Science is Shaped by Wikipedia: Evidence From a Randomized Control Trial*

Neil C. Thompson
MIT Sloan School of Management &
MIT Computer Science and Artificial Intelligence Lab

Douglas Hanley
University of Pittsburgh

Abstract

"I sometimes think that general and popular treatises are almost as important for the progress of science as original work." — Charles Darwin, 1865

As the largest encyclopedia in the world, it is not surprising that Wikipedia *reflects* the state of scientific knowledge. However, Wikipedia is also one of the most accessed websites in the world, including by scientists, which suggests that it also has the potential to *shape* science. This paper shows that it does.

Incorporating ideas into Wikipedia leads to those ideas being used more in the scientific literature. We provide ***correlational*** evidence of this across thousands of Wikipedia articles and ***causal*** evidence of it through a randomized control trial where we add new scientific content to Wikipedia. We find that the causal impact is strong, with Wikipedia influencing roughly one in every ∼830 words in related scientific journal articles. We also find causal evidence that the scientific articles referenced in Wikipedia receive more citations, suggesting that Wikipedia complements the traditional journal system by pointing researchers to key underlying scientific articles.

Our findings speak not only to the influence of Wikipedia, but more broadly to the influence of repositories of scientific knowledge and the role that they play in the creation of scientific knowledge. *JEL Codes:* O31, O33, O32

*The authors would like to thank MIT for research funding and for the provision of a ridiculous amount of computing resources, Elsevier for access to full-text journal data, Dario Taraborelli at the Wikimedia Foundation for guidance, and Caroline Fry for excellent research assistance.

1 Introduction

In a letter to fellow biologist T.H. Huxley in 1865, Charles Darwin wrote "I sometimes think that general and popular treatises are almost as important for the progress of science as original work" (Lightman 2007, p 355). And, tellingly, *On the Origin of Species* was both a seminal scientific work and a bestseller (Radford 2008).

This paper asks whether "general and popular treatises" themselves feed back into science and help shape it. Rephrasing this into the language of economics, we ask whether the provision of known scientific knowledge in an open, accessible repository can shape the scientific discussion of those ideas – and, in particular, whether Wikipedia already does. This is an important public policy question because it has been known since at least Samuelson (1954) that public goods, of which public repositories of knowledge are a good example, are underprovisioned in a market setting. They are thus good candidates for welfare-improving interventions by governments, organizations, and public-spirited individuals.

Governments already embrace the role of providing public goods for science in a number of contexts by funding scientific repositories. These include repositories of physical objects, like seed banks (NCGRP 2005) and model organism repositories (MMRRC 2017), and there is good evidence that this promotes scientific activity (Furman and Stern, 2011). Governments also fund some informational repositories, for example those related to the human genome project (NIH 2017). Many repositories are also run by organizations or individuals. For example, StackOverflow.com, is a widely used question-and-answer repository for knowledge about computer programming.

Conversely, the most extensive repositories of scientific knowledge – academic journals – remain overwhelmingly financed by subscription fees, thereby restricting access. But what if many of the key insights from those journal articles were also available in an easily accessible public repository?

Wikipedia is one of the largest informational public goods providers for science. It is freely available, easily accessible, and is the 5^{th} most visited website in the world (Alexa 2017). A wide variety of scientific topics are covered on Wikipedia, and a substantial fraction of Wikipedia articles are on scientific topics. Depending on the definition and methods used, Wikipedia has 0.5-1.0 million scientific articles, representing one article for every $\sim$120 scientific journal articles. The scientific sophistication of these articles can be substantial. Based on spot testing in Chemistry, we find that Wikipedia covers more than 90% of the topics discussed at the undergraduate level at top-tier research universities, and about half of those covered at the introductory graduate level.

Given this extensive coverage, it is clear that Wikipedia reflects science. But does it also shape science? Do scientists read Wikipedia articles and encounter news ideas? Or perhaps scientists encounter ideas on

Wikipedia that they are already aware of, but which are brought together in a way that influences how they think about them? One could imagine, for example, that in a broad academic field, a concept from one part of the literature might not have been encountered by people from another until it is seen on Wikipedia.[1] A further possibility is that a scientist could lack access to costly journals, and thus the appearance of an idea on Wikipedia could be that person's only access to that scientific knowledge.

To assess the influence of Wikipedia we need a way to measure the impact that it is having on the academic literature. A traditional way of measuring this would be to count academic citations, the acknowledgements that the scientists themselves make in their publications. Unfortunately, measuring the impact of Wikipedia using citations is difficult for two reasons. First, purported experts might be reluctant to admit that they referenced an encyclopedia for their knowledge, and thus not cite Wikipedia even if they used it. Indeed, university guidelines specifically discourage the citation of Wikipedia, as MIT citation guidelines make clear (MIT, 2017):

> *"**Wikipedia is Not a Reliable Academic Source***
>
> *Many of us use Wikipedia as a source of information when we want a quick explanation of something. However, Wikipedia or other wikis, collaborative information sites contributed to by a variety of people, are not considered reliable sources for academic citation, and you should not use them as sources in an academic paper."*

A second challenge to measuring the impact of Wikipedia with citations is that, even if an author were willing to cite an encyclopedia, they might not feel a need to. As Princeton's Academic Integrity Statement advises (Princeton, 2017):

> *"If the fact or information is generally known and accepted—for example, that Woodrow Wilson served as president of both Princeton University and the United States, or that Avogadro's number is 6.02x10^{23}—you do not need to cite a source"*

It is quite plausible that a researcher, finding that a fact is present in an encyclopaedia, might conclude that the fact is "generally known" and therefore would not feel obliged to cite it. Together, these challenges suggest that citations will not be an accurate way to assess Wikipedia's impact.

We measure the impact of Wikipedia on academic science in two ways: (i) a Big Data approach, and (ii) an experimental approach. Our Big Data approach identifies word-usage in Wikipedia and looks for similar patterns in the full text of academic journal articles. We do this using a full edit-history of Wikipedia (20 terabytes) and full-text versions of every article from 1995 onward from more than 2,000 Elsevier academic journals (0.6 terabytes). This allows us to look at the addition of *any* Wikipedia article and to ask if afterwards the prose in the scientific literature echoes the Wikipedia article's. The advantage of this approach

[1] This happened to one of the authors (Thompson) with regard to the *many* flavors of t-tests. He was reminded of the panoply here: https://en.wikipedia.org/wiki/Student's_t-test

is that we can look very broadly across Wikipedia articles. The disadvantage is that our results are only correlational; they cannot establish causality. This is an important weakness since it cannot rule out the plausible alternatives, particularly mutual causation. In this case, that a breakthrough scientific article would generate both a Wikipedia article and follow-on articles in the scientific literature. This would induce correlation between Wikipedia and the follow-on articles, but it would *not* indicate that it is Wikipedia that is *shaping* science. It seems obvious that this mechanism is occurring. The interesting question thus becomes if there is an additional impact that Wikipedia itself is causing.

To establish the *causal* impact of Wikipedia, we performed an experiment. We commissioned subject matter experts to create new Wikipedia articles on scientific topics not covered in Wikipedia. These newly-created articles were randomized, with half being added to Wikipedia and half being held back as a control group.[2] If Wikipedia shapes the scientific literature, then the text from the treatment group articles should appear more often in the scientific literature than the text from the control articles. We find exactly that; the word-usage patterns from the treatment group show up more in the prose in the scientific literature than do those from the control group and that these effects are large.

2 Public Goods in Science

The underprovision problem of public goods is a well-researched topic. Since at least Samuelson (1954), it has been known that private incentives are less than the welfare-maximizing level because they fail to capture the spillover benefits to others. Under these conditions, there is underprovision of the public good absent intervention by governments, organizations, or public-spirited individuals. The underprovision of *information goods* is particularly worrisome because the welfare losses can be larger and more common. The welfare losses can be larger because of a "long-tail" of users with low individual values for the public could. Because information goods can be costlessly copied, providing these goods to the many low valuation individuals could still create a large surplus. Conversely, if these goods aren't provided, this could could collectively represent a substantial welfare loss. The underprovision problem might also be more common with information goods because free-riding on information goods may be easier than on other public goods, leading to fewer initial contributions.

A common way of resolving public goods problems is to make information excludable, for example by putting information into for-pay journals.[3] Under these circumstances, those benefiting from positive spillovers will not be able to free-ride, potentially leading to better incentives for private provision, though at the cost of excluding some consumers from the market. For example, these restrictions could exclude either customers who don't value the good very much or those who value it highly but are budget constrained. The latter would be particularly worrisome since it would represent a larger welfare loss and exacerbate inequity.

[2] Note: *both* sets of articles need to be written, since the analysis is lexical, and thus the wording of the control articles is important.

[3] In this case, the goods should technically be called "club goods"

The challenge of informational public goods for the scientific literature is, however, worse than the analysis above might suggest. This is because, absent actually reading a scientific article, it may be hard to assess its value to you – that is, due to Arrow's Information Paradox (Arrow, 1962):

> *"there is a fundamental paradox in the determination of demand for information; its value for the purchaser is not known until he has the information, but then he has in effect acquired it without cost"*

So, to avoid giving away their content for free, journals need to prevent potential consumers from reading an article before they purchase it. But being unable able to read articles, it might be hard for consumers to determine their valuation (e.g. whether the article will help solve a problem). This will render consumers unwilling to pay their full marginal value, reflecting the uncertainty of whether or not the article will be valuable. As a result, even consumers with marginal values higher than the cost might choose not to purchase the article, magnifying the welfare loss.

There are several distressing implications that arise from Arrow's Information Paradox and raising the price of scientific information goods above marginal cost. First, information is likely to become siloed, with only the most valuable articles from one area crossing over to another. This is a natural consequence of the discussion above, since information in neighboring fields is likely to be less valuable and the probability of recognizing a good article is also likely to fall. Thus only the highest-quality work is likely to be paid for, and much of the potentially useful sharing of knowledge between fields will be stifled.[4]

Even within a field of knowledge, the implications of this matching process between scholars and articles are likely to be troubling. It is known that the citation patterns of scientists follow a power law, and thus there exists a long tail of articles that are seldom cited.[5] For some articles, a lack of citations probably indicates a lack of quality. But other seldom-cited articles may be of high quality but targeted to a limited readership, perhaps to specialists in their narrow field.[6] In these instances, siloing is again likely to be caused by restricted access and Arrow's information paradox.

Together these examples highlight a difficult problem with the dissemination of scientific information. A fully open-source model will have too few private incentives, and will require substantial market intervention to avoid underprovision. In contrast, a fully-closed model is likely to substantially curtail welfare-improving spillovers both to those with low marginal value and those who are unable to tell if the underlying articles are worth paying for.

The model of scientific information sharing embodied in Wikipedia is a middle ground between these. It provides a free, widely-accessible summary of the scientific findings with links to the underlying papers, perhaps similar to what one might find in a review article. Because the summaries are free but provide more detail than many scientific abstracts, such a repository may facilitate many of the positive spillovers to those

[4] This also corresponds to the experience of one of the authors (Thompson) when he was in the business world and was unable to gather insights from academic articles, because of the fees associated with accessing them.

