Online Appendix for "Externalities in Knowledge Production: Evidence from a Randomized Field Experiment" at Experimental Economics

Marit Hinnosaar* Toomas Hinnosaar† Michael E. Kummer‡
Olga Slivko§

August 21, 2021

A 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:

$$\Delta logLength_i = \beta_0 + \beta_1 TreatmentGroup_i + LanguageFE_i + \varepsilon_i \quad (1)$$

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

*University of Nottingham, Collegio Carlo Alberto, and CEPR, marit.hinnosaar@gmail.com.
†University of Nottingham and CEPR, toomas@hinnosaar.net.
‡University of East Anglia, Georgia Institute of Technology, and ZEW, M.Kummer@uea.ac.uk.
§Rotterdam School of Management, Erasmus University, slivko@rsm.nl.

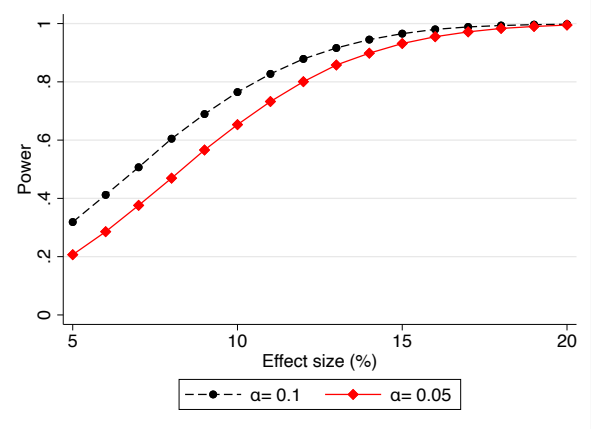
effects:

$$Users_{2014,i} = \beta_0 + \beta_1 TreatmentGroup_i + \beta_2 Users_{2011,i} + LanguageFE_i + \varepsilon_i \quad (2)$$

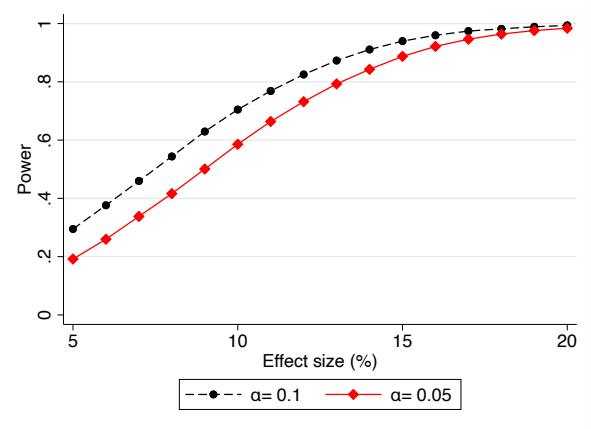
The outcome variable $Users_{2014,i}$ is the yearly average of the number of monthly users from August 2013 to July 2014, and we calculate $Users_{2011,i}$ analogously. Note that since this cross-section regression includes the lagged outcome variable $Users_{2011,i}$, we expect the estimates to be similar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

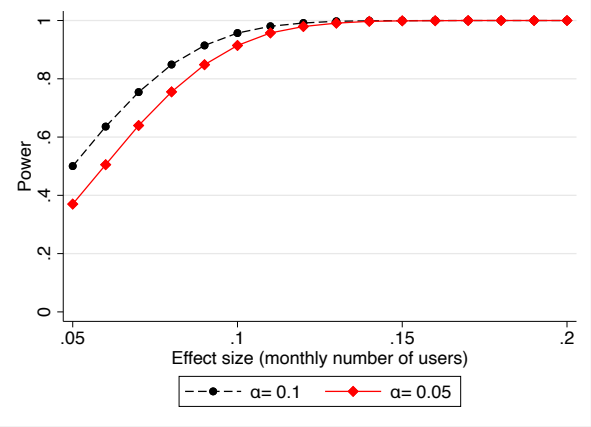
To summarize, our study is underpowered to detect small long-term effects on page length, but we can detect even half of the effect-size suggested by Aaltonen and Seiler (2016). On the number of users, our experiment has relatively more power, being able to detect the effect sizes suggested in the literature.



