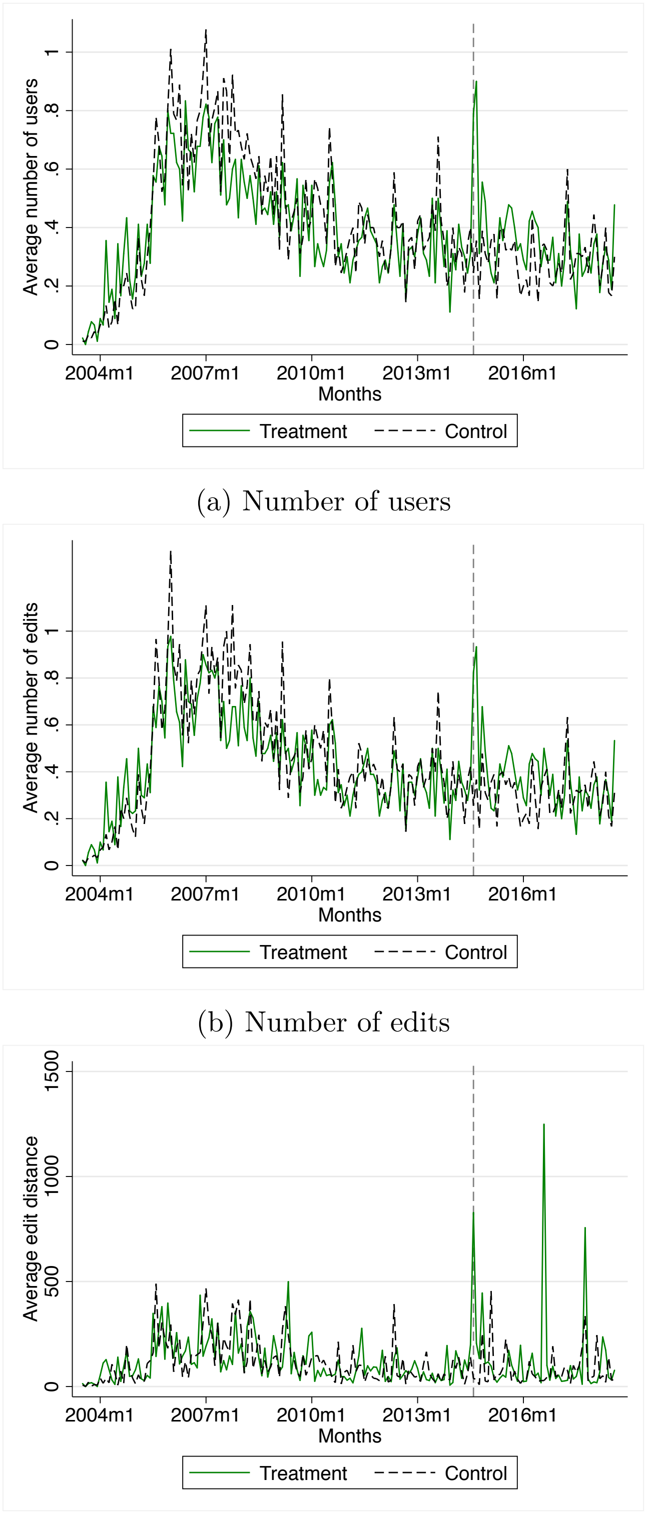
(c) Similarity to Spanish versus log. length

(d) Completeness versus similarity to Spanish

Online Appendix Figure B.6: Quality, completeness (compared to English), similarity (compared to Spanish), and log. length

Notes: We group the characteristic on the horizontal axes into quintiles. For each quantile, the graph presents the median (as a horizontal line) and the interval from the 25th to the 75th percentile (as the box) of the variable on the vertical axes. All measures are from pre-treatment (in August 2014.