[5] The exact type of power law is debated, as can be seen here: https://arxiv.org/abs/1402.3890

[6] Increasing specialization, as observed by Wuchty et al. (2007) suggests that this issue could be becoming more prevalent.

outside the field or those whose marginal value would otherwise be low. They may also help to resolve some of Arrow's Information Paradox, by pointing researchers to the key articles that would be worth paying for – that is, by acting as a filter about which scientific articles on the topic are important and certifying their value.

In this paper we test several of the implications about public goods in science, in particular (i) is Wikipedia used to inform academic research, (ii) does it generate demand for the underlying papers, and (iii) are those with less access to journals more likely to benefit.

3 Wikipedia

Wikipedia is a user-generated and edited online encyclopedia, currently the largest of its kind. It was founded by Jimmy Wales and Larry Sanger in early 2001 and has seen continual growth since that time. Though it was originally launched in English, it currently has wikis in over 250 languages. For the purposes of this study, we focus only on English-language Wikipedia.[7]

Wikipedia has 5.3 million articles. These were written and are edited by a total of about 30 million registered editors of whom roughly 120 thousand are currently active (Wikipedia). In the past decade, there has been a consistent average of 30 million edits per year (authors' calculation), which includes both the creation of new articles and development of existing ones. Not surprisingly, a small number of very active editors contribute an outsize share of edits. Suh et al. (2009) find that editors averaging more than 1000 edits per month account for only 1% of editors but make 55% of edits.

Editors of Wikipedia are not representative of the general population. For example, there is a widely discussed gender gap. An opt-in survey of visitors done by Glott et al. (2010) found that only 31% of readers and 13% of editors are female. The Wikimedia Foundation (which operates Wikipedia) has taken steps to correct this but has thus far not succeeded to a substantial extent.

The editing community enforces certain codified rules designed to ensure accuracy and prevent bias in articles. A study comparing the accuracy of various scientific topics in Wikipedia and Encyclopaedia Britannica found that they occurred at similar rates between the two (Giles, 2005). In particular, a Wikipedia science article contained an average of four "inaccuracies," while an Encyclopaedia Britannica article contained only three. While the error rates between the two may be comparable, the volume of scientific information available on them is not: Encyclopaedia Britannica currently has about 65,000 articles totalling 40 million words (Wikipedia), while English Wikipedia has about 5.3 million articles totalling 1.8 billion words.

Wikipedia is very widely read. According to Alexa, a major web analytics company, Wikipedia is the fifth most visited website on the internet, both globally and when restricting to only the US. As of 2014, it served a total of 18 billion page views to 500 million unique visitors each month, suggesting that a substantial

[7] For the experiment, we checked to see whether our articles were translated into other languages, which might have made looking at them interesting as well. We find almost no translations that happen this quickly.

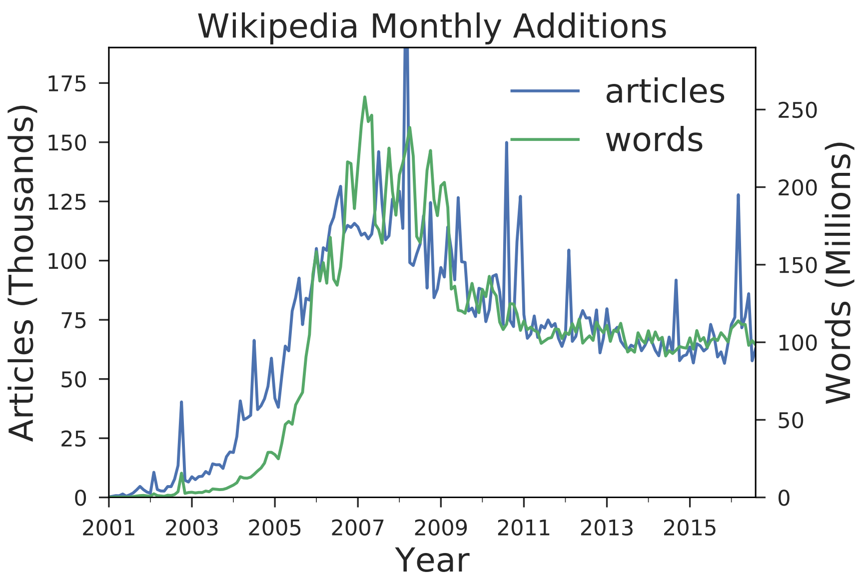

Figure 1: Words and articles added to Wikipedia since its inception

fraction of humanity is using Wikipedia.

The Wikimedia Foundation is a non-profit that operates Wikipedia, as well as numerous related projects such as Wikidata (for structured data), Wikisource (a repository for original source texts), and Wiktionary (an open dictionary). It currently has over 200 employees. The website is run using open-source software, much of it developed in house in the form of the MediaWiki platform. This platform has come to be widely used by other wikis, including those not associated with the Wikimedia Foundation.

In the 2015-2016 fiscal year, the Wikimedia foundation had $82 million in revenue and $66 million in expenses. To put these numbers into perspective, the American Type Culture Collection (a major biological research center) has a budget of $92 million (GuideStar, 2017), and Addgene (the non-profit plasmid repository) has a budget of $8.5 million (D&B Hoovers, 2017).

A wide variety of scientific topics are covered on Wikipedia, and a substantial fraction of Wikipedia articles are on scientific topics. Determining exactly which articles do or do not constitute science is somewhat subjective. Depending on the definition and methods used, roughly 10-20% of Wikipedia articles are on scientific topics (between 0.5-1.0 million out of a total of about 5 million).[8] Based on spot testing in Chemistry, we observe that Wikipedia covers more than 90% of the topics discussed at the undergraduate level at top-tier research universities, but only about half of those covered at the introductory graduate level. There exists substantial interest in the open-source committee for continuing to deepen the scientific knowledge on Wikipedia (Shafee et al., 2017).

[8] To determine which articles are considered Chemistry, we rely on Wikipedia's user generated category system. This tends to pull in far too many articles though, so we take the additional steps of paring the category tree using a PageRank criterion and hand classifying a subsample of candidate Chemistry articles and using them to train a text-based Support Vector Classifier.

Wikipedia is also used by professionals, for *scientific* information. For example, a 2009 study of junior physicians found that in a given week 70% checked Wikipedia *for medical information* and that those same physicians checked Wikipedia for 26% of their cases (Hughes et al, 2009).

Previous research by Biasi and Moser on German textbooks in WWII (2017) showed that lowering the cost of scientific information (and thus making it more accessible) led to substantial changes in scientific publishing. Since Wikipedia is also making scientific information cheaper and more widely accessible, we would expect that it too would have an influence on the scientific literature. However, evidence of this effect is largely absent from the usual place where one would look for it: citations from the academic literature. Tomaszewski and MacDonald (2016) find that only 0.01% of scientific articles directly cite Wikipedia entries.

We hypothesize that this is not because Wikipedia doesn't have an effect, but rather that academic citations are not capturing the effect that Wikipedia has. To test this, we develop a text-based measure, where we can measure this effect directly in the words used by scientists.

4 Data

This paper relies on four major sources of data. The first is a complete edit history of Wikipedia, which includes every change to every page since Wikipedia's inception. The second is a full-text version of all articles since 1995 across more than 2,061 Elsevier journals, which we use to represent the state of the scientific literature. The third is data on citations to academic journals, which we get from Web of Science. These three sources are described in this Section. The fourth data source is a set of Wikipedia articles created as part of the randomized control experiment. We discuss these as part of the experimental design in Section 7.

4.1 Wikipedia

The Wikimedia Foundation provides the full history of all edits to each article on Wikipedia. This includes a variety of projects run by the foundation, in particular, the numerous languages in which Wikipedia is published. For the purposes of this study, we focus only on the English corpus, as it is the largest and most widely used.

Even restricting to English Wikipedia, there are numerous non-article pages seldom seen by readers. This includes user pages, where registered users can create their own personalized presence; talk pages, one for each article, where editors can discuss and debate article content and editing decisions; pages associated with hosted media files such as images, audio, and video; and much more.[9]

The edit history of Wikipedia is a series of XML files containing information on the evolution of each article. This constitutes an entry for every revision of an article. For each revision, one sees the exact date and time that the revision occurred (a "timestamp"), the username of the editor who committed the change,

[9]There are also redirect pages that allow for multiple name variants for a single source page. These are safely ignored.

and the full text of the article at that particular state. The article content is stored in an internal wiki markup language designed to be easily edited and read in raw text form.

The edit history covers 5.1 million articles, 353 million edits, and 17.4 billion words. The entire database is 20 TB[10] although there is considerable duplication because, with each revision, no matter how minor, a full copy of the article is stored. To get an idea of the information content, using advanced compression algorithms, one can reduce the size to 83 GB. For our analysis, we reduce the history to a stream of new words added and deleted over time. This method reduces the corpus to only 118 GB.

4.1.1 Article Creation

Every month thousands of new Wikipedia articles are created. Figure 2 plots these (and the corresponding word additions) across all of Wikipedia and for the two scientific disciplines that will be relevant for our randomized control trial: chemistry and econometrics.

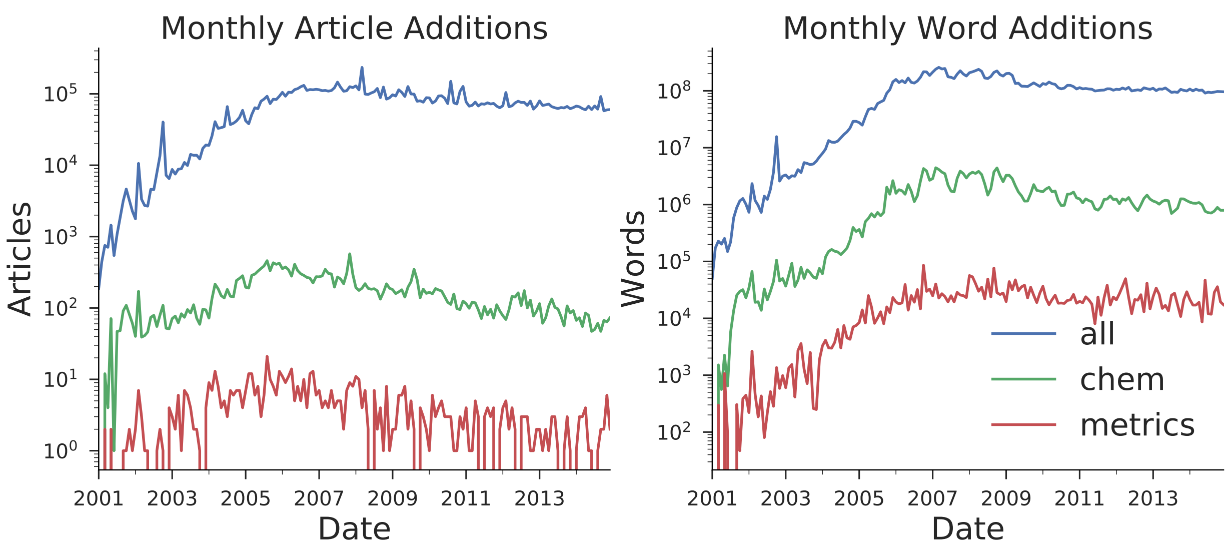

Figure 2: Monthly article and word additions to Wikipedia

Each of the three time series show very similar profiles in terms of growth, but at different scales. Chemistry articles are added in greater numbers and with greater total volume than econometrics. Interestingly, this gap in Wikipedia additions of 10-100× is significantly larger than the difference in the number of practitioners of the two fields, where Chemists only outnumber Economists ∼4.5× (U.S. Bureau of Labor Statistics, 2016). This indicates that, for whatever reason, economists seem to be less good at capturing their knowledge in Wikipedia.

