# Wikipedia as the Great Equalizer

Kris Gulati*

April 26, 2022

*Department of Economics, University of California at Merced. Email:kgulati@ucmerced.edu

## 1 Introduction

*"The advancement and diffusion of knowledge is the only guardian of true liberty."*- James Madison

An objective view of recent history highlights the unprecedented growth in the acces sibility to knowledge, replacing historical barriers as well as reducing the influence of the gatekeepers of information. Gradually information is permeating the world, in an ever important open-source environment. Wikipedia is the largest encyclopedia and with that comes its imperative role in the diffusion of knowledge. As frequently dis cussed within the overarching goals are implementing the broader aims of Wikipedia by providing access to disadvantaged groups. The open-source nature of Wikipedia enables the platform to provide information to a more representative global audience, bypassing pay-walls that prohibit many individuals from accessing information and the novel creation of knowledge. This research proposal aims to see if these obstacles can be tackled, using a rigorous randomized control trial.

As we know, several disparities remain in the progression of science. For example, several scientific groups suffer from inequalities: women, non-English speakers, and researchers from developing countries. These individuals are likely less to be repre sented in academia, despite their contributions to science (Exley & Kessler, 2019). For example, Azoulay et al. (forthcoming) find that relative to non-Chinese project investigators(PIs), Chinese PIs receive 32 % fewer citations from US researchers.

Furthermore, scholars in developing countries face additional barriers to the flow of knowledge and are often severely disadvantaged, given that most hubs of knowledge in the world are in the developed world. These limitations hinder the true poten tial of a representative scholarly population, hindering the progress of science. For example (Agarwal & Gaule, 2020), document that equally talented researchers orig inating from developing countries suffer from a large scientific productivity penalty. Our study aims to use Wikipedia to understand and close these gaps.

This project aims to accentuate the role Wikipedia has in contributing to the progress of science by providing opportunities in expanding the coverage of neglected scholars, regardless of historical disparities. In addition, it also provides insights on the role of Wikipedia as the unofficial “handbook” of global knowledge. Additionally, we hope to document how the diffusion of this new information may spur applied science, via increased patenting activity in non-English speaking countries. To the best of our knowledge, we would be the first economics paper to research how Wikipedia could be used as a tool to support disadvantaged researchers. Additionally to the best of our knowledge, we would be the first paper to see how Wikipedia shapes applied science int he form of patenting.

In summary, this project aims to answer two questions via a randomized control trial:

Q1: Does the creation of Wikipedia profiles for disadvantaged groups (female scientists, non-English speaking scientists, and scientists from developing countries) increase their coverage and the number of citations, helping to alleviate some of the hurdles they face?

Q2: Does translating Wikipedia articles to non-English languages increase the levels of innovation (proxied by patent citations) in developing countries?

Additionally, we hope to investigate the following questions with a non-causal empir ical approach

Q3: To what extent are disadvantaged scholars (women, non-English speaking scientists, and researchers from developing countries) currently excluded from Wikipedia?

Q4: Reanalyzing (Thompson & Hanley, 2018), did their randomized control trial have any impact on applied science, measured by patent citations

To execute the first two research questions, this project will focus on randomly creat ing Wikipedia profiles for disadvantaged groups (the treatment arm of the random ized control trial). To do this, we will identify scholars who do not currently have Wikipedia profiles and match them on a number of characteristics, such as number of publications, institution of employment, and citations. These individuals will be carefully matched using *sosia* a software designed to find similar scientists according to a number of carefully selected observables (Rose & Baruffaldi, 2020). After this careful matching process is complete, half of the authors will be assigned a Wikipedia page (i.e. the treatment group) and the remainder serve as a control group. Such a careful randomized control trial, allows for precise causal inference estimates on the impact of Wikipedia on science. If we find a positive effect of treatment, we believe this could spur further usage of Wikipedia as a tool to support disadvantaged scholars.

In addition, we want to understand the effects Wikipedia has in breaking the barriers to knowledge, particularly in developing countries. The vast majority of scholarly topics of Wikipedia are in English, excluding valuable information to many promising but disadvantaged scholars. We want to understand if Wikipedia can serve as a vehicle to break these gaps by first collecting a series of related topics, and then subsequently randomizing which topics get to be translated to a language used in a developing country.

In many developing countries contain entire populations or regions of people only speak native languages, such as Tulugu, Urdu, or Quechua. These languages are not as easily translatable yet contain a large number of speakers. These groups also tend to be marginalized, and perhaps have the most to gain by having scholarly informa tion translated to their own language. We can observe these outcomes by capturing the changes in patent citations related to the topics that were randomly translated.