Electronic copy available at: https://ssrn.com/abstract=3341630



(a) Number of users

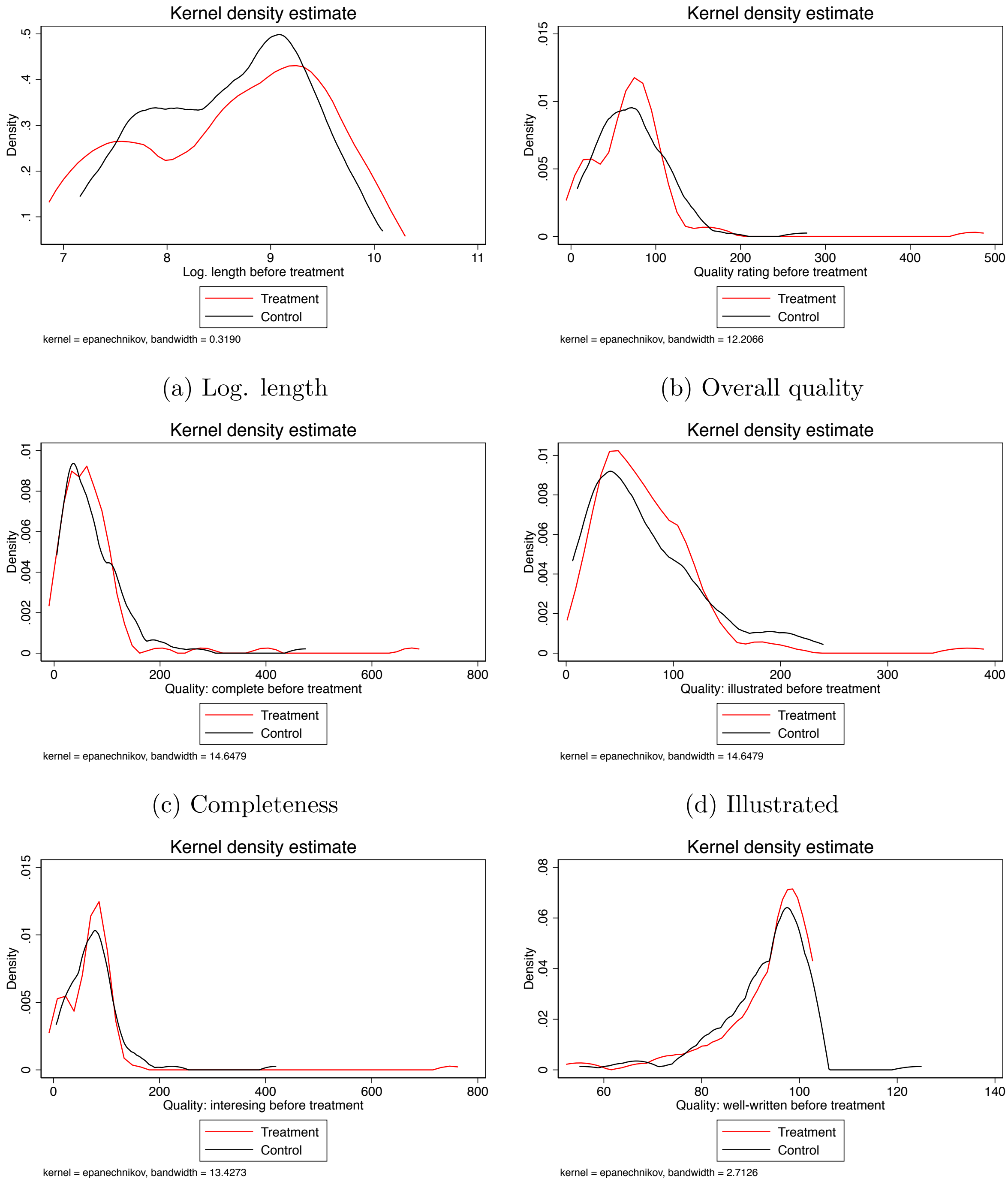
(b) Number of edits

(c) Edit distance

Online Appendix Figure B.7: Average input measures in the treatment and control groups per month

Notes: The number of observations is 90 in the control and 90 in the treatment groups. The experiment month (August 2014) is marked by dashed vertical line.