[10]Technically all the units in this section are in base 2 units, so for example the entire database is 20 tebibytes (TiB). This unit is closely related to the terabyte, which readers may be more familiar with (and which for the sake of our broad description above is sufficiently accurate), but accounts for the binary nature of computer memory which means that there are 210=1024 gibibytes per tebibyte.

Generally speaking, new Wikipedia articles start out quite small and grow slowly over time. Roughly 70% of articles are less than 20 words long upon creation, reflecting the fact that many article begin as a “stub” – a short article, perhaps just a title and single descriptive sentence, that is intended to be built upon in the future. Figure 3 shows an example of an early edit of the Magnesium Sulfate stub, where new additions are underlined and deletions are struck through.

“Magnesium sulfate ~~,~~ ” ~~$MgSO_4$~~, (commonly ~~known as~~ called“ Epsom ~~salts~~ salt” in hydrated form) is ~~used as a therapeutic bath~~ ~~a chemical compound with formula $MgSO_4$~~,

Epsom salt was originally prepared by boiling down mineral waters at Epsom, England and afterwards prepared from sea water. In more recent times, these salts are obtained from certain minerals such as siliceous hydrate of magnesia.

Figure 3: Example of the early editing on the Magnesium Sulfate article

Figure 4 plots the size distribution of newly created articles that are longer than 20 words. Here we can see that the bulk of articles begin at less than 200 words. There is some mass in the tails of the distribution, though this may be due to the renaming or reallocation of large existing articles.

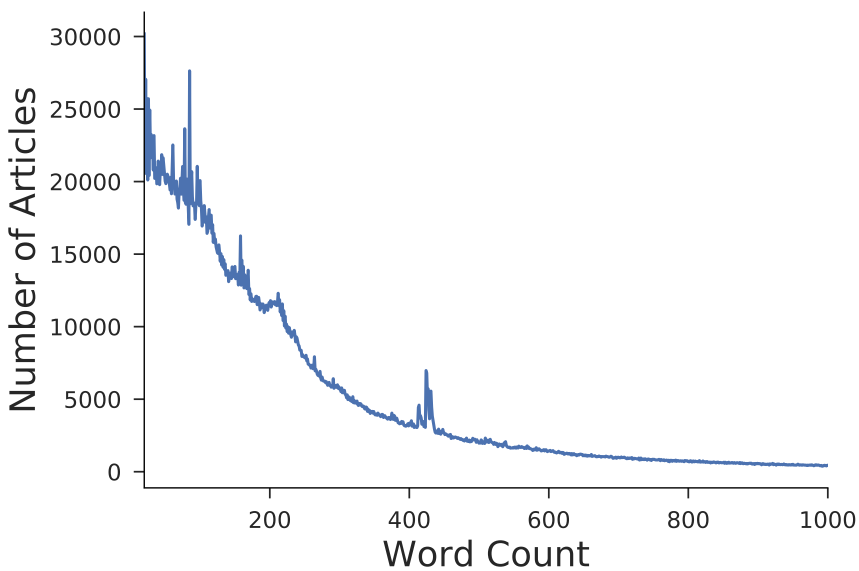

Figure 4: Size distribution of new articles longer than 20 words

In Figure 2 there was some evidence of tapering off in the number of chemistry and econometrics articles being created. This is likely because many of the most important topics in these fields have already been created. Figure 5 corroborates this by plotting how articles grow on average. Interestingly, each of the three cohorts average approximately 250 words when first written. Lengths expand significantly after this, but particularly so for articles earlier – again suggesting that these were on broader, more important topics.

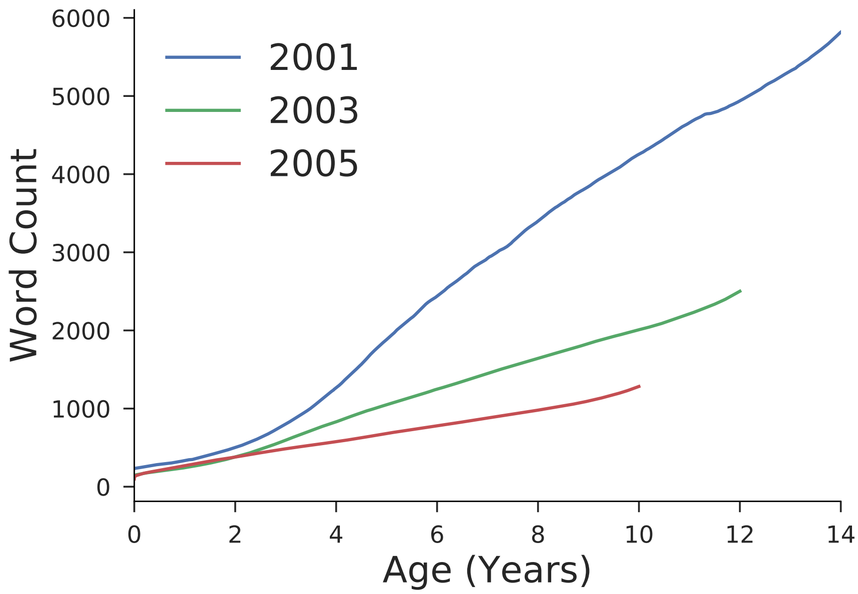

Figure 5: Average size of articles conditional on age (daily)

Finally, in Figure 6 we present the current distribution of article size conditional on being larger than 30 words. Here we see the characteristic long tail extending nearly linearly in log-log space. There are also a large number of articles with very few words. We exclude such “stub” articles from our analysis by imposing a minimum of 500 characters (about 100 words) in each article for inclusion in our sample.

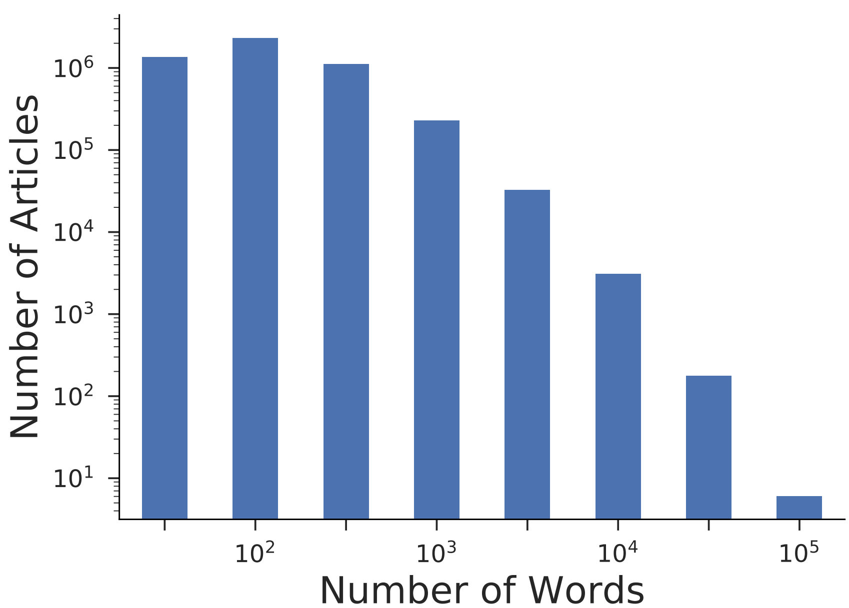

Figure 6: Distribution of article size for articles larger than 30 words

4.1.2 Word Coverage

For our word-level analysis we focus on the Top 90% most common unique words in the scientific journals. The entire vocabulary that we could potentially use contains ∼1.2M words, of which we focus on the most-used ∼1.1M in science. This serves two purposes: (1) it eliminates noise from words with single digit frequencies (and thus where there are large proportional swings in usage), and (2) it avoids issues arising from errors in parsing, non-content strings (such as URLs), and misspellings in the source text. It should be noted that this set of words will include very common ones such as "the" and "a." In subsequent analysis, we use inverse document frequency weighting, which ensures that our results are not being driven by these words. Even after such a cull, the words represented here account for 99% of word usage in science and 72% of word usage in the Chemistry pages of Wikipedia.

We can also consider the overlap in the two vocabularies. We see that they are similar, but also have substantial differences. About 61% of the words in the scientific literature appear in Wikipedia, while amongst the set of words appearing in Wikipedia, about 63% appear in science. The following provides some context for the relative frequency of the words in our data:

Table 1: Comparison of word usage between Wikipedia and scientific literature

Word	**Literature Rank**	**Wikipedia Rank**
Acid	57	65
Reaction	34	132
Graphene	1,975	3,099
Photovoltaic	5,618	7,651
Gravity	6,206	2,375
Pokemon	749,131	14,485

4.1.3 Scientific Fields in Wikipedia

We are interested in investigating the effects that a Wikipedia article has on the corresponding areas in the scientific literature. Doing this requires assigning Wikipedia articles into scientific fields, which is not a trivial undertaking. Our first step is to take advantage of a user-generated categorization scheme, in which editors can tag articles with a particular category, to generate a hierarchical relationship between articles. This induces a category tree.[11] To generate a list of articles in a particular field, we simply look at the top level category (say, Chemistry), find all of its descendant subcategories, then find all pages belonging to such categories.

Unfortunately, this pulls in a large number of false positives, as can be seen by tracing descendents from Chemistry: Chemistry > Laboratory Techniques > Distillation > Distilleries > Yerevan Brandy Company (an Armenian cognac producer).

[11] Surprisingly, this is in fact a cyclic graph. However, it can be pared down to a tree using a small number of edge deletions. In particular, we calculate the PageRank of each node (category) in this graph and eliminate edges from low PageRank nodes to high PageRank nodes. This eliminates only 1% of nodes and renders the graph acyclic (a tree).

To correct for false positives, we hand classify a set of 500 articles and use these to train a support vector machine (SVM) classifier. The SVM maps vectors of word frequencies into a binary classification (in the field or not). The SVM is standard technique in machine learning for tackling high dimensional classification problems such as this one. In the case of chemistry, this process narrows the set of 158,000 potential articles to 27,000 likely chemistry articles in Wikipedia.

4.2 Scientific Literature

The data on the scientific literature is provided by Elsevier and includes the *full text* of articles published in their journals. This is useful for us, since it allows us to look for the words used in the scientific literature and whether they reflect those used in Wikipedia.

In addition, we make use of the article metadata provided, such as author and publication date. The entire dataset includes 2,061 journals over many years. Since we are interested in the interaction of the scientific literature with Wikipedia, we use only data from 2000 onward.

