Readme File

Data and code

Externalities in Knowledge Production: Evidence from a Randomized Field Experiment

Marit Hinnosaar, Toomas Hinnosaar, Michael E. Kummer, Olga Slivko

A: Raw and processed data files

We have three main sources of data. The first data source is Wikipedia editing histories. Specifically, Wikipedia German, French, Italian, Spanish, English, and Dutch editions. The data is publicly available to download. We translated some of the Wikipedia data to English using Yandex Translate (https://translate.yandex.com/). The translated data is also included in the replication files. Second, we use data from Hinnosaar, Hinnosaar, Kummer, Slivko (2021). Third, we obtained Wikipedia page quality ratings for German, French, and Italian pages by two independent raters fluent in the corresponding language. The lists of all the data files are in tables 1 and 2.

In the php/input folder, each file includes three variables: Wikidata ID which identifies the Wikipedia page, it is the same id in each language (*page*); *date*; English text of the Wikipedia pages, translated to English if the page is from the Wikipedia edition other than English (*plaintext*).

The file stata_analysis/input/editdistance_character_minoredits_numericuserid.csv includes eight variables: *date*, Wikidata ID which identifies the Wikipedia page, it is the same id in each language (*page*); the language of Wikipedia edition (*lang*); the type of Wikipedia user editing the page (*usertype*); number of characters the user inserted (*inserted*) and number of characters the user deleted (*deleted*); an indicator whether the edit was marked as minor (minor); user ID (*numericuserid*).

The file stata_analysis/input/pagelength_alllanguages.csv includes four variables: *date*; Wikidata ID which identifies the Wikipedia page, it is the same id in each language (*page*); the language of Wikipedia edition (*lang*); Wikipedia page length (*length*).

The file stata_analysis/input/pagelength_images.csv includes six variables: *date*; Wikidata ID which identifies the Wikipedia page, it is the same id in each language (*page*); the language of Wikipedia edition (*lang*);

Wikipedia page length (*length*); the length of plain text of the Wikipedia page; which is, html elements removed from the text (*plain*); number of images on the Wikipedia page (*images*).

The file stata_analysis/input/quality_wikipedia_offical_english_ratings.csv includes five variables: Wikidata ID which identifies the Wikipedia page, it is the same id in each language (*page*); page quality rating in English language Wikipedia in the project Spain (*wp_spain_class*); page quality rating in English language Wikipedia in the project Cities (*wp_cities_class*); the importance of the page in English language Wikipedia in the project Spain (*wp_spain_importance*); the importance of the page in English language Wikipedia in the project Cities (*wp_cities_importance*).

The file stata_analysis/input/randomization group.csv includes three variables: Wikidata ID which identifies the Wikipedia page/city, it is the same id in each language (*page*); number of tourists in the city (*vNumberOfTourists*), the number of the stratification group to which the city belonged in the experiment (*vGroup*).

The file stata_analysis/input/randomization treatment.csv includes three variables: Wikidata ID which identifies the Wikipedia page/city, it is the same id in each language (*page*); the language of Wikipedia edition (*lang*); an indicator for whether the page-language pair was randomized to the treatment group or not (*treatment*).

The file stata_analysis/input/tversky_en.csv includes five variables: Wikidata ID which identifies the Wikipedia page/city, it is the same id in each language (*page*); the language of Wikipedia edition (*lang*); *year*, *month*, Tversky similarity index measuring similarity between the page in the language and the same page in English (*similarity*).

The file stata_analysis/input/tversky_es.csv includes five variables: Wikidata ID which identifies the Wikipedia page/city, it is the same id in each language (*page*); the language of Wikipedia edition (*lang*); *year*, *month*, Tversky similarity index measuring similarity between the page in the language and the same page in Spanish (*similarity*).

The file stata_analysis/input/users_edits.csv includes nine variables: month of Wikipedia editing activity (*date*); Wikidata ID which identifies the Wikipedia page/city, it is the same id in each language (*page*); the language of Wikipedia edition (*lang*); number of registered Wikipedia users editing the page in that language during the month (*user_edit_users*); number of anonymous users editing it (*anonymous_edit_users*); number of regular user days editing it (*user_edit_days*), number of anonymous user edit days (*anonymous_edit_days*); number of regular user edit days to text that was not added by the treatment (*user_edit_days_nontt*); number of anonymous user edit days to text that was not added by the treatment (*anonymous_edit_days_nontt*).

The file stata_analysis/input/watchers.csv includes seven variables: Wikidata ID which identifies the Wikipedia page/city, it is the same id in each language (*page*); number of watchers, that is, the number of Wikipedia editors that get notified of all the edits of page in the English language edition (watchers_en), and similarly in the French (watchers_fr), German (watchers_de), Italian (watchers_it), Dutch (watchers_nl), and Spanish (watchers_es) language editions.