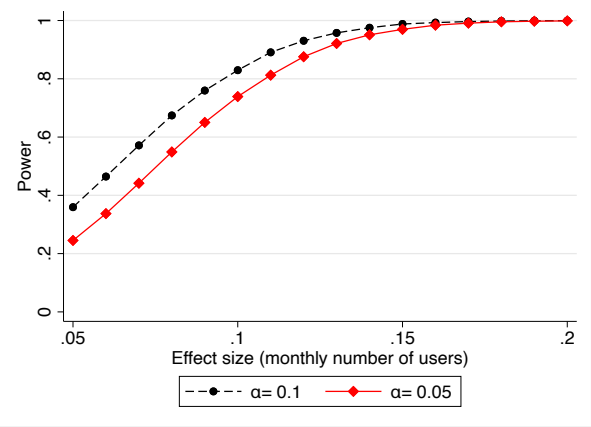
(a) Page length (with Dutch)



(b) Page length (without Dutch)



(c) Number of users (with Dutch)

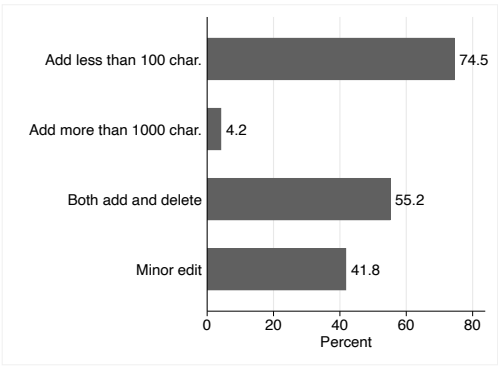


(d) Number of users (without Dutch)

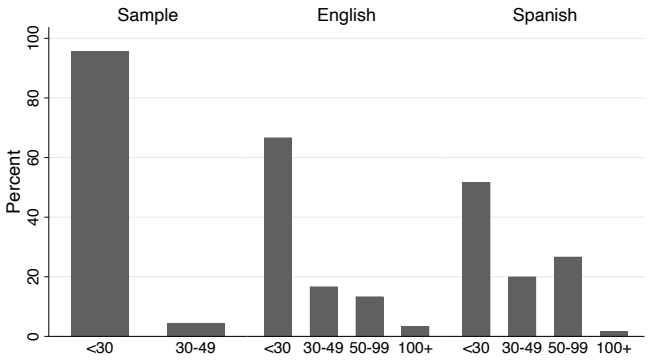
Figure A.1: Power analysis for the effect on the page length and the number of users

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

B Additional figures and tables



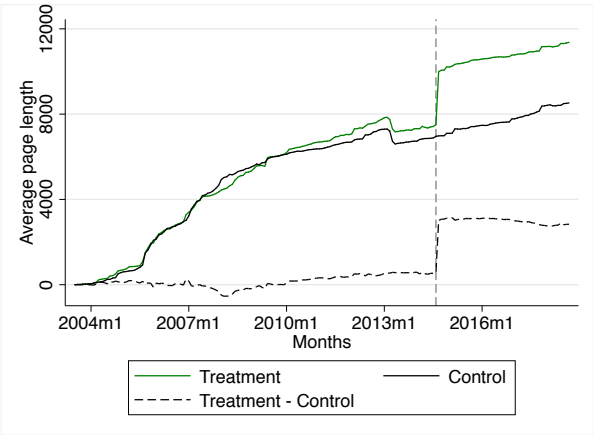
(a) Edits by type (as a percentage)



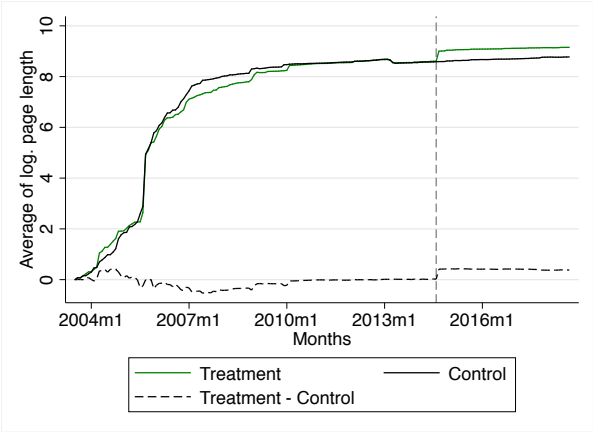
(b) Distribution of pages by the number of watchers

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.



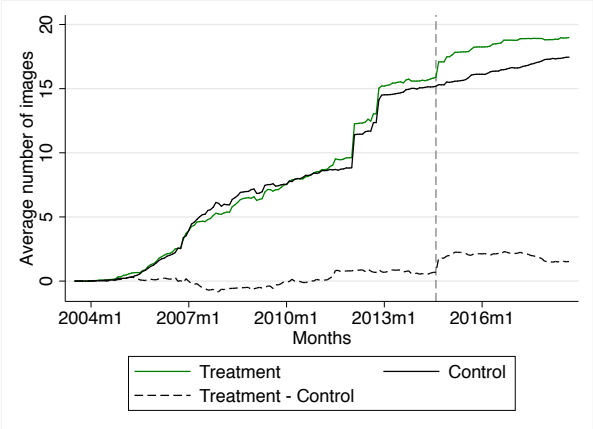
(a) Page length without Cordoba in French