For each article, we observe the journal that it is published in, the year of publication, the journal volume and issue numbers, the title and author of the article, and the full text. We don't make use of any image data representing figures or charts, and equations, since our analysis is word-based. Finally, since journal publication time is often poorly documented (saying, for example, "Spring 2009"), we hand collect this info at the journal issue level for the journals we use.

Focusing specifically on the chemistry literature, which we examine in particular detail, we look at 50 of the highest impact journals, constituting 745,000 articles. Of these, we focus on the 326,000 that are from after 2000.

4.3 Web of Science Citation Data

The data on academic citations is provided by Web of Science. It provides directional links, indicating which papers cite which other ones. This information is also aggregated to provide total monthly citation counts for each paper.

5 Observational Analysis Methodology

The purpose of this first analysis is to establish the broad correlations between word usage in Wikipedia and word usage in the scientific literature. The intent in this section is *not* to establish causation, but rather to establish whether there are contemporaneous changes across Wikipedia and Science over large numbers of articles on many topics. In Sections 6.3 and 8, we return to the question of causality to try to understand how much of these effects are *because* Wikipedia is shaping Science.

5.1 Word Co-occurrence

5.1.1 Documents

In addition to analyzing the usage of individual words, we also take advantage of their arrangement into documents in both corpora. Given a certain set of possible words (a vocabulary) of size K, each document can be represented by a K dimensional vector in which each entry denotes the number of appearances of a particular word. This is referred to as a bag-of-words model, because information on word positions within the text are discarded. These vectors are generally extremely sparse, since only a small number of words are represented in any document.

We can now define the cosine similarity metric between two documents with vectors v_1 and v_2 as

$$d(v_1, v_2) = \sqrt{\frac{v_1 \cdot v_2}{\|v_1\| \ \|v_2\|}}$$

where $\|v\| = \sqrt{v_1 \cdot v_2}$. This satisfies the natural properties that: (1) $d(v_1, v_2) \in [0, 1]$; (2) $d(v, v) = 1$, and (3) $d(v_1, v_2) = 0$ when v_1 and v_2 have non-overlapping bases.

To account for the fact that some words carry more meaning than others, we utilize term frequency-inverse document frequency (tf-idf) weighting to inflate the relative weight given to rarer (and presumably more important) words. In particular, this scheme weights tokens by the inverse of the fraction of all documents that the token appears in. This is a standard metric used text analysis problems.

Most articles in the scientific literature are not that similar to any given Wikipedia article. Figure 7 shows this empirically, plotting the average similarity between all pairs of Wikipedia and scientific articles in our Chemistry sample.

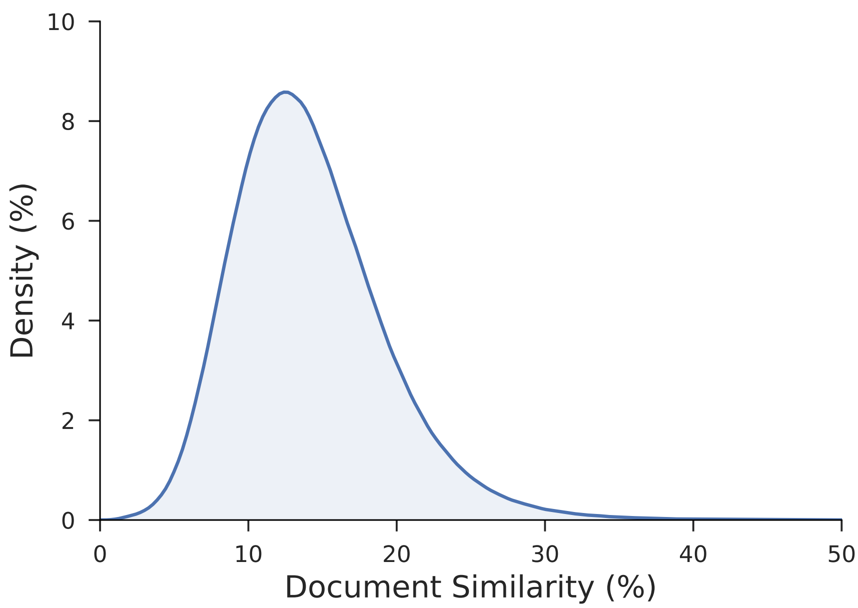

Figure 7: Density of similarity between Wikipedia and scientific articles (all pairs shown)

To estimate the effect of Wikipedia on Science from our observational data requires two elements: an observed correlation between Wikipedia and the change in language use in the scientific literature (raw effect), and a counterfactual about how word usage would have changed absent the Wikipedia article. We calculate the latter first, by considering language drift.

5.1.2 Language Drift

There is a natural ebb and flow to the usage of words in scientific writing over time. Sometimes this is due to the advent of genuinely new concepts or discoveries, such as the "CRISPR" gene-editing technique in biology, or the increased usage of particular scientific techniques. Other times, it may just reflect linguistic "fashion", with some terminology coming into vogue, and others becoming passé. In order to assess the relationship between science and Wikipedia at the document level, we must first characterize what we call the baseline "drift" of word usage frequency.

To accomplish this, we track word usage frequencies in science over time and characterize their evolution. That is, we calculate the probability of a word moving from a particular frequency f_t one year to another potentially different frequency f_{t+1} the next year. In practice, we group words into very finely spaced frequency bins in each year and calculate their Markov transition matrix. We do this without any dependence on the Wikipedia data.

Once we've accomplished this, we can then simulate large numbers of documents to see what it implies for the evolution of document similarities. Specifically, starting at a given time period t, we simulate a large batch of documents using word frequencies from science and Wikipedia.[12] We then generate a new word frequency vector for period $t+1$ using the aforementioned Markov matrix. Finally, we use this vector to generate a sample of new science articles.

As this shows, there is a natural drift (δ) away from the language of previous articles, resulting in lower document similarity as time passes. On average, this effect is a reduction in similarity of 0.074 percentage points over the same time period as we will measure the treatment effect (described below, roughly nine months). The biggest effects are from a fall in the number of documents at similarity $\sim$20%+ down to a level of $\sim$16%. The distribution provided by this drift analysis provides a baseline against which we can measure the raw observational correlation associated with adding a Wikipedia article.

5.1.3 Specifications

We calculate the (raw) effect of adding a Wikipedia article using a regression approach. Let us denote the cosine similarity between Wikipedia article i and scientific article j at time t as d_{ijt}. This notational will include all articles pairs, even those where the scientific article was published before the Wikipedia article.

[12] We assume no correlation between words within documents. In reality, of course, this is not the case. However, entertaining more elaborate theories of synthetic document generation is beyond the scope of this paper and would not substantially impact the contribution of this paper to the literature.

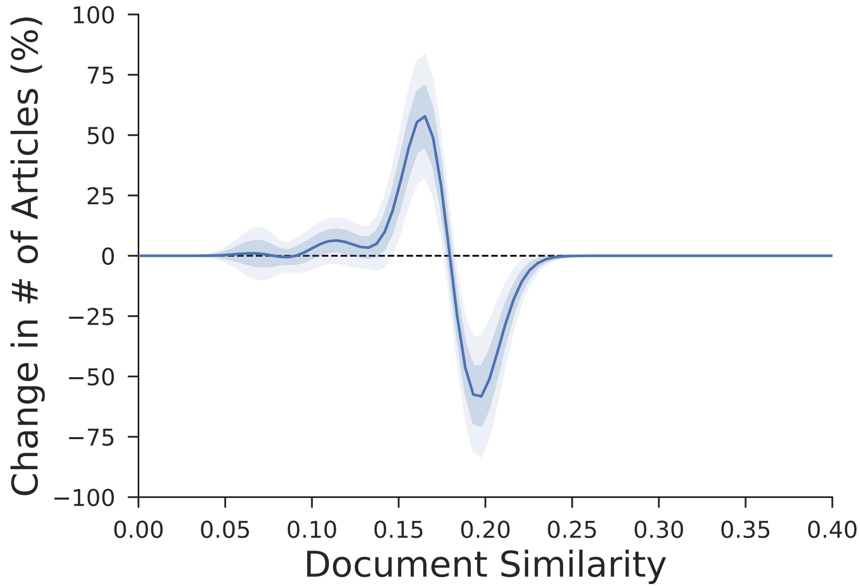

Figure 8: Simulated drift in science-Wikipedia document similarity

Thus, let us also denote by w_{ijt} the binary indicator of whether scientific article j was written after Wikipedia article i.

With our notation defined, we can state the precise specification we use.

$$d_{ijt} \sim \alpha + \tau * w_{ijt}$$

This is essentially a difference in means that comparaes document similarity before and after the Wikipedia article is created. By adding this estimate of the raw treatment effect, τ,to this our synthetic counterfactual to take account of natural drift in scientific lanugage, δ, we get our observational estimator for the net effect on scientific lanugage of adding a Wikipedia article: $\omega = \tau - \delta$.

5.2 Measurement Timeline

In order to examine the relationship between Wikipedia and science, we look at scientific articles shortly before and shortly after the appearance of new Wikipedia article. Our hypothesis is that if Wikipedia has an impact on the progression of the literature, science published after the creation of the Wikipedia article will bear a closer similarity to the article than the science published before it did.

We consider that new article creation in Wikipedia happens over a three-month period, and use the wording at the end of those three months as the "new" article language. This approach to reflect a common article creation sequence in Wikipedia where someone (such as an editor) indicates that a new page *should* be written and creates a placeholder for it (a "stub"), after which subsequent edits are made to fill in the page (Figure 3 shows an example of this). Such stubs are a prevalent phenomenon on Wikipedia (Shafee et al., 2016), and absent this choice we would have a large number of articles "created" with almost no content.[13]

We look for effects of the Wikipedia article on science at two time windows around article creation, one six-month window preceding it and one six-month window starting three months after it. The latter delay also accounts for publication lags in science. The following diagram explains this:

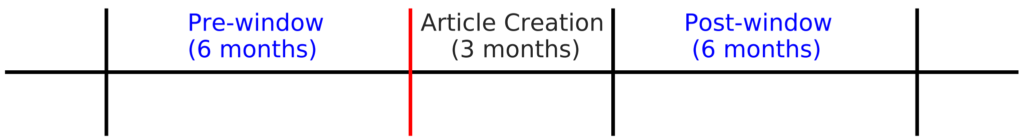

Figure 9: Measurement timeline

For each Wikipedia article, there is a certain set of scientific articles associated with the pre and post windows, respectively. This induces a distribution of similarities (pre and post) for each Wikipedia article. In our analysis, we look at the average difference between these pre and post distributions. If the post distribution shifts closer to the Wikipedia article, it suggests an increased correlation between Wikipedia and the scientific articles.

6 Observational Analysis Results

6.1 Overall correlations

Figure 10 plots the log frequencies of tokens with above median frequency in both Wikipedia and science. The red line shows an OLS regression fit, indicating a rather strong relationship between the relative frequencies of words in the two corpora, but with considerable variance around it.