Electronic copy available at: https://ssrn.com/abstract=3341630



(a) Log. length

(b) Overall quality

(c) Completeness

(d) Illustrated

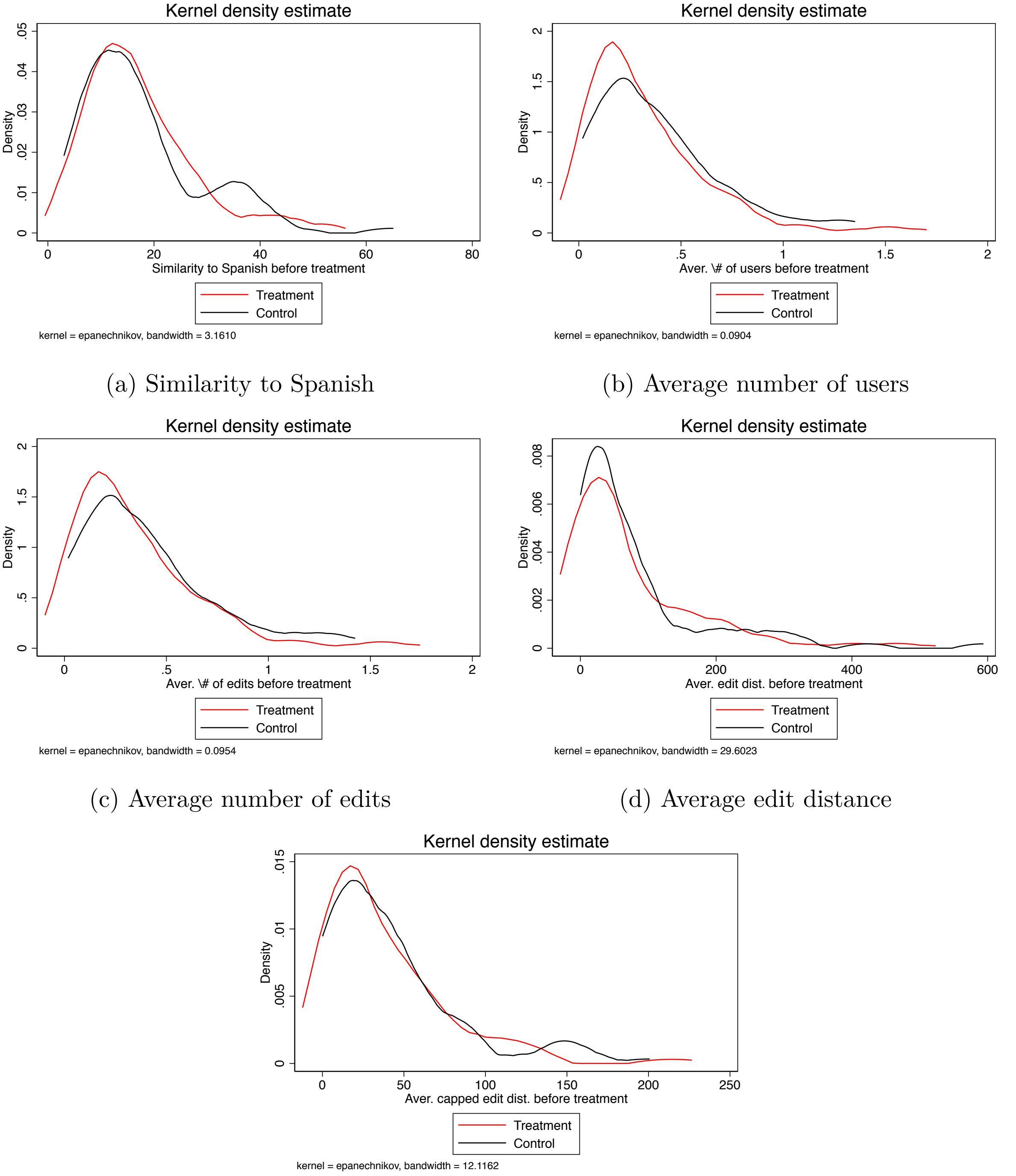
(e) Interesting

(f) Well-written

Online Appendix Figure B.8: Distributions of page length and quality before treatment, separately by treatment and control groups

Notes: Kernel density estimates of the pre-treatment distributions, separately for the control and treatment groups.