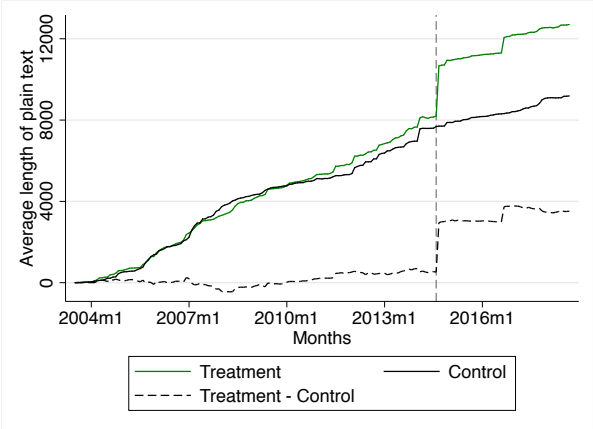
(b) Logarithm of page length

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.



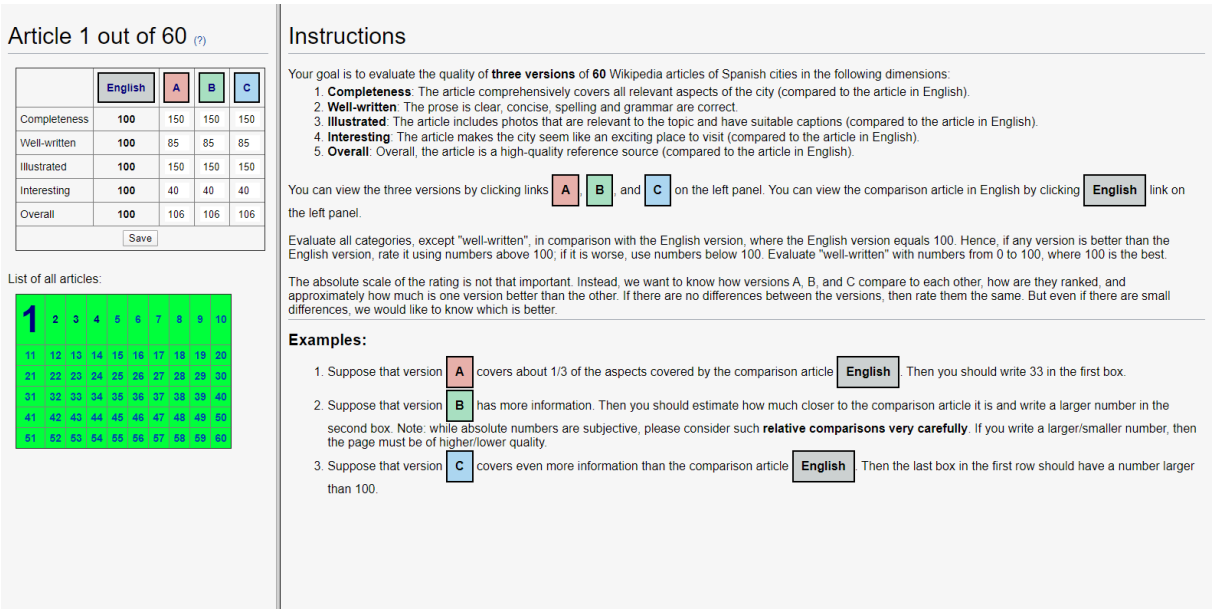
(a) Images



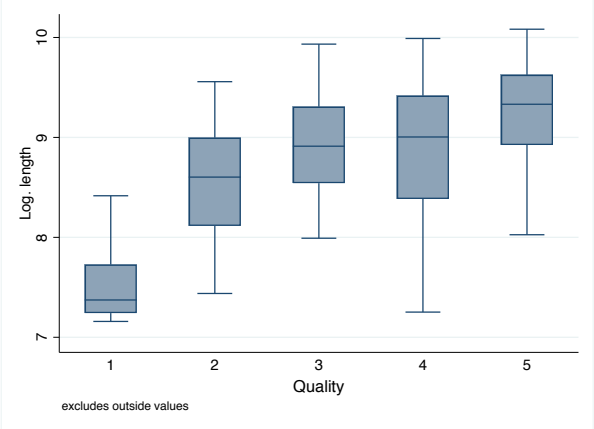
(b) Plain text

Online Appendix Figure B.3: Other output measures

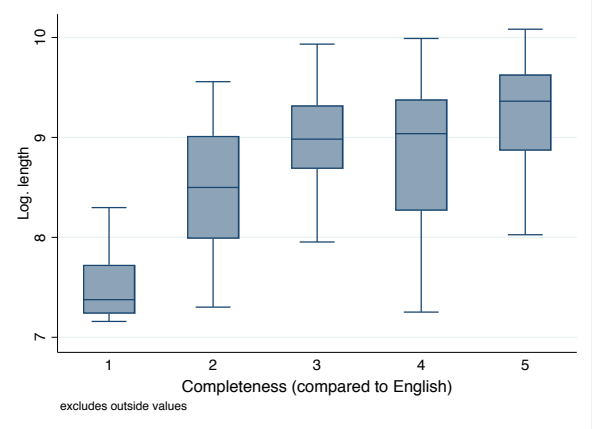
Notes: The number of observations is 90 in the control and 90 in the treatment groups. The experiment month (August 2014) is marked by dashed vertical line. *Plain text* is obtained by removing html elements from the parsed text.