6.2 Event studies

Figure 11 shows some examples of token frequency time series in the present vocabulary. Each is shown starting from 2001, when Wikipedia started, until nearly the present day. From these we can see that there are words, like "ozone" that seem to exhibit strong correlations in usage between Wikipedia and Science, but others like "reaction" whose trends seem mostly unrelated.

Naturally, coverage in Wikipedia does not always begin immediately, but for most tokens coverage begins between 2005 and 2007. In terms of trends, it is quite common to see a large amount of editing activity near

[13] An example of a newly-created article with almost no content can be seen at https://en.wikipedia.org/wiki/Paracamelus, as of April 2017.

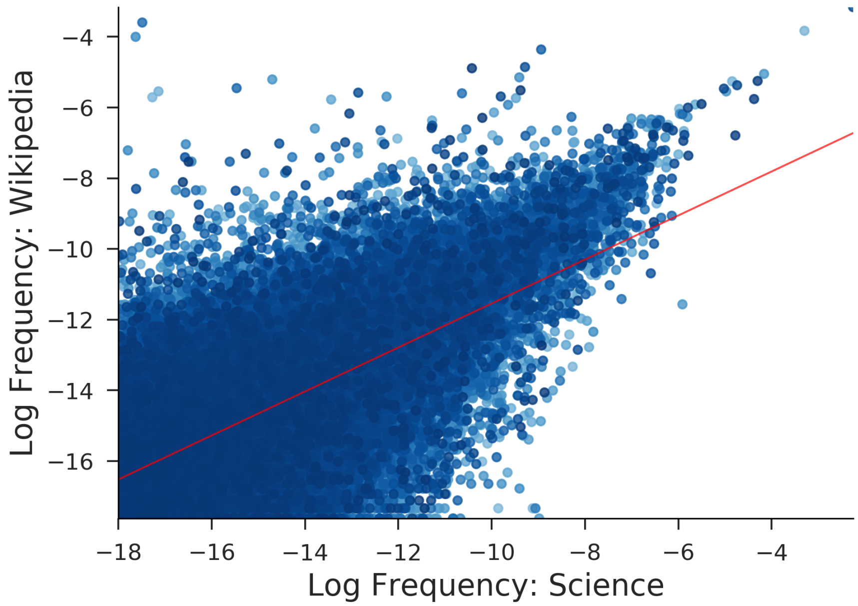

Figure 10: Token frequencies in Wikipedia and Science (above median tokens)

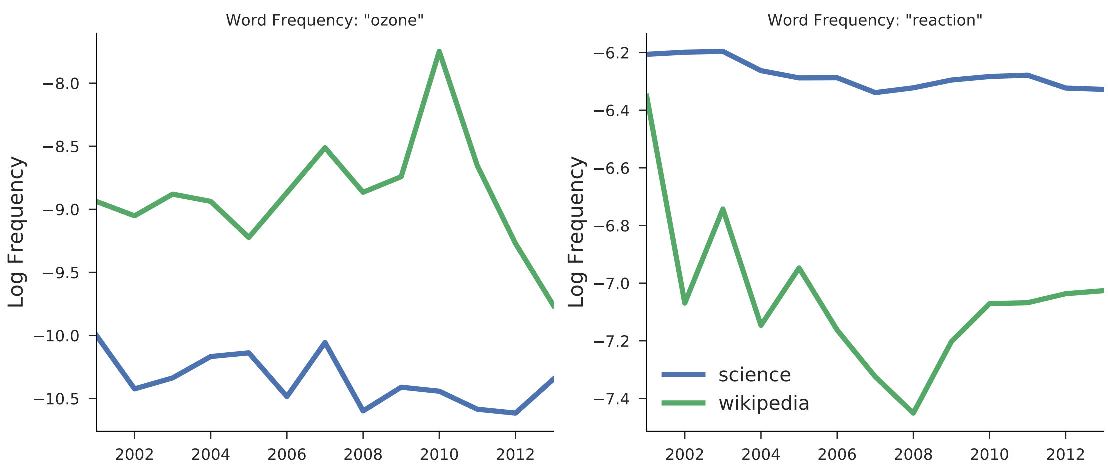

Figure 11: Frequencies of Chemistry words in Wikipedia and in Science

when the token is first introduced, as a main article is built up, followed by a reduction in editing as the content matures. After this, there may or may not be future increases, presumably depending on whether the term enjoys more relevance to research in the future.

We explore the co-occurrence of words in the two corpora in greater detail in Appendix B.

6.3 New Wikipedia Articles

In this section we analyze how similar the scientific literature is to Wikipedia articles when they are created. Our hypothesis is that the scientific literature after the Wikipedia article will be more similar to it that the scientific literature before.

We first look at the raw average effect – are scientific articles published after the Wikipedia article more similar to it?

Table 2: Observational Effect of new Wikipedia Article (not accounting for language drift)

	Similarity (%)
Intercept	14.2494***
	(0.0608)
After	0.0692***
	(0.0261)
N	835379
R^2	0.0000
Adjusted R^2	0.0000
F Statistic	34.9453
Note:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

The positive and highly statistically significant coefficient on "After" in the regression confirms that articles published afterwards are indeed more similar.[14] Having observed the average effect, we now examine the distributional changes in similarity. This approach closely mimics that used (later) in our analysis of the experimental effect, and hence also provides a useful point of comparison for that discussion.

To expose where these differences are occurring, we calculate the probability density function (pdf) for the similarity between the Wikipedia articles and the pre and post scientific articles. We then plot the difference in these densities, Figure 12. Because both pre and post are densities, this differential must integrate to zero, reflecting where the relative densities have risen and fallen.

Figure 12 shows the shift from the lower similarity levels, below ∼12% to higher similarity levels. Here, and in later figures, the estimates are shown as the solid line, and bands of 1 and 2 standard error are shown in increasingly light shading. Showing the absolute change in densities makes it hard to appreciate the scale of these effects, since a small increase in the number of similarity 30% articles may represent a large percentage increase. To make the interpretation of these effects more straightforward, Figure 13 presents the effects as the proportional change at each level of similarity. Thus, for example, the estimate of the small peak at similarity 27% implies an increase of 2.5% in the number of articles of this level of similarity. Recall that while these effects are not individually significant in most narrow windows of document similarity, they are collectively statistically significant as Table 2 showed.

Recall that the estimate from Table 2 represents only the raw effect, of 0.069pp. This effect needs to be contrasted with the counterfactual drift value calculated in Section 5.1.2 of 0.074pp. Thus we get a

[14]Since the unit of analysis here is a Wikipedia-science article pair, we use dyadic standard errors, as discussed in more detail in Section 8.

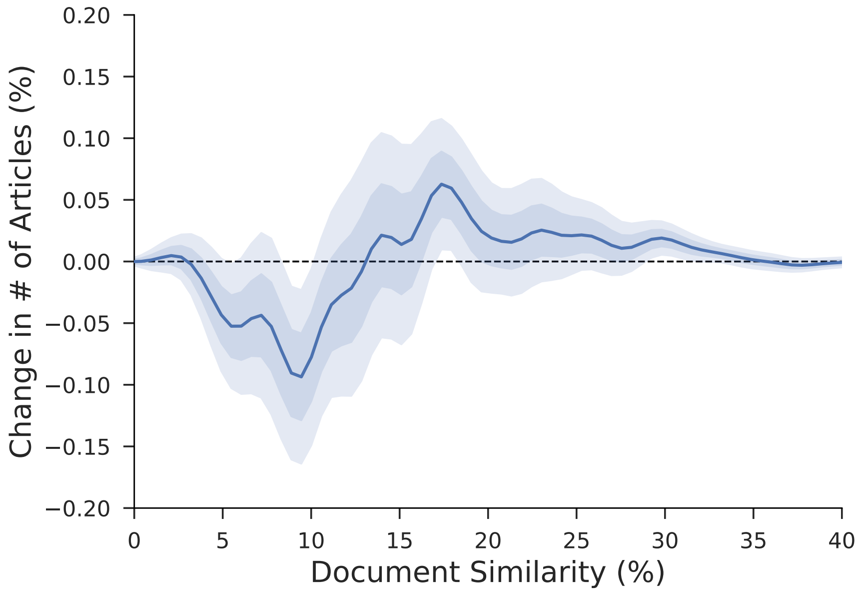

Figure 12: Evolution of the similarity between Wikipedia and science after a new Wikipedia article (absolute change in density), not accounting for language drift

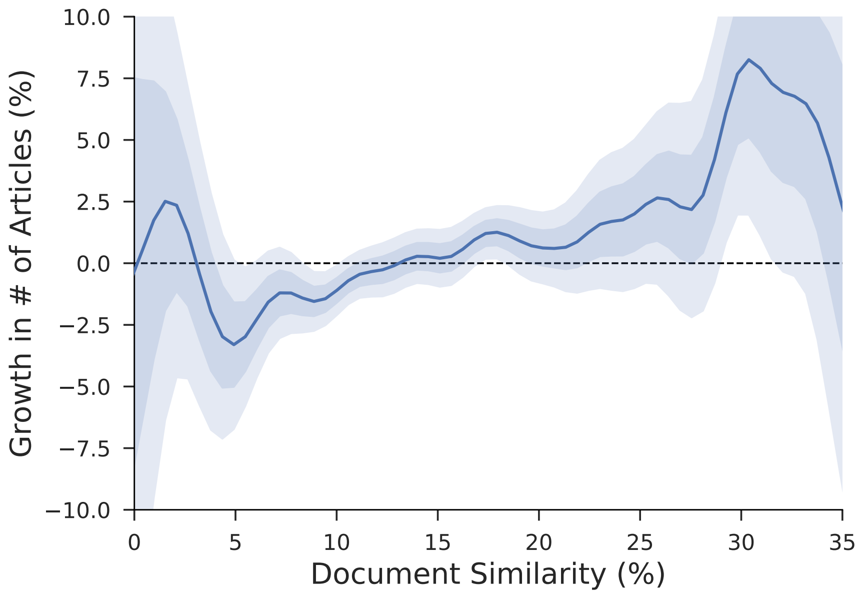

Figure 13: Proportional change in density of similarity between pre and post windows, not accounting for language drift

correlational effect of $\omega = 0.069pp - (-0.074pp) = 0.14pp$.[15] In words: the addition of a Wikipedia article is correlated with a 0.14 percentage point movement of language in the scientific literature towards it.

[15] Our experimental setup, presented next, will not require this drift adjustment because the control group can be used directly as the counterfactual.

To help put this finding in context, we can compare it with another effect: the change in the academic literature that arises when a scientific journal publishes a review article. Figure 14 compares these, showing the observational effect from the Wikipedia article and a comparable line for review articles. The latter is calculated by re-running of the observational analysis, but substituting contemporaneous review articles, instead of Wikipedia articles, as the treatment.[16]

These curves suggest that the effect on the scientific literature from a Wikipedia article is similar, but weaker, to the effect from a review article. Interestingly, both are notably different from the effect of non-review journal article (not shown), which show relatively little effect.