Why is a randomized control trial necessary for the first two questions? A randomized control trial provides an exogenous shock to the knowledge contained in Wikipedia, allowing for precise estimates of it's effect on disadvantaged scholars and applied sci ence. We will be able to come up with precise estimates of the impact on Wikipedia on science. Furthermore, methodologically a randomized control trial overcomes en dogeneity concerns. If Wikipedia editors write about the articles they are interested in, they are selecting certain for certain characteristics. This would bias econometric specifications.

Although we're confident that the randomized control trial will produce results, we also have short-run questions that can be answered in our research project, i.e. ques tions 3 and 4. To answer question 3, we will see to what extent are current Wikipedia profiles biased against women, non-English speaking scientists, and developing coun try scholars. We will do this by matching scholars along a number of characteristics (as described above) and evaluate if scholars have a presence on Wikipedia. As well as the *quantity* of the articles, i.e. whether a profile exists. We will also measure the *quality* of the articles, i.e. the length and depth of articles.

Finally, question four can be answered by re-analyzing the existing data by (Thompson & Hanley, 2018). Again, this is a short-run research question that can be implemented relatively quickly. Here, we will see if the treatment articles in (Thompson & Hanley, 2018) also result in increased patent citations in the relevant areas compared to the control group. Thus, we would potentially be able to see a direct impact of Wikipedia on applied science.

# 2 Impact on Wikipedia

Another point worth mentioning is tying the framework of this proposal to current Wikimedia projects. This proposal contributes to several ideas that encompass the philosophy of several Wikimedia projects, such as having a research question that deals closely with 'Knowledge Gaps'. The logic being that many people try to identify information on scientists by searching for scientists’ names. Thus, perhaps the creation of these new profiles demonstrates how Wikipedia can shape science.

Additionally, the research project helps to deal with ’Representation Gaps’. Because our methodology is a randomized control trial, we can see whether Wikipedia can be used as a tool to support disadvantaged scholars.

Lastly, this projects also creates a new framework in how Wikimedia can innovate free knowledge, by creating a methodology in which creating access to information in non-English languages can help break the barriers associated with an internal cost to accessing knowledge

## 3 Intellectual Merit

To the best of our knowledge, no prior study has captured the causal impact of Wikipedia to support disadvantaged scholars or applied science. Even though this paper builds upon the framework from (Thompson & Hanley, 2018), this project aims to establish a different argument in which Wikipedia is not just a framework that expands knowledge, but also in its leverage as an equalizer of social disparities in science.

In addition, this will be the first paper that will show the effects of Wikipedia on applied science in developing countries. Whereas (Thompson & Hanley, 2018) focuses on the impact on academic citations, we measure its impact on patents. This has direct practical and financial impact. The increased exposure of knowledge translated into a native language can increase the pool of ideas for inventors, enabling the creation of products that can reduce poverty and increase the standard of living for people in developing countries. For example, agricultural technology may improve in India if monolingual native speakers have access to information regarding pedology or modern irrigation techniques, enabling them to innovate.

In addition, we also further shed light on Wikipedia as a public good that has an affect on science. (Thompson & Hanley, 2018) demonstrate this and we think our

paper can add to their results through another mechanism.

# 4 Social Impact

This project has the potential to display the importance of open-source knowledge and how Wikipedia can proactively reduce gaps amongst disadvantaged scholars as well as provide more access to non-English speaking developing countries. If the find ings suggest that creating Wikipedia profiles for these groups is positive then this suggests closing scientific gaps may be achieved at a relatively low cost. This may help spur further innovation in these areas.

Additionally, if we find that Wikipedia impacts applied science, then it may support innovation and science in developing countries.

Lastly, it is plausible that these results have increasing returns over time. For example, the translation of articles can not only have a direct effect in the short-run but it can catapult an exponential increase in the network of ideas in a developing countries (Hinnosaar, Hinnosaar, Kummer, & Slivko, 2021).

# 5 Data & Methodology

## 5.1 Data

We will collect scientific information on academics using OpenAlex.

## 5.2 Hypotheses

We believe that it is beneficial for disadvantaged groups to be self-promoted through Wikipedia, given the platform’s role in diffusing information.

Hypothesis 1: The creation of disadvantaged Wikipedia pages will have a positive affect on academic citations, supporting the diffusion of knowledge.

Hypothesis 2: Wikipedia positively impacts applied science and can be used as a tool to support innovation in developing countries.