Electronic copy available at: https://ssrn.com/abstract=3341630



(a) Similarity to Spanish

(b) Average number of users

(c) Average number of edits

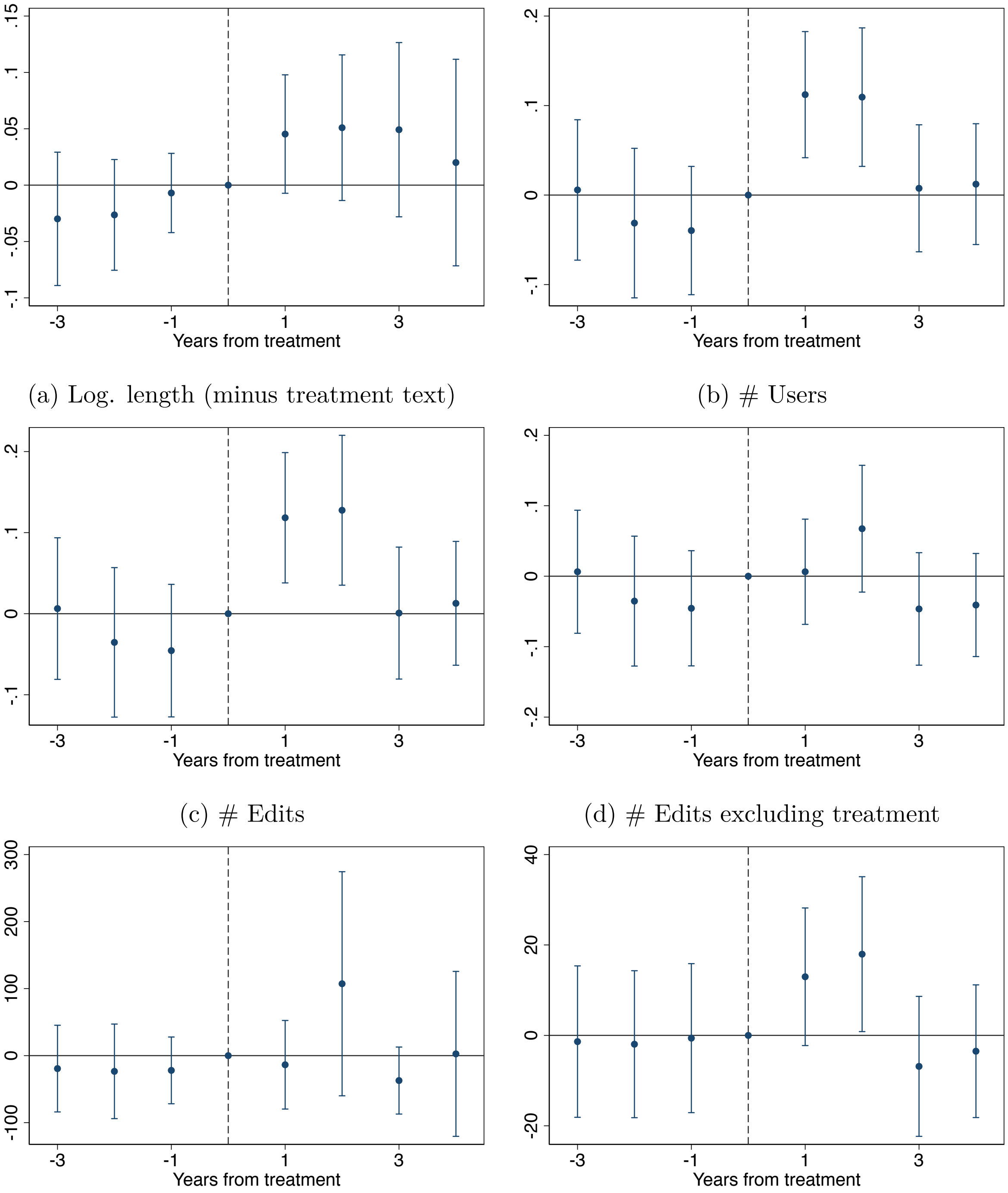
(d) Average edit distance

(e) Average capped edit distance

Online Appendix Figure B.9: Distributions of page similarity to Spanish and editing activity before treatment, separately by treatment and control groups

Notes: Kernel density estimates of the pre-treatment distributions, separately for the control and treatment groups.

Electronic copy available at: https://ssrn.com/abstract=3341630



(a) Log. length (minus treatment text)

(b) # Users

(c) # Edits

(d) # Edits excluding treatment

(e) Edit distance

(f) Capped edit distance

Online Appendix Figure B.10: Panel

Notes: Point estimates and 90% confidence intervals from regressions in table 5 of the treatment group and years since treatment interactions. A unit of observation is a page-month pair. Regressions include page fixed effects and month fixed effects. The sample is a balanced sample from September 2010 to August 2018, excluding the treatment month August 2014.

Electronic copy available at: https://ssrn.com/abstract=3341630

Online Appendix Table B.1: Foreign language reading skills, % of population

	Reading Spanish	Reading English
French	9	32
Germans	2	33
Italians	4	26
	Spanish reading other languages	
Reading English	15	
Reading French	7	
Reading German	1	
Reading Italian	2	

Source: Eurobarometer (2012)

Electronic copy available at: https://ssrn.com/abstract=3341630