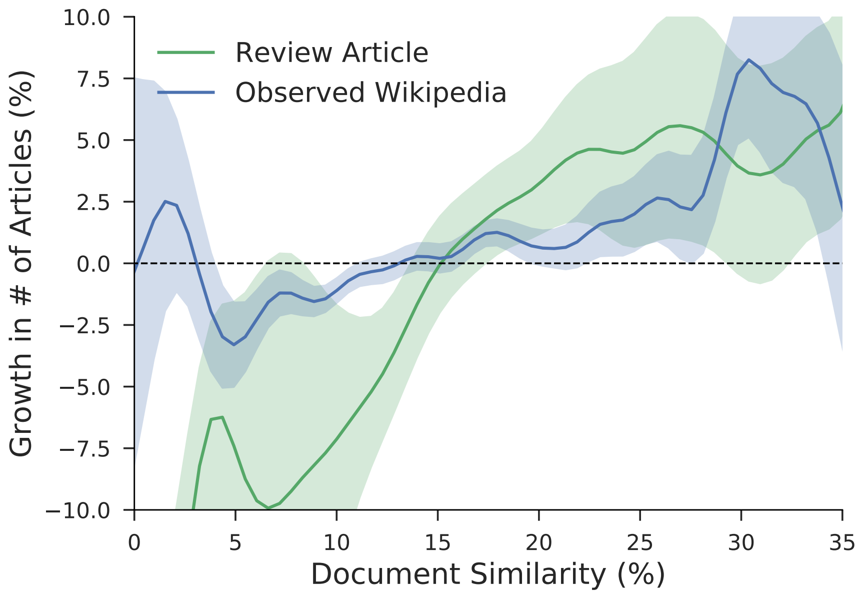

Figure 14: Proportional change in density of similarity after scientific review articles and observed Wikipedia articles

Although the correlations presented in this section are suggestive, they are not causal. It is possible that they represent an effect that Wikipedia is having on the scientific literature. But such effects are indistinguishable from a different causal pathway, mutual causation, in which new scientific breakthroughs generate both a Wikipedia article and more follow-on work. Establishing the causal effect of a Wikipedia article requires turning to our experiment.

7 Experimental Design

From 2013-2016 we ran an experiment to ascertain the *causal* impact of Wikipedia on academic science. New scientific articles were written by PhD students in those areas. Half of those articles were uploaded to

[16]We look at only the review articles published at the time of the Wikipedia articles so that they are both being tested against the same scientific literature, and hence are comparable analyses.

Wikipedia, while half were held back.[17] We then considered the differential impact that adding these articles to Wikipedia had on the scientific literature.

The experiment was run in two waves, first a wave in Chemistry (January 2015 - 43 articles created) and then in Econometrics (November 2015 - 45 articles created). The main text of this article concerns only the Chemistry wave. It turns out that the rest of the world was less excited by econometrics than the authors of this paper, and so the average views of the Chemistry articles were more than thirty times those of the Econometrics pages! With so few views by the Econometrics community the second experimental wave is underpowered and thus we do not discuss it here (although for the sake of full disclosure, we do report additional details in Appendix A).

7.1 Article Creation

To create the Wikipedia articles for this experiment we followed the following process:

1. Generate a list of potential Wikipedia article topics of science from textbooks or course syllabi from leading universities

2. Have subject experts check whether the potential topics were already present in existing Wikipedia pages

3. Commission subject experts to create new articles for the topics not already covered in Wikipedia

Using personal connections and online research we located textbooks and course syllabi for upper-level undergraduate and introductory graduate level classes from several prominent universities (Harvard, MIT, Berkeley, and Cambridge). PhD students in Chemistry then reviewed Wikipedia to see if those topics were already covered. Table 3 shows the percentage of topics from these textbooks / syllabi that were already covered in Wikipedia.

Table 3: Wikipedia coverage of science

	Chemistry Topics in Wikipedia
Upper-level undergraduate	600 / 646 (93%)
Graduate – Masters level	64 / 136 (47%)

Because we are interested in the effect of future deepening of the scientific content on Wikipedia, we focused the experiment on the graduate level topics – which represent nearly all the opportunity for new scientific Wikipedia articles.

Within these potential articles, there were differences in the breadth of applicability: some represented a topic on their own, while others only covered a narrow aspect of a topic. We focused on broader topics since

[17] At the end of the experiment, these were also uploaded to Wikipedia to deepen the knowledge available to the public.

our journal-level analysis was also broad-based.[18] Here are some examples of the graduate articles that were identified as missing from Wikipedia and which we targeted for article creation:

Table 4: Examples of New Wikipedia Articles Created

Chemistry
• Synthesis of Hydrastine
• Multiple Michael/Aldol Reaction
• Reagent control: chiral electrophiles

PhD students with expertise in these fields then drafted articles on these topics, basing them both on their own knowledge and research conducted during writing.

In total, 43 new Chemistry articles were written for this experiment. These articles then became the "at-risk" set for being randomized into treatment and thus uploaded to Wikipedia.

As we did with the observational analysis, we can characterize the distance between these articles and articles in the scientific literature using cosine similarity. Figure 15 shows the distribution of similarities for the experimental articles.

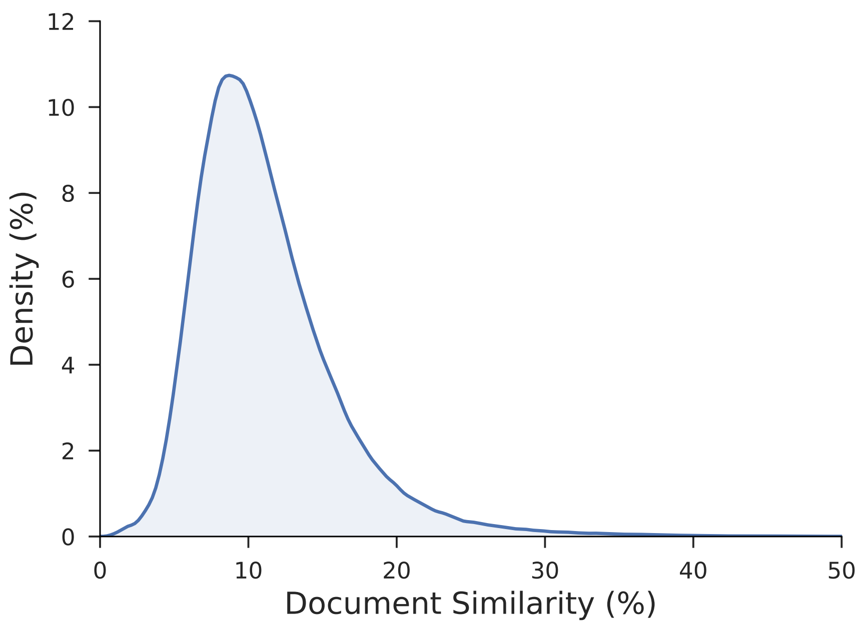

Figure 15: Density of similarity metric for all pairs of experimental Wikipedia and scientific articles

The document similarities shown for the experimental article in Figure 15 are lower those from the observational analysis shown in Figure 7. This is probably because our experimental articles are on fairly specific topics and thus will share less overlap than many of the early Wikipedia Chemistry articles on broader, introductory topics. This difference in document similarity distributions implies that while the

[18]We have no reason to believe that the effect from narrower topics would have a smaller per-article effect, but they would likely manifest across fewer scientific articles, which could make effect detection harder.

observational and experimental estimates for the effect of Wikipedia will be relevant to each other, they are not directly comparable.

7.2 Article Stratification and Randomization

To maximize the statistical power of the experiment, we stratified the at-risk set of articles with a block randomized design. We stratified on the following:

- Article Author – to control for differences in topic area / article quality / article readability
- Branch of knowledge (e.g. Organic vs. Inorganic Chemistry)
- Types of topics (e.g. general chemical principles vs specific reactions)

Within this block design we did complete randomization, assigning 50% (or the nearest integer) to treatment and 50% to control. To ensure that our randomization yielded covariate balance, we compare the following characteristics of the treatment and control groups: (1) number of words in the article, (2) number links in the article, (3) number of figures in the article, (4) number of academic references cited in the article, (5) number of non-academic references cited in the article, (6) number of web-of-science articles written on that topic (searched with title only), (7) number of google hits when searching for that topic (searched with title only).

The following tests show the balance using both a t-test (comparing differences in means) and a Kolmogorov-Smirnov test (comparing for differences in distribution):

Table 5: Covariate Balance

	Treatment (mean)	**Control (mean)**	**T-test (p-value)**	**KS-test (p-value)**
# **words**	241	243	0.47	0.16
# **links**	11.1	10.9	0.82	0.99
# **figures**	1.9	1.9	0.98	1.00
# **academic refs**	3.0	2.4	0.26	0.99
# **non-academic refs**	0.0	0.2	0.10	0.98
# **web-of-science articles**	858	1934	0.52	0.74
# **google hits (million)**	1.9	4.3	0.32	0.08*

As Table 5 shows, the covariate balance is excellent across both sets of tests, particularly on variables that pertain to the article itself (# words, # links, # academic / non-academic references). This mitigates concerns of selection effects biasing our results. Our articles lengths are also very consistent with those of an average article, ~250 words, as discussed in Section 4.1.1.

7.3 Implementation

The treatment articles were uploaded to Wikipedia in January 2015.[19] All the articles were initially uploaded as unique pages. After this point, the self-governing, open-source nature of Wikipedia became important for these articles. Based on the editors' views these articles were variously (i) accepted, (ii) rejected for rewriting (e.g. for being too technical), (iii) added as sub-sections of other pages. Rejected articles were revised in light of the editor's comments and then re-submitted.

Because the Wikipedia editor intervention happened after the randomization, it only applied to treatment articles, and thus it is impossible to establish the counter-factual effect that editor intervention would have had on the control articles. As a result, we estimate our effects as an intent-to-treat – that is, we consider the timing and article content to be that from the initial upload. We do not include any changes due to the editors or our revisions based on editor comments.

These articles (or the page that they were added to) received an enormous amount of interest, with each article averaging over 4,400 views *per month* since they were uploaded. In total, by February 2017 the pages from the experiment had accumulated over 2 million views. This makes it plausible that the causal chain of interest to us (new Wikipedia article → scientists reading the articles → effect on the scientific literature) is sufficiently strong for our treatment articles to have an impact on the scientific literature.

Data on the content of the scientific literature through November 2016 was then used to look for impacts from the treatment articles.

7.4 Outcome Measures

To interpret the experimental results, we perform the same analysis as in the observational section. In particular, we construct pre and post windows around the creation of each Wikipedia article and compare document similarity before and after.

In contrast to the observational analysis, where we needed to simulate language drift, the presence of the control group (with excellent covariate balance and only random differences) allows for much more precise measurement of the counterfactual. As such, it is possible to directly compare the results from the treatment and control groups and ascribe the difference to a *causal* effect.

8 Experimental Results

8.1 Causal Effect of Adding a Wikipedia Article

Recall that our estimator for the treatment effect is a difference in differences, comparing the similarity of articles in the scientific literature after the Wikipedia articles, as compared to beforehand (first difference),

[19] One article was uploaded earlier, in September 2014 as a pilot to test the review process.

and then comparing across treatment and control (second difference). Before showing the net effect, we present the first differences for each of treatment and control in Figure 16.