Data: As mentioned in the preceding section, we intend to search for candidates who are women, non-English speakers, and scholars from developing countries. In addition to laying out a plan in which we will select these candidates based on field of study, publication record, and number of current citations, another way to collect our data is to utilize already existing mechanisms, such as Wikipedia's Women in Science Edit a-thon, whose purpose is to increase female representation by creating profiles for women in science. We can advance mutual objectives in which not only do we create a potential pipeline of potential candidates who are qualified for a Wikipedia profile, we also gain by designing the experiment in which we have a control and treatment group. In the absence of a randomized control trial, this will be the method of data collection. Although this is strictly less preferred to the randomized control trial due it it's lower level of experimental control.

Once the data is collected, we will randomly assign a treatment and control group and implement a randomized control trial using Ordinary Least Squares

Model Specification The Model will take the following form:

$$\Delta citations_i = \beta_1 + \beta_2 T\ reatment_{i,t} + \epsilon_{i,t}$$

Where citations is the change in the outcome variable. We are interested in a post period *t+k* for a candidate *i*. We are then running an OLS regression to compare the means of the treated articles, so $\beta * T\ reatment_{i,t}$ is our parameter of interest.

Q2: Does translating Wikipedia articles to non-English languages increase the levels of innovation proxied by patent citations in developing countries?

We believe that translating articles in science to languages spoken in developing countries can significantly increase the current pool of ideas and increase innovation.

Hypothesis 1:Topics that are encompassed in the Wikipedia articles that will be translated will have a positive effect in the country's innovation, leading to more patent citations to the areas of the translated topics.

Data: To collect a pool of potential articles that will be translated for the implemen tation of non-English speaking countries, we will need to identify the current gaps

in what is and what is not already translated. Since we are interested in identifying the causal effects of article translation, we need to isolate topics to those that have little exposure to speakers of a certain language. One concern is that a non-English language also needs to have a large pool of Wikipedia readers. According to the Wikimedia traffic analysis, India has a significant amount of Hindi readers. With that said, it is likely that there are better chances to find translators from English Hindi. We will then likely start with translations to Hindi, but are willing to expand to other languages if we find qualified translators and a sizeable pool of readers.

Model Specification The Model will take the following form:

$$\Delta patentcitations_i = \beta_1 + \beta_2 T\ reatment_{i,t} + \epsilon_{i,t}$$

Where *patentcitations* is change the outcome variable we are interested in a post period *t+k* for a candidate *i*. We are then running an OLS regression to compare the means of the treated articles, so $\beta * T\ reatment_{i,t}$ is our parameter of interest.

.

# 6 Timeline

The timeline of this project will come in three stages:

1. Stage 1: Preparation. This will take 6 months to commence the pre-training of a PhD student or alternatively hire research assistants to create Wikipedia articles on our behalf as per (Thompson & Hanley, 2018).

   Additionally, we will construct descriptive statistics on past profile creation of disadvantaged groups as well as past article translations, this will serve as preliminary results as well as motivate our RCT design. Questions 3 and 4 above demonstrate our short-run results. In the initial six months, we will implement these, ensuring our project has some tangible results, that don't rely on a randomixed control trial.

2. Stage 2: Implementation. This will take 3 months, given the preparation in Stage 1.This stage will be where Wikipedia profiles will be created and articles translated.
3. Stage 3. This will take 3 months. We will prepare the code so its ready for statistical analysis, once a sufficient amount of time has passed since the completion of the randomized control trial

## 7 Ethical Compliance

We will ensure that we respect all privacy conditions. Any profile creation or transla tion of article will not violate the terms of Wikipedia. In the next section, there will be greater elaboration on the ethical concerns reviewers had in the previous stage.

## 8 Response to Reviewers Comments

Response to ethical concerns: This project will require that all profile creations will be on the right side of WP:NOTLAB. Given that there will not be any editing of an existing Wikipedia article, rather only the creation of new profiles as well as the addition of newly constructed articles on certain scholarly topics. Given this, we still will provide a framework in which we prevent going against Wikipedia norms, such as identifying the candidates whose profile will be created as well, ensuring that all information posted is academically related only. In addition we will identify early-on any topics that will be translated into a non-English language. For the translated articles, these will be referenced to the original English article but in no case will there be any editing to the original articles.

Response to ethical concerns: In addition, any concerns will be willingly discussed on Wikipedia’s Village pump to also hear additional advice. Specific training will be given to the individuals responsible for creating the new profiles, with preference in delegating this task to individuals who has prior experience in tasks related to Wikipedia.