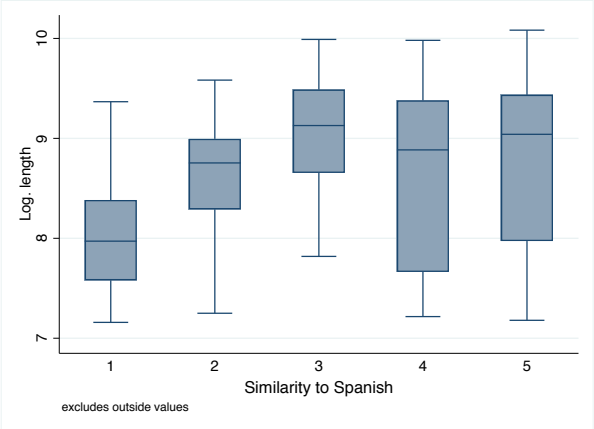
Online Appendix Figure B.4: A screenshot with instructions given to research assistants who rated the quality of articles as described in section 3.2



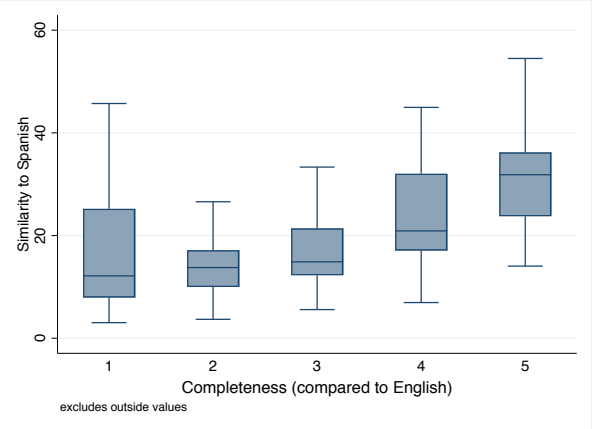
(a) Quality versus log. length



(b) Completeness versus log. length



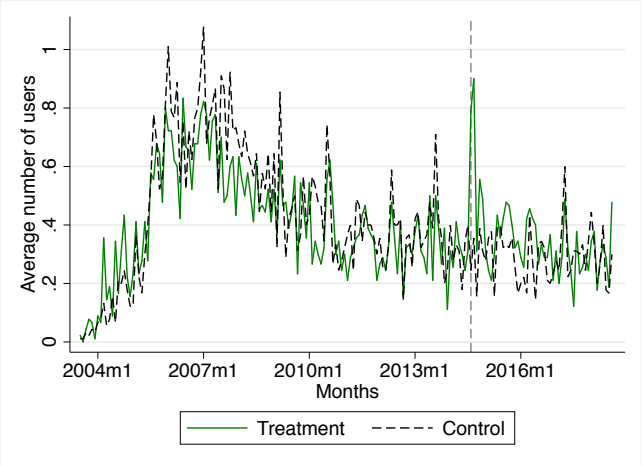
(c) Similarity to Spanish versus log. length



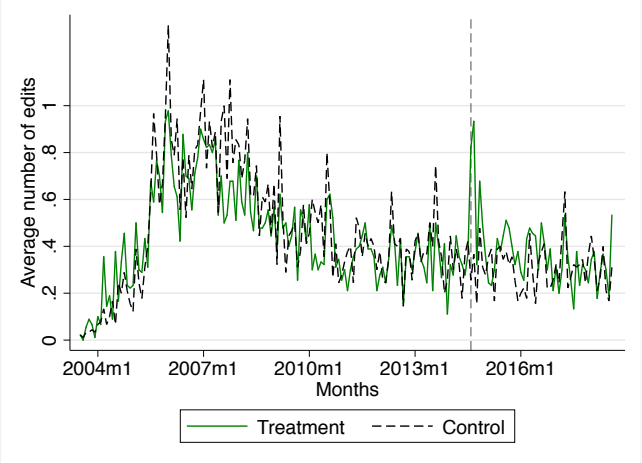
(d) Completeness versus similarity to Spanish