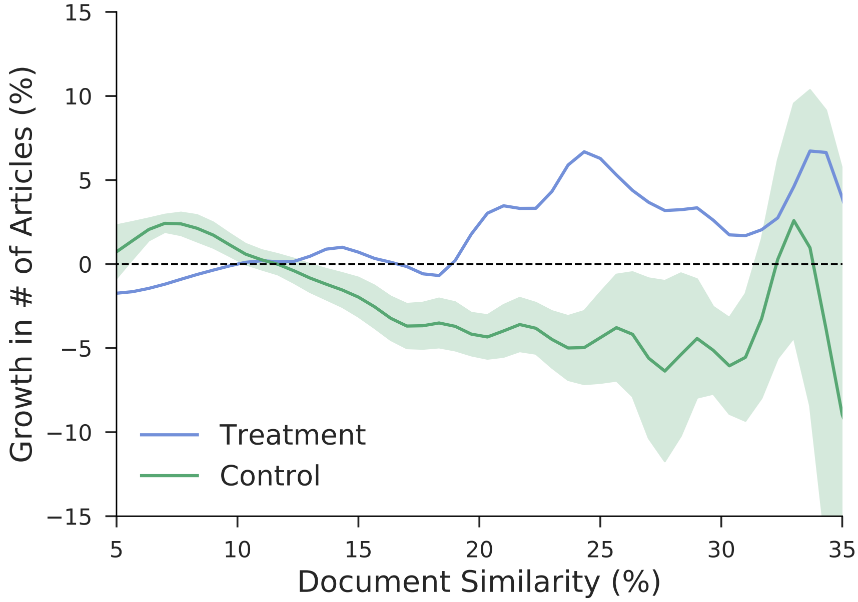

Figure 16: Treatment and Control proportional density differentials

Here we can see that there is a sizeable difference between the response of the scientific literature to the treatment and control articles. For example, the control group shows a rise in the number of low-similarity articles and a drop in the number of high-similarity ones. Since the control articles were not actually uploaded to Wikipedia, this distribution represents the baseline of change in the scientific literature – i.e. the natural evolution in the topics covered and words used in them. This is the same overall pattern we seen in in Figure 8, which showed our simulation of language drift in Science. In contrast, the treatment articles show the opposite pattern: with fewer low similarity articles and more high similarity ones. As one example, the peak at similarity ∼24% suggests an increase of ∼7% in the number of articles at that level of similarity. Thus, the divergent responses in the scientific literature to the treatment and control articles in Wikipedia already makes it clear that Wikipedia is having a causal impact.

The net effect can be seen by looking at the full difference-in-differences (i.e. by comparing to the baseline represented by the control articles). Figure 17 shows that the effect is concentrated in the high similarity region, as was true in the observational analysis. This is what we would expect, since the vast majority of articles are of low similarity, and we would not expect our articles to have broad effects across other areas of Chemistry.

Having presented our results visually, we now present them in regression framework. It is important to note that running a simple OLS regression will not be sufficient to calculate the standard errors correctly as all data are dyadic: one Wikipedia article to one scientific literature article. This implies strong correlations

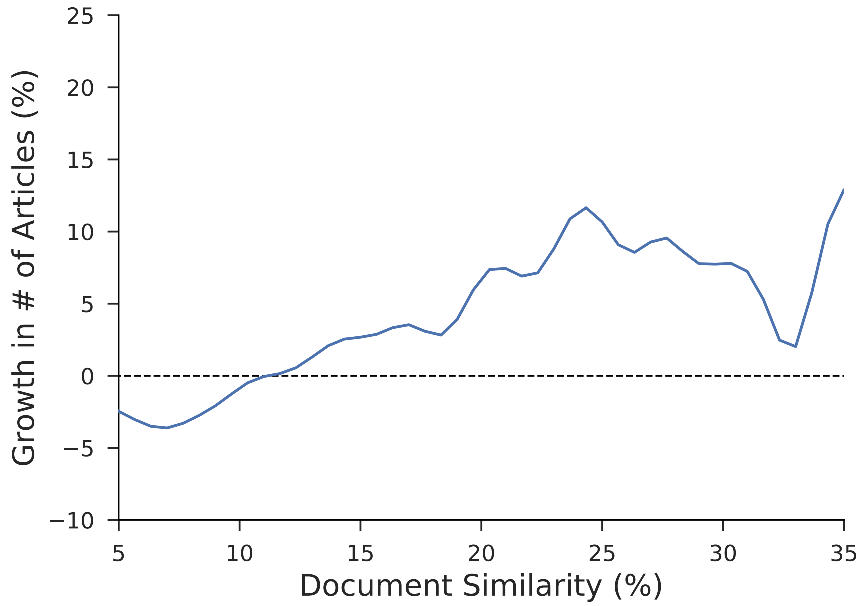

Figure 17: Effect of treatment on scientific article similarity to Wikipedia (the shaded areas indicate 1 and 2 standard errors, respectively)

between errors. To account for this, we use a two-way cluster robust estimator (Cameron and Miller, 2014) to calculate the (dyadic) standard errors for the mean effects. We bootstrap the standard errors for our quantile regressions.

Table 6: Experimental Results

	Similarity (OLS)	Similarity (q=25%)	Similarity (q=50%)	Similarity (q=75%)
Intercept	11.2404*** (0.3778)	8.0502*** (0.3456)	10.4781*** (0.3719)	13.6000*** (0.3934)
Treated	−0.1367 (0.5859)	−0.2383 (0.3982)	−0.4068 (0.4865)	−0.4743 (0.6423)
After	−0.0768*** (0.0192)	−0.0499** (0.0201)	−0.0715*** (0.0253)	−0.1103*** (0.0399)
Treated x After	0.1181*** (0.0358)	0.0804*** (0.0263)	0.1041** (0.0412)	0.1815*** (0.0604)
N	887159	887159	887159	887159
R^2	0.0001			
Adjusted R^2	0.0001			
F Statistic	28.3082			
Note:				$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

The coefficient estimate for the “Treated” indicator in Column 1 in Table 6 shows that, *prior to the* *intervention*, there was no statistically significant difference between the similarity of the scientific literature to the treatment Wikipedia articles or the control Wikipedia articles. This implies that our randomization worked. The coefficient on "After" shows a background level of language drift of 0.08pp*** towards lower similarity. Our coefficient of interest, "Treated x After," shows that adding a Wikipedia on that topic moves the later scientific literature by 0.12pp***. This effect is a change in the cosine similarity, which is difficult to interpret intuitively. However, simulations that we have done suggests that for small changes this moves almost one-for-one with the fraction of meaningful words changed in a document. That is, we find that on average the presence of a Wikipedia article changes 0.12% (or about 1 in 830) of the meaningful words in the scientific article.

As Figure 17 showed, the treatment effect is concentrated in a smaller number of highly similar articles, where the effect is substantially larger. Table 6 also shows this in regression form, with quantile regressions at the 25$^{\text{th}}$, 50$^{\text{th}}$, and 75$^{\text{th}}$ percentiles. It shows that for the lowest similarity articles the effect is 0.08pp***, about two thirds of the overall effect. This effect grows as similarity does, rising to 0.18pp*** for the 75$^{\text{th}}$ percentile.

In order to better understand the mechanism through which Wikipedia is impacting scientific articles, we repeat this same analysis but instead of considering all words in the scientific article we subset to particular subsections. In the case of Chemistry, it happens that the structure of articles is quite standardized. The vast majority follow the layout: abstract, introduction, methods, results, conclusion (or some minor variation thereof).

Table 7: Effect of Wikipedia by Article Section

	Abstract	Introduction	Methods	Results	Conclusion
Intercept	7.4550*** (0.2336)	10.3081*** (0.3888)	9.2365*** (0.3492)	10.4816*** (0.3666)	8.1863*** (0.2937)
Treated	−0.0220 (0.3658)	−0.2167 (0.5702)	−0.3550 (0.5292)	−0.0409 (0.5718)	−0.1998 (0.4206)
After	0.1604*** (0.0337)	0.0403* (0.0223)	0.0179 (0.0207)	−0.0778*** (0.0174)	−0.0772*** (0.0256)
Treated x After	−0.0306 (0.0508)	0.1100*** (0.0419)	0.0516 (0.0412)	0.1050*** (0.0342)	0.0875** (0.0347)
N	535824	828510	649646	740053	725903
R^2	0.0003	0.0004	0.0017	0.0000	0.0003
Adjusted R^2	0.0003	0.0004	0.0017	0.0000	0.0003
F Statistic	50.2205	113.8606	363.8574	8.3242	71.0333

Note: $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

The results for this analysis are displayed in Table 7. Here we can see that an effect of roughly similar magnitude to the entire article results arises for the introduction, results, and conclusion section, while there is no significant effect on the abstract or the methods section. This suggests that our Wikipedia articles are not affecting the types of experiments that are carried out, but the contextualization of them and the connections that the authors are making to the rest of the field. This latter finding is very consistent with our hypothesis that a Wikipedia article is essentially an easily-accessible review article. The lack of an effect in the Methods sections *might* indicate that scientists do not plan their experiments based on the content of Wikipedia. Alternatively, it could just indicate that there is not sufficient time for them to do so in the 3 to 9 month window after Wikipedia publication that we are examining. Put another way, if the methods planning for a paper happens more than 9 months before publication (which seems plausible for most projects), we would be unable to observe it using our analysis. Hence we interpret the lack of effect on methods as an absence of evidence, but not as evidence of an absence.

Our results indicate that Wikipedia articles *causally* affect the content of scientific articles. But is this influence positive? Revealed preference suggests "yes". Authors choosing to use the Wikipedia information are indicating that they deem the Wikipedia information to be "better" in some way. This could be because the information is of higher quality, because it is more easily accessible, or because it is easier to understand. In any case, revealed preference suggests that the scientists view it as preferable.

Notwithstanding this revealed preference argument, as a public policy matter we might want to know if the usage of the Wikipedia information improves (or harms) the quality or influence of scientific articles that use it. To evaluate this question, we compare the academic citations that accrue to two sets of articles published after our experimental intervention: one "related" to the treatment group and one "related" to the control group. Here "related" is subjective, since the effect we observe is statistical; we can't ascribe any single paper as being influenced by Wikipedia, only that on average they are more similar. We find no evidence that Wikipedia-influenced articles accrue either more or less citations that non-Wikipedia influenced ones.

We also examine if our estimated effects vary by journal quality. To do this, we disaggregate the treatment effect into four quartiles, diving the 50 journals of our sample based on the average impact factor of the journal. Re-running our analysis on these quartiles reveals no significant differences between these journal groups in terms of the size of the treatment effect.

8.2 Wikipedia points scientists to the scientific literature

We have hypothesized that Wikipedia acts like a review article, in that it discusses topics broadly and accessibly, referencing primary sources as the scientific support. In this section we explore whether the scientific articles referenced in our Wikipedia article receive more citations as a result of being randomly assigned into being posted to Wikipedia. There is significant potential for this effect, as "Wikipedia is the 6$^{\text{th}}$ highest referrer of DOI links (the unique hyperlinks assigned to academic articles)" (AOASG, 2017).