The files in folder stata_analysis/input/quality_manualratings include the following variables: order (#); Wikidata ID which identifies the Wikipedia page/city, it is the same id in each language (id); the language of Wikipedia edition (*lang*); rater ID (*rater*); rating of completeness before treatment in August 1st, 2014 (*Completeness_1*), after treatment in September 1st, 2014 (*Completeness_2*), and four years after treatment in September 1st, 2018 (*Completeness_3*); similarly three ratings for well-written (*Well-written*), illustrated (*Illustrated*), interesting (*Interesting*), overall quality (*Overall*).

Table 1: Dataset list in folder stata_analysis/input

Data file	Source
stata_analysis/input/editdistance_character_minoredits_numericuserid.csv	Wikipedia
stata_analysis/input/pagelength_alllanguages.csv	Wikipedia
stata_analysis/input/pagelength_images.csv	Wikipedia
stata_analysis/input/quality_wikipedia_offical_english_ratings.csv	Wikipedia
stata_analysis/input/randomization_group.csv	Hinnosaar, Hinnosaar, Kummer, Slivko (2021)
stata_analysis/input/randomization_treatment.csv	Hinnosaar, Hinnosaar, Kummer, Slivko (2021)
stata_analysis/input/tversky_en.csv	Wikipedia, processed data
stata_analysis/input/tversky_es.csv	Wikipedia, processed data
stata_analysis/input/users_edits.csv	Wikipedia
stata_analysis/input/watchers.csv	Wikipedia
quality_manualratings/save_69ee8c.csv	Ratings collected in this study
quality_manualratings/save_903c9a.csv	Ratings collected in this study
quality_manualratings/save_ad000b.csv	Ratings collected in this study
quality_manualratings/save_1af85e.csv	Ratings collected in this study
quality_manualratings/save_17d3b2.csv	Ratings collected in this study
quality_manualratings/save_400b4c.csv	Ratings collected in this study

Table 2: Dataset list in folder php/input

Data file	Source
php/input/articles_en_y2018.csv	Wikipedia
php/input/after_translation/articles_translated_de_y2014.csv	Wikipedia, translated using Yandex Translate
php/input/after_translation/articles_translated_de_y2018.csv	Wikipedia, translated using Yandex Translate
php/input/after_translation/articles_translated_es_y2018.csv	Wikipedia, translated using Yandex Translate
php/input/after_translation/articles_translated_fr_y2014.csv	Wikipedia, translated using Yandex Translate
php/input/after_translation/articles_translated_fr_y2018.csv	Wikipedia, translated using Yandex Translate
php/input/after_translation/articles_translated_it_y2014.csv	Wikipedia, translated using Yandex Translate
php/input/after_translation/articles_translated_it_y2018.csv	Wikipedia, translated using Yandex Translate

B: Programs used to perform the data analysis

B.1: Computational requirements

- Stata (version 16)
- Stata user-written programs that can be installed from the Boston College Statistical Software Components (SSC) archive:
    - wyoung (version 1.0.2, 14 May 2018) by Julian Reif
    - estout (version 3.21, 19 August 2016) by Ben Jann
    - coefplot (version 1.8.4, 17 December 2020) by Ben Jann
- PHP (version 7.3.21)

B.2: Description of programs

The programs are in two folders: php/code and stata_analysis/code.

In the folder php/code, the file index.php calculates the Tversky index comparing the translated text from Wikipedia pages to that in either English or translated Spanish texts. It uses input files from php/input folder and additional functions in class.stemmer.inc. It generates two output files: tversky_en.csv and tversky_es.csv saving those into the php/output folder. These two files will be used as input in the remaining analysis.

In the folder stata_analysis/code, the files 01_get_data.do and 02_get_data_quality.do prepare the datasets for the analysis. Using the data generated by the two files, the files 03_figures_length_edits.do, 04_regressions_cross_section.do, 05_regressions_panel.do, 06_other_results.do, 07_multiple_hypothesis_testing.do generate descriptive statistics figures and tables, estimate the regressions and generate output tables and figures. The file 08_power_analysis_run.do runs the power analysis using power_analysis_montecarlo.do. Finally, 09_power_analysis_graphs.do generates power analysis figures. The file run_directory.do executes all the above-described Stata codes and generates the output tables and figures saving those to folder stata_analysis/output. However, each of the numbered files could also be run separately in the given numbered order.

B.3: Instructions

Step 1: Unzip the file with the replication files. There will be two main folders php and stata_analysis.