Online Appendix Figure B.5: Quality, completeness (compared to English), similarity (compared to Spanish), and log. length

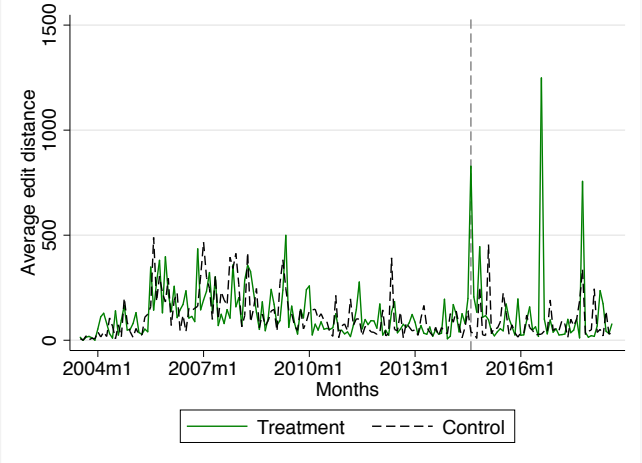
Notes: We group the characteristic on the horizontal axes into quintiles. For each quantile, the graph presents the median (as a horizontal line) and the interval from the 25th to the 75th percentile (as the box) of the variable on the vertical axes. All measures are from pre-treatment (in August 2014).



(a) Number of users



(b) Number of edits



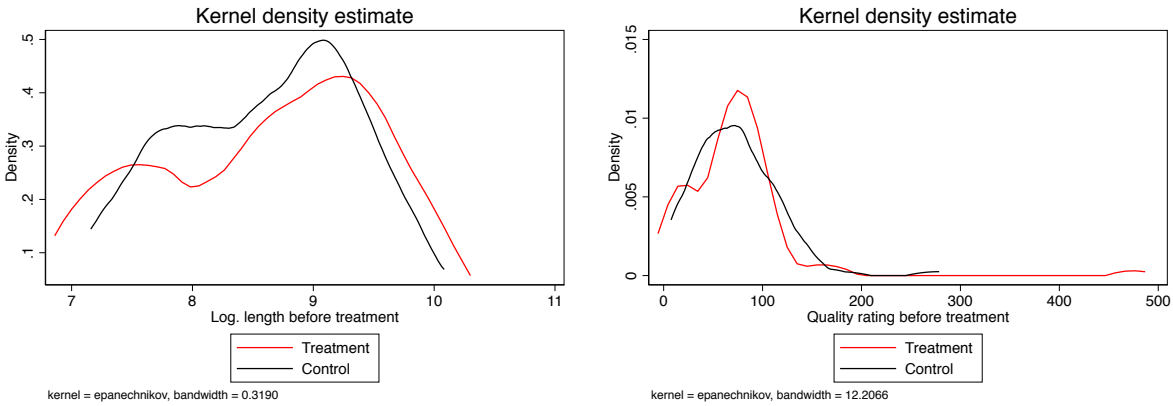
(c) Edit distance

Online Appendix Figure B.6: 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.

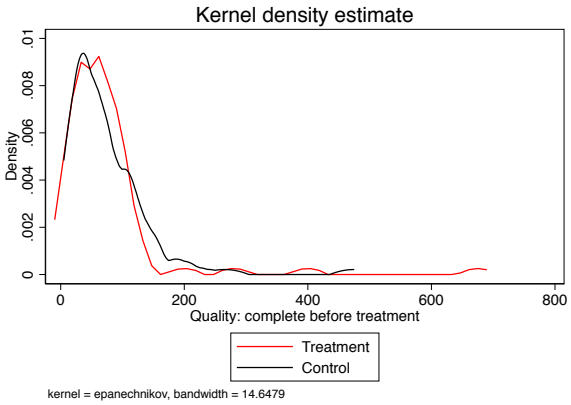
'ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there', 'about', 'once', 'during', 'out', 'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some', 'for', 'do', 'its', 'yours', 'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's', 'am', 'or', 'who', 'as', 'from', 'him', 'each', '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.7: List of stop words used to clean the text for computing the Tversky similarity measure in section 3.4