Online Appendix Table B.2: Robustness of the effect of treatment on subsequent growth in page length and quality, alternative controls. Dependent variable: $y_{2018Sep} - y_{2014Sep}$.

	Change in page length or quality ($y_{2018Sep} - y_{2014Sep}$)								
	ΔLog. page length			ΔQuality rating			ΔSimilarity to Spanish		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment group	0.012	0.012	0.012	-0.867	-0.867	-0.867	-0.478	-0.478	-0.478
	(0.053)	(0.053)	(0.048)	(1.786)	(1.802)	(1.683)	(0.658)	(0.657)	(0.611)
Group FE	No	Yes	No	No	Yes	No	No	Yes	No
Language FE	No	No	Yes	No	No	Yes	No	No	Yes
Mean dep. var.	0.190	0.190	0.190	6.589	6.589	6.589	2.032	2.032	2.032
SD dep. var.	0.353	0.353	0.353	11.958	11.958	11.958	4.409	4.409	4.409
Observations	180	180	180	180	180	180	180	180	180

Notes: Each column presents estimates from a separate cross-section regression of 180 Wikipedia pages. The dependent variable is the change in logarithm of page length (columns 1-3), change in the overall quality rating (columns 4-6), and the change in the similarity to the corresponding Spanish Wikipedia article (columns 7-9). Regressions in columns 1, 4, and 7 don't include any controls besides the indicator for the treatment group, regressions in columns 2, 5, and 8 include group dummies, and regressions in columns 3, 6, and 9 include language fixed effects. Standard errors are reported in parentheses.

Online Appendix Table B.3: Robustness of the effect of treatment on subsequent growth in page length and quality, Dutch pages included. Dependent variable: $y_{2018Sep} - y_{2014Sep}$.