Step 2: Copy php directory to webserver with php 7.0 or higher. Open index.php?base=en and index.php?base=es to generate the two output files tversky_en.csv and tversky_es.csv.

Step 3: Set the stata_analysis/code folder as the working directory. Run run_directory.do which runs all the remaining do-files (01-09) in the folder. It saves temporary files into the stata_analysis/temp folder and all the output tables and figures into the stata_analysis/output folder. See tables 3 and 4 for the complete list of all the tables and figures and corresponding do-files.

B.4: List of tables and figures and the corresponding programs

The full list of tables, figures, and programs are reported in Tables 3 and 4.

Table 3: List of tables and figures in the main text and the corresponding programs

Figure/table	Output file	Program
Table 1	tBalanceTest.tex	04_regressions_cross_section.do
Table 2	tCompleteness.tex	04_regressions_cross_section.do
Table 3	tRegrLengthQuality.tex	04_regressions_cross_section.do
Table 4	tRegrManualRatingsDetailed.tex	04_regressions_cross_section.do
Table 5	tRegrPanel.tex	05_regressions_panel.do
Figure 1	gLength.pdf; 03_figures_length_edits	03_figures_length_edits.do
Figures 2a-2d	gLength_fr.pdf, gLength_de.pdf, gLength_it.pdf, gLength_nl.pdf	03_figures_length_edits.do
Figures 3a-3f	gChange_Quality_overall.pdf, gChange_Quality_completeness.pdf, gChange_Quality_illustrated.pdf, gChange_Quality_interesting.pdf, gChange_Quality_wellwritten.pdf, gChange_Similarity.pdf	04_regressions_cross_section.do
Figures 4a-4d	out_con.pdf, out_lin.pdf, out_lind.pdf, out_pol3.pdf	External source (theoretical model)

Table 4: List of tables and figures in the online appendix and the corresponding programs

Figure/table	Output file	Program
Table C1	tEurobarometer.tex	External source (Eurobarometer)
Table C2	tRobustRegrLengthQuality_Controls.tex	04_regressions_cross_section.do
Table C3	tRobustRegrLengthQuality_Dutch.tex	04_regressions_cross_section.do
Table C4	tRobustLengthQuality_HeterQuality.tex	04_regressions_cross_section.do
Table C5	tRobustLengthQuality_HeterLengthAge.tex	04_regressions_cross_section.do
Table C6	tMultipleHypothesisTesting_LengthQuality.tex	07_multiple_hypothesis_testing.do
Table C7	tMultipleHypothesisTesting_ShortTermEditing.tex	07_multiple_hypothesis_testing.do
Table C8	tMultipleHypothesisTesting_LongTermEditing.tex	07_multiple_hypothesis_testing.do
Figure A1a-A1d	gPower_Length_240.pdf, gPower_Length_180.pdf, gPower_Users_240.pdf, gPower_Users_180.pdf	09_power_analysis_graphs.do
Figure C1a-C1b	gEditTypes.pdf, gWatchers.pdf	06_other_results.do
Figure C2a-C2b	gLength_wo_FrenchCordoba.pdf, gLogLength.pdf	03_figures_length_edits.do
Figure C3a-C3b	gImages.pdf, gLengthPlainText.pdf	03_figures_length_edits.do
Figure C4	rating_instructions.png	External source (screenshot)
Figure C5a-C5d	gLength_overall.pdf, gLength_completeness.pdf, gLength_similarity.pdf, gSimilarity_completeness.pdf	03_figures_length_edits.do
Figure C6a-C6c	gUsers.pdf, gEditDays.pdf, gEditDist.pdf	04_regressions_cross_section.do
Figure C7		External source (text)
Figure C8a-C8f	gBC_vLogLength2014Aug.pdf, gBC_overall_1.pdf, gBC_completeness_1.pdf, gBC_illustrated_1.pdf, gBC_interesting_1.pdf, gBC_wellwritten_1.pdf	04_regressions_cross_section.do
Figure C9a-C9e	gBC_vSimilarity_2014Aug.pdf, gBC_vAverUsersPreTreatment.pdf, gBC_vAverEditDaysPreTreatment.pdf, gBC_vAverEditDistPreTreatment.pdf, gBC_vAverEditDistDCapPreTreatment.pdf	04_regressions_cross_section.do
Figure C10	gEnglishQualityOfficial.pdf	06_other_results.do
Figure C11a-C11e	gRegrPanel_vFLogLengthMinusTreatment.pdf, gRegrPanel_vUsers.pdf, gRegrPanel_vEditDays.pdf, gRegrPanel_vEditsNotOurText.pdf, gRegrPanel_vEditDist.pdf, gRegrPanel_vEditDistDailyCap.pdf	05_regressions_panel.do