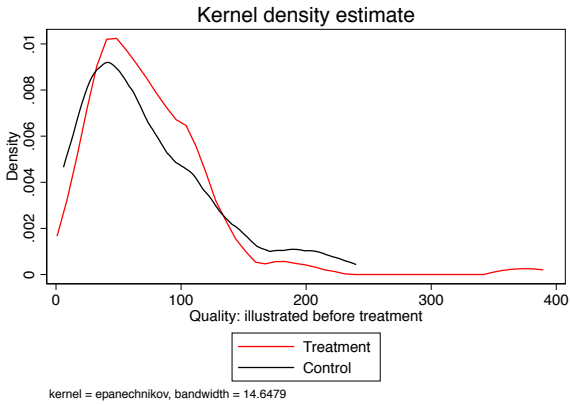


(a) Log. length

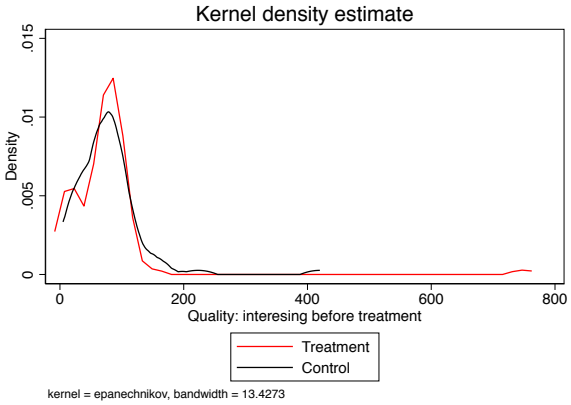
(b) Overall quality



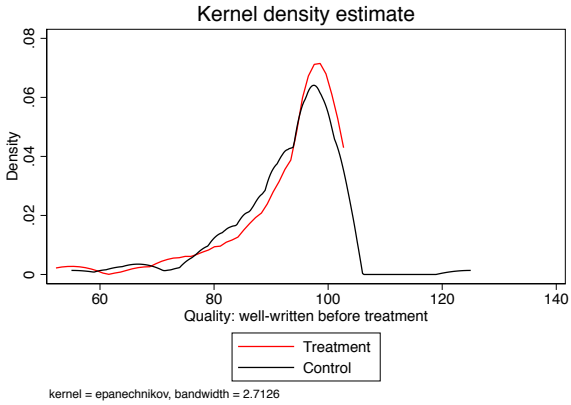
(c) Completeness



(d) Illustrated



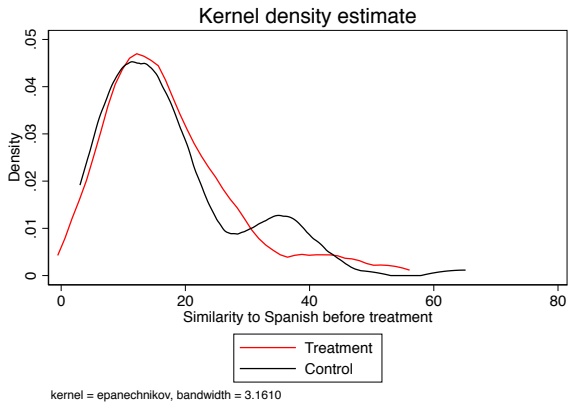
(e) Interesting



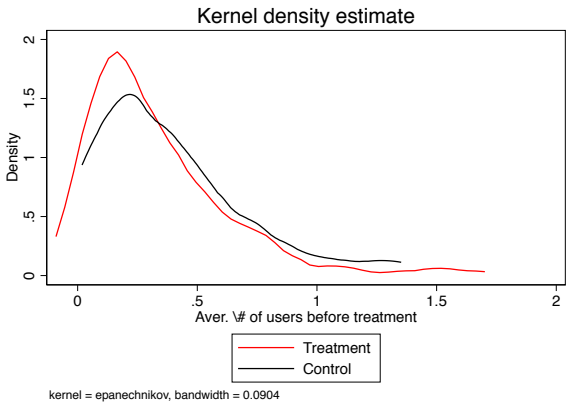
(f) Well-written

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

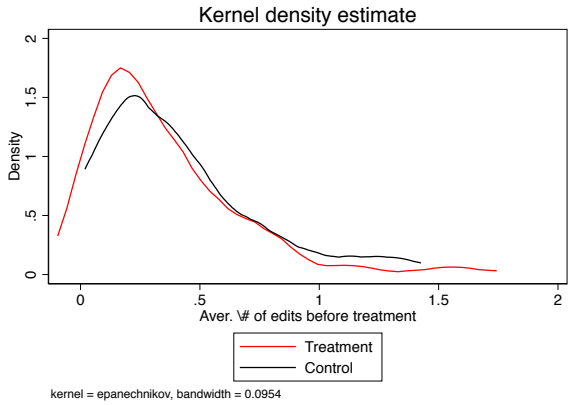
Notes: Kernel density estimates of the pre-treatment distributions, separately for the control and treatment groups.