	Change in page log. length ($y_{2018Sep} - y_{2014Sep}$)	
	Dutch in control gr.	Intention to Treat
	(1)	(2)
Treatment group	0.021	0.004
	(0.043)	(0.036)
Language FE	Yes	Yes
City FE	Yes	Yes
Mean dep. var.	0.156	0.156
SD dep. var.	0.319	0.319
Adj. R-squared	0.248	0.247
Observations	240	240

Notes: Each column presents estimates from a separate cross-section regression of 240 Wikipedia pages. The dependent variable is the change in logarithm of page length. All regressions include language fixed effects and city fixed effects. In column 1, all Dutch pages are assigned to the control group. Column 2 presents the intention-to-treat estimate. Standard errors are reported in parentheses.

Electronic copy available at: https://ssrn.com/abstract=3341630

Online Appendix Table B.4: The effect of treatment on subsequent growth in page length and quality, heterogeneity by page quality. Dependent variable: $y_{2018Sep} - y_{2014Sep}$.

	Change in page length or quality ($y_{2018Sep} - y_{2014Sep}$)		
	ΔLog. page length	ΔQuality rating	ΔSimilarity to Spanish
	(1)	(2)	(3)
	Panel A: Heterogeneity by page quality		
Treatment group	-0.034	-1.564	-0.901
	(0.073)	(2.743)	(0.975)
Treatment group	0.118	2.340	0.612
× Below median	(0.111)	(4.129)	(1.467)
Below median	0.079	3.084	0.260
	(0.084)	(3.125)	(1.111)
Language FE	Yes	Yes	Yes
City FE	Yes	Yes	Yes
Mean dep. var.	0.190	6.589	2.032
SD dep. var.	0.353	11.958	4.409
Adj. R-squared	0.282	0.128	0.189
Observations	180	180	180
	Panel B: Heterogeneity by page completeness		
Treatment group	-0.002	-1.310	-0.768
	(0.072)	(2.732)	(0.969)
Treatment group	0.052	1.788	0.342
× Below median	(0.110)	(4.131)	(1.465)
Below median	0.144*	3.153	0.557
	(0.082)	(3.094)	(1.097)
Language FE	Yes	Yes	Yes
City FE	Yes	Yes	Yes
Mean dep. var.	0.190	6.589	2.032
SD dep. var.	0.353	11.958	4.409
Adj. R-squared	0.294	0.125	0.191
Observations	180	180	180

Notes: Each column presents estimates from a separate cross-section regression of 180 Wikipedia pages. The dependent variable is the change in logarithm of page length (column 1), change in the overall quality rating (column 2), and the change in the similarity to the corresponding Spanish Wikipedia article (column 3). All regressions include language fixed effects and city fixed effects. An indicator for treatment group is interacted with an indicator whether either page pre-treatment overall quality (panel A) or completeness (panel B) is below the median. Standard errors are reported in parentheses.

Electronic copy available at: https://ssrn.com/abstract=3341630

Online Appendix Table B.5: The effect of treatment on subsequent growth in page length and quality, heterogeneity by page length and page age. Dependent variable: $y_{2018Sep} - y_{2014Sep}$.

	Change in page length or quality ($y_{2018Sep} - y_{2014Sep}$)		
	ΔLog. page length	ΔQuality rating	ΔSimilarity to Spanish
	(1)	(2)	(3)
	Panel A: Heterogeneity by page relative length (to Spanish)		
Treatment group	0.043	-0.608	-1.157
	(0.070)	(2.599)	(0.914)
Treatment group	0.000	1.557	1.427
× Below median	(0.099)	(3.655)	(1.285)
Below median	0.221**	6.928*	1.644
	(0.099)	(3.680)	(1.294)
Language FE	Yes	Yes	Yes
City FE	Yes	Yes	Yes
Mean dep. var.	0.190	6.589	2.032
SD dep. var.	0.353	11.958	4.409
Adj. R-squared	0.289	0.148	0.225
Observations	180	180	180
	Panel B: Heterogeneity by page age		
Treatment group	0.053	0.676	-0.158
	(0.062)	(2.302)	(0.807)
Treatment group	-0.084	-3.234	-1.328
× Below median	(0.121)	(4.459)	(1.564)
Below median	0.038	1.011	0.458
	(0.103)	(3.816)	(1.339)
Language FE	Yes	Yes	Yes
City FE	Yes	Yes	Yes
Mean dep. var.	0.190	6.589	2.032
SD dep. var.	0.353	11.958	4.409
Adj. R-squared	0.250	0.105	0.190
Observations	180	180	180