To test this effect, we look at average monthly citations in the 2 year windows before and after our Wikipedia intervention. Table 8 reports four specifications for estimating the effect on citations in this window. In specification 1 we see that scientific articles referenced in the treatment Wikipedia articles get

$\sim$21% more citations than those referenced in the control Wikipedia articles, although this difference is not statistically significant. Specification 2, rather than using a yes-no treatment indicator, uses the number of pageviews for the Wikipedia article as a measure of the intensity of the treatment. By definition, control Wikipedia articles that were not posted have zero pageviews. The estimate for specification 2 shows that getting 1% more pageviews increases citations by 0.07%*. Specification 3 adds in a control for log pre-cites - the log of the number of monthly citations that the articles were acruing before the Wikipedia article. This continues to show a positive effect from pageviews, although it again becomes statistically insignificant. Finally, specification 4 considers whether articles that were more highly-cited beforehand benefit more from Wikipedia pageviews, as measured by the interaction effect of log views and log pre-cites. This interaction effect is both positive and statistically significant. If we assume that pre-citations are a good indication of the scientific importance of a publication, then this should be interpreted as saying that the more important is a scientific finding, the greater the bump in citations it receives from being referenced in Wikipedia. This is consistent with authors becoming aware of scientific work through Wikipedia but only citing it if they deem it to be notable or of high quality.

Collectively, our results are unanimous in finding a positive citation effect, but with only some specifications being statistically significant. As such, we interpret them as providing suggestive, but not conclusive evidence, that referencing a scientific article in Wikipedia generates more citations for it.

Table 8: Effect of Wikipedia on Article Citations

	(1)	(2)	(3)	(4)
Treatment	0.2131 (0.2391)			
Log views		0.0690^{*} (0.0386)	0.0102 (0.0130)	-0.0343 (0.0217)
Log pre cites			0.9170^{***} (0.0316)	0.8236^{***} (0.0481)
Log views X Log pre cites				0.0244^{**} (0.0096)
N	107	107	107	107
R^2	0.0075	0.0296	0.8935	0.8998
Adjusted R^2	-0.0019	0.0203	0.8915	0.8968
F Statistic	0.7945	3.1978	436.3892	308.1585
Note:				$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

The interpretation of our citation findings could be positive or neutral for welfare. The positive interpretation is that Wikipedia is used to find scientific articles, which are then read and referenced. A less charitable interpretation would be that authors cite the underlying work having only read the Wikipedia article. In this second case, the citations are just a secondary indicator of the effect of the Wikipedia article, but don't represent any additional knowledge gleaned from the scientific literature.

8.3 Distributional Effects

One might imagine that public repositories of knowledge would be particularly valuable to those with less access to non-public repositories, in this case those with less access to formal journals, for example developing country scientists. However, these same scientists might also have less access to the journal articles that Wikipedia is referencing and thus might benefit less from pointers to them.

We test for the net impact of such effects by considering the differential effects of our experiment based on the GDP per capita of the modal home country of the scientific authors (assuming that those with lower GDP per capita will have less access). Our results are presented in Table 9.

Table 9: Experimental Results (Distributional Effects)

	Quartile 1 (Lowest)	Quartile 2	Quartile 3	Quartile 4 (Highest)
Intercept	10.0401*** (0.3923)	12.1339*** (0.4764)	11.2624*** (0.3741)	11.4241*** (0.3840)
Treated	0.3739 (0.5614)	−0.0793 (0.7304)	−0.1631 (0.5892)	−0.1450 (0.5924)
After	1.1193*** (0.1640)	0.0112 (0.0444)	−0.0017 (0.0215)	−0.2039*** (0.0323)
Treated x After	−0.2117 (0.2417)	0.0504 (0.0699)	0.0705* (0.0392)	0.2219*** (0.0496)
N	5381	62408	364596	413810
R^2	0.0106	0.0000	0.0002	0.0002
Adjusted R^2	0.0101	-0.0000	0.0002	0.0002
F Statistic	19.2688	0.8881	24.1986	30.9601

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Interestingly, we find, that wealthier countries benefit much more than poorer countries. This is presented graphically in Figure 18 (error bars are *one* standard error).

These results suggest that there are likely positive benefits to most developing-world scientists, but that the biggest benefits accrue to the highest-GDP countries. There are multiple plausible explanations for this finding. As suggested earlier, it could be that access to academic journals is necessary to get the full benefit of a Wikipedia article. Alternatively, it could just be that the Chemistry topics covered by our experiments were of more relevance to developed-world Chemists – recall: the topics were selected from the content at developed-world universities. We're not able to distinguish between these effects, so we just note them as potential hypotheses.

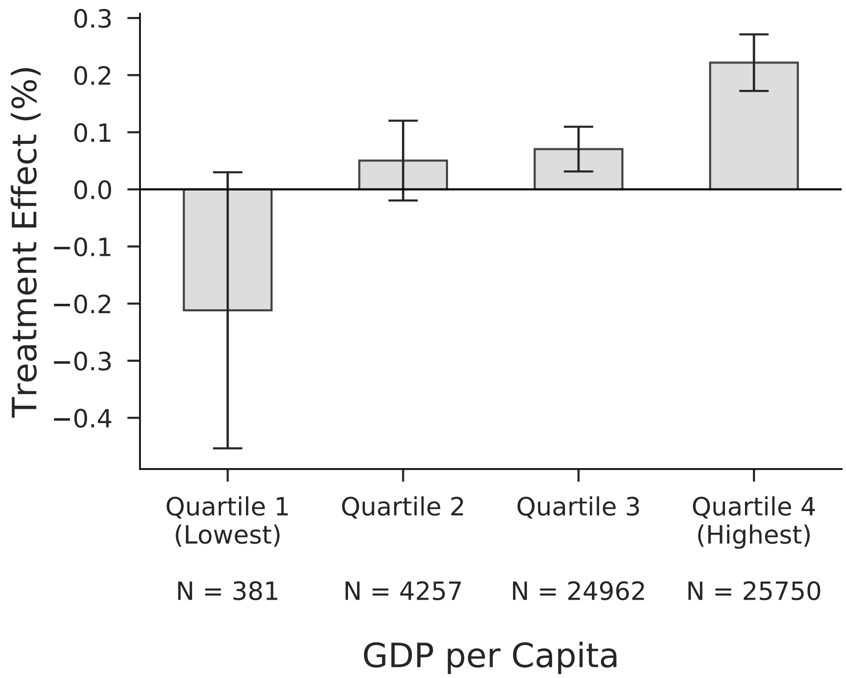

Figure 18: Effect of Wikipedia across Countries with varying access to other resources

9 Discussion

When Darwin extoled the importance "general and popular treatises," he wasn't praising their effect on popular understanding, but on "the progress of science". Our results paint a coherent image of how one of the biggest modern repositories of general and popular treatises does this. We find that Wikipedia seems to act as a collection of review articles, helping to shape how scientists contextualize their own research and pointing them to the most important scientific articles that relate to their question.

We hope that our findings on the effect of Wikipedia for the dissemination of knowledge will be sufficient to motivate scientists to undertake new initiatives to contribute articles and edits to Wikipedia. But as a society we needn't limit ourselves to individual action. Public policy interventions could encourage the development of these public resources for science. For example, a National Institutes of Health or National Science Foundation grant could require scientists to make commensurate edits to Wikipedia (or another repository). Alternatively, extra credit might be given on grant applications for those that promote science in this way. Grants could also be given directly to these scientific repositories to help with their operating costs.

Professional societies could also organize their members to develop comprehensive online repositories of knowledge, either within Wikipedia or a separate repository. A very successful existing example of this is the Stanford Encyclopedia of Philosophy. It is our hope that other groups will undertake similar initiatives.

To judge whether such interventions would be welfare-improving, it is important to investigate both the benefits and the costs. For example, if we required grantees to edit / write a Wikipedia article, this initiative would essentially be a tax on their time (or their students'). The key question is how high a tax would be needed, and whether it would be justified by the social benefit of the additional dissemination of knowledge. We consider two approaches for answering these questions, with the disclaimer that both are back-of-the-envelope calculations designed to show orders of magnitude effects. They are not intended to be precise, and they don't need to be to show the policy conclusion.

Consider first a traditional funding approach to these questions. Currently, the average NIH grant is for \$500,000 for 4.5 years, or ∼110,000 per year (U.S. Department of Health & Human Services, 2017). If one assumes that such a grant produces one paper every two years, then the approximate cost of producing one such paper is \$220,000. For our research we paid students \$100 per article. Assuming one Wikipedia article (or equivalent contribution) per research paper, the implicit tax on research would be $(\frac{\$100}{\$220{,}000}) = 0.05$. This seems like a relatively small, reasonable number, but a comparison of effectiveness is still lacking.

A second approach asks if the creation of a Wikipedia article is cost effective compared to the small fraction of grant funding that goes to promote the dissemination of knowledge. To do this cost-benefit calculation, we'll assume the cost of such dissemination is 1% of a grant (for attending conferences, copy-editing, etc.), although this assumption could be varied significantly without affecting the overall conclusion.

To calculate the benefits, we consider how the dissemination of knowledge influences later research. For this, we want to calculate a measure of how much an average scientific paper influences later ones. For simplicity, assume that this can be measured accurately by citations (although that won't actually be required for the argument). Then, if our paper generates N citations[20], we should attribute some fraction of each of those papers, s, as influence due to our paper. Thus, our measure for its impact should be $N * s$. But what is s? If we assume that there are the same number of papers giving and receiving citations, it must be at most $\frac{1}{N}$, because any number larger than that would imply that the contributions of all the papers receiving citations totalled than 100% of the value of all those giving them![21,22] And thus we conclude something which is fairly intuitive: an average paper cannot influence more than $\left(\frac{N*1}{N}\right) = 1$ other papers. We could also represent this effect in words. In our sample the average Chemistry article in our sample has ∼3,900 words, and thus 3,900 words of influence is the maximum influence that an article could have. We now have both the cost of disseminating information and the influence is has, and thus we get our (rough) estimate of the cost as $\frac{\$220{,}000*1}{3{,}900} = \$0.56$ per word. [23]

[20]According to Thomson Reuters, a typical 2000 paper in Chemistry receives ∼19 citations over a ten-year period (Times Higher Education, 2011).

[21]In actual practice the number of papers giving and receiving citations are not equal because of the growth of the number of publications over time. But this effect would not materially change our conclusions, so we omit it from our calculations for simplicity.

[22]Of course, multiple pieces of research could influence some particular set of words, but once credit was apportioned, this statement would continue to be true. Similarly, research could provide influence without receiving a citation, but such influence would need to be subtracted from other articles, so again this statement would remain true.

[23]This number could be an over- or under-estimate for many reasons. It could be an underestimate because NIH-funded work

A similar analysis can be done for the influence of the Wikipedia article. Based on the estimates from Section