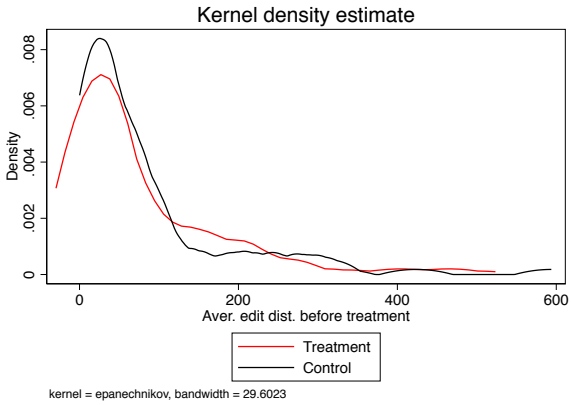
(a) Similarity to Spanish



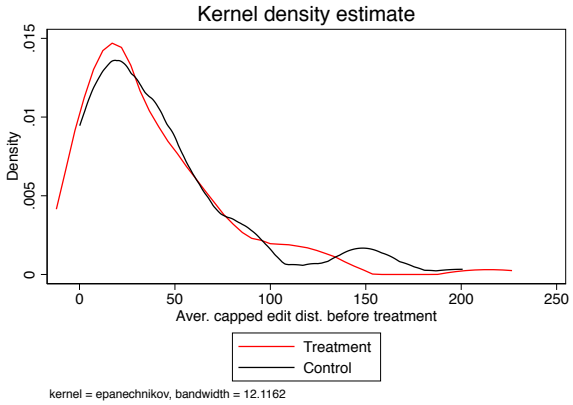
(b) Average number of users



(c) Average number of edits



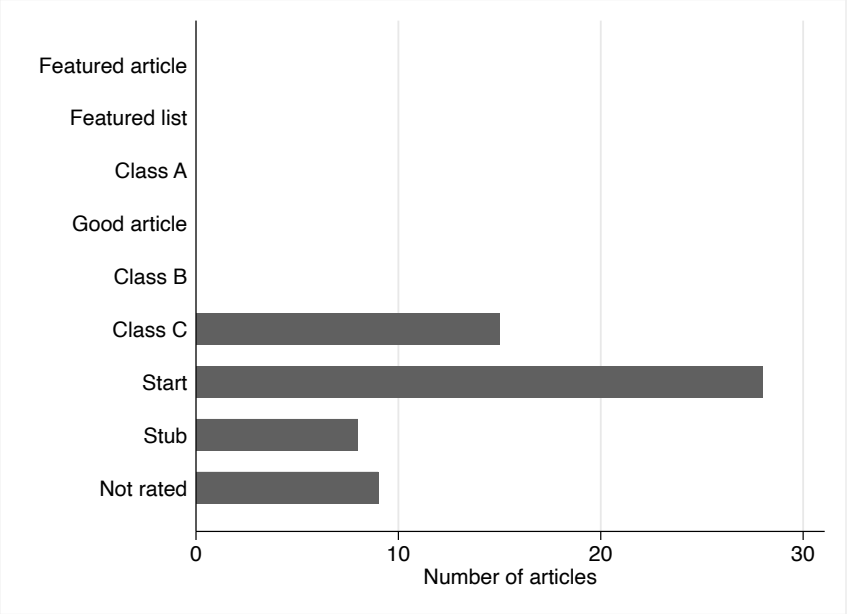
(d) Average edit distance



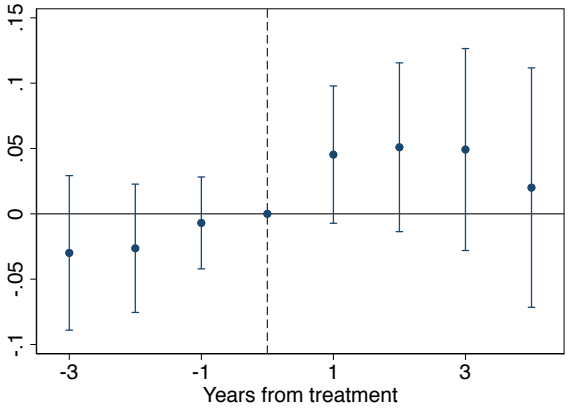
(e) Average capped edit distance

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

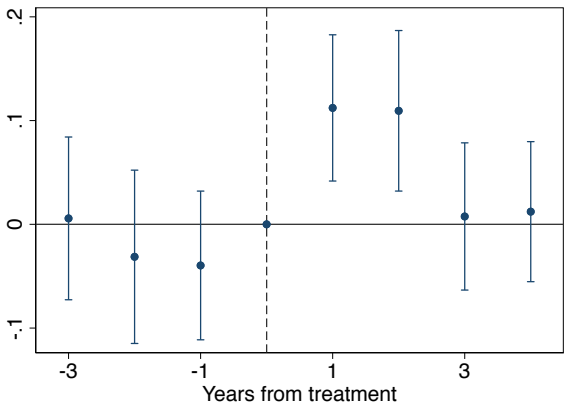
Notes: Kernel density estimates of the pre-treatment distributions, separately for the control and treatment groups.