Response to metrics used for analysis: The long-term effect sees the marginal change in citations for the treated individuals. It also can measure the increase of patent citations and the marginal effect of a casual shock in information to individuals who otherwise could not access information on scientific topics.

Response to suggested partners: Additionally any project affiliations would be accepted and would be happy to work on projects with individuals such as Emmanuel Dabo. If successful in the grant, we will certainly reach out to Emmanuel for collab oration

Response to reinforcing inequalities: Although, it is likely that the creation of scholars considered ”eminent” can reinforce existing inequities on an academic dimension, any negative effect can by offset by the creation of eminent scholars of disadvantaged groups who in-turn receive more citations. Current inequities already come in the form of under-representation of disadvantaged groups and lack of representa tion of “less eminent” scholars, however by potentially solving the problem of under representation, we are creating a pathway of a more meritocratic promotion of schol ars.

Response to use of budget: Both projects as mentioned above will utilize the budget to either train to understand the Wikipedia processes or alternatively hire an experienced Wikipedian with a skill in the Hindi Wiki, although we can expand to other languages as well. The majority of the budget will be used for research assistants to ensure the project is executed in a timely and high quality fashion.

# 9 Budgeting

As laid out in the budget section, the majority of the grant (35,000 USD) will be used to hire research assistants or a PhD student for one dedicated year to implement the construction of the Wikipedia proposals. This requires extensive training or the hiring of research assistants who are well acquainted with Wikipedia’s protocols.

In addition, the PI will dedicate three months of full time work during the Summer of 2022 to the project. This will cost 5000USD.

The remainder of the costs will be allocated to presenting the paper at conferences, open source journal costs, and computer equipment needed for computationally in tensive statistical/econometric analysis.

## 10 Dissemination of Scientific Findings

I plan to undertake the following tasks in order to boost the dissemination of the newly created data sets and research findings.

- Workshop: One-day workshop in the Fall of 2025 for academics and policymak ers on the diffusion of knowledge for basic and applied science.
- Seminar Presentations: the paper will be presented at UC Merced's internal seminar. Additionally, I have close connections with UC Berkely, and can reach out to present my work their

- Publications: I will aim to publish one paper at either a management or eco nomics journal. This will be open source, to ensure the knowledge is appropri ately diffused.

## 11 Impacting Graduate Student's Future Careers

This project will be conducted from beginning to end by Kris Gulati and research assistants. He is a PhD student in Economics at the University of California, Merced (UC Merced). He will play an important part in managing the data collection, data cleaning, as well as formalizing the statistical analysis. He will be advised by his cur rent academic mentor Christian Fons-Rosen. Christian is a professor of Economics at UC Merced, specializing in research relating to innovation and science. He has pub lished in numerous academic journals not limited to the *American Economic Review, Management Science, European Economic Review* and *The Journal of International Economics*.

Kris is a 1st year economics graduate student at UC Merced. He has a BA (Hons) from the University of London: Goldsmiths College, an MSc in Econometrics and Economics from The University of Nottingham (with Distinction), and the first year of a Mathematics BSc from The Open University (First Class). He also has research assistance experience at various top academic institutions: The London School of Economics, The University of Nottingham, and University College London. This experience is advantageous given that prior work on Wikipedia as a tool for infor mation diffusion such as that from Toomas and Marit Hinnosaar are also based at the University of Nottingham, an institution Kris can reconnect with for advice on implementing the project.

He is also a first-generation university student and originates from a single-parent low socio-economic background, which provides a unique perspective advantage when addressing the pressing issues of inequality in a research setting.

## References

Agarwal, R., & Gaule, P. (2020). Invisible geniuses: Could the knowledge frontier advance faster? *American Economic Review: Insights*, *2* (4), 409–24. Exley, C. L., & Kessler, J. B. (2019). *The gender gap in self-promotion* (Tech. Rep.). National Bureau of Economic Research.

Hinnosaar, M., Hinnosaar, T., Kummer, M. E., & Slivko, O. (2021). Externali ties in knowledge production: Evidence from a randomized field experiment. *Experimental Economics*, 1–28.

Rose, M. E., & Baruffaldi, S. (2020). Finding doppelg¨angers in scopus: How to build scientists control groups using sosia. *Max Planck Institute for Innovation & Competition Research Paper* (20-20).

Thompson, N., & Hanley, D. (2018). Science is shaped by wikipedia: Evidence from a randomized control trial.