Notes: Each column presents estimates from a separate cross-section regression of 180 Wikipedia pages. The dependent variable is the change in logarithm of page length (column 1), change in the overall quality rating (column 2), and the change in the similarity to the corresponding Spanish Wikipedia article (column 3). All regressions include language fixed effects and city fixed effects. An indicator for treatment group is interacted with an indicator whether either page pre-treatment length (panel A) or age (panel B) is below the median. Standard errors are reported in parentheses.

Electronic copy available at: https://ssrn.com/abstract=3341630

Online Appendix Table B.6: The effect of treatment on subsequent growth in page length and quality, p-values adjusted for multiple hypothesis testing

Outcome variable	Coef.	SE	Unadj. p-value	Adj. p-value
	(1)	(2)	(3)	(4)
Δ log. length	0.026	0.048	0.586	0.974
Δ quality	-0.375	1.777	0.833	0.974
Δ similarity	-0.590	0.624	0.346	0.952
Δ complete	-0.967	2.254	0.669	0.974
Δ interesting	-0.150	1.851	0.936	0.974
Δ illustrated	-2.037	3.752	0.588	0.974
Δ well-written	4.658	5.980	0.438	0.970

Notes: Westfall and Young (1993) multiple hypothesis p-value adjustment as implemented by Jones et al. (2019) employing 10,000 bootstrap draws.

Online Appendix Table B.7: Short-term effects of treatment on subsequent editing activity, p-values adjusted for multiple hypothesis testing

Outcome variable	Coef.	St.er.	Unadj. p-value	Adj. p-value
	(1)	(2)	(3)	(4)
# users: year 1	0.122	0.030	0.000	0.310
# users: year 2	0.119	0.035	0.001	0.385
# edits: year 1	0.138	0.032	0.000	0.294
# edits: year 2	0.140	0.043	0.002	0.406
# edits excl. treatment: year 1	0.022	0.030	0.462	0.873
# edits excl. treatment: year 2	0.072	0.043	0.099	0.668
Edit distance: year 1	12.806	31.301	0.683	0.873
Edit distance: year 2	100.284	105.570	0.344	0.873
Capped edit distance: year 1	17.282	5.609	0.003	0.420
Capped edit distance: year 2	20.512	8.654	0.019	0.511

Notes: Westfall and Young (1993) multiple hypothesis p-value adjustment as implemented by Jones et al. (2019) employing 10,000 bootstrap draws.

Electronic copy available at: https://ssrn.com/abstract=3341630

Online Appendix Table B.8: Long-term effects of treatment on subsequent editing activity, p-values adjusted for multiple hypothesis testing

Outcome variable	Coef.	St.er.	Unadj. p-value	Adj. p-value
	(1)	(2)	(3)	(4)
# users: year 3	0.014	0.028	0.621	0.983
# users: year 4	0.024	0.033	0.480	0.982
# edits: year 3	0.011	0.032	0.726	0.988
# edits: year 4	0.029	0.036	0.424	0.979
# edits excl. treatment: year 3	-0.035	0.031	0.269	0.910
# edits excl. treatment: year 4	-0.026	0.033	0.443	0.982
Edit distance: year 3	-13.534	21.692	0.534	0.983
Edit distance: year 4	1.296	79.265	0.987	0.996
Capped edit distance: year 3	-4.158	5.294	0.434	0.982
Capped edit distance: year 4	0.485	5.343	0.928	0.996

Notes: Westfall and Young (1993) multiple hypothesis p-value adjustment as implemented by Jones et al. (2019) employing 10,000 bootstrap draws.

Electronic copy available at: https://ssrn.com/abstract=3341630