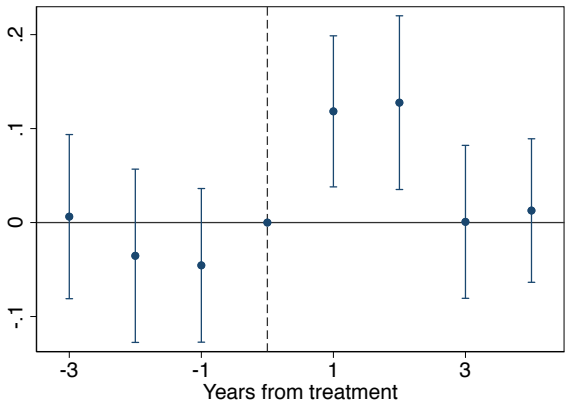
Online Appendix Figure B.10: Quality of the 60 city pages (corresponding to our sample) in the English Wikipedia



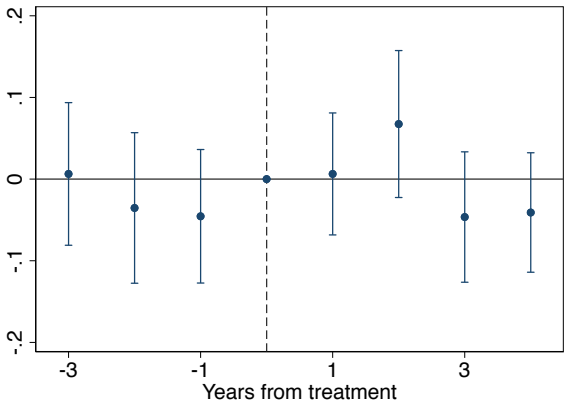
(a) Log. length (minus treatment text)



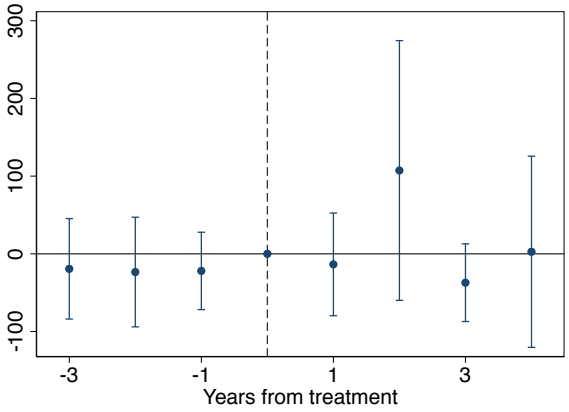
(b) # Users



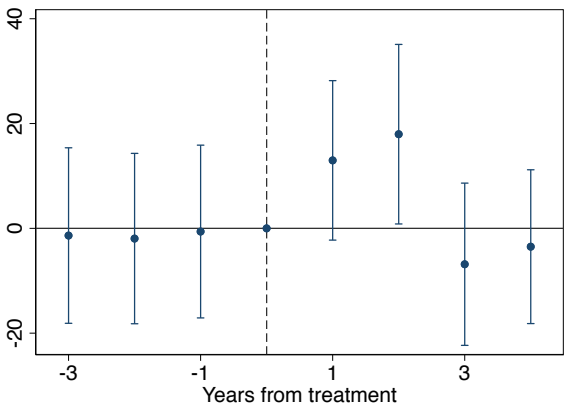
(c) # Edits



(d) # Edits excluding treatment



(e) Edit distance



(f) Capped edit distance

Online Appendix Figure B.11: 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.

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)

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.

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.

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.

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.

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.

C Theoretical framework

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.[1] Agent t gets payoff

$$u_t(x_t, X_{t-1}) = v_t(X_t) + w_t(x_t) - c_t(x_t, X_{t-1}), \tag{3}$$

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.[2]

Let us first consider a benchmark with no externalities, i.e., $\mu(X_{t-1}) = 0$. Then the optimal contribution of agent t that maximizes (3) 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 4a.

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

$$\frac{\partial c_t}{\partial x_t} = \gamma_t \left(x_t - \mu(X_{t-1})\right). \tag{4}$$

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

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

[2] 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.

find something to contribute.

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

$$x_t^*(X_{t-1}) = \max\left\{0, \frac{\alpha_t + \beta_t}{\gamma_t} + \mu(X_{t-1})\right\} \tag{5}$$

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 4b 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 4c 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 4d illustrates this case.

References

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

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

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

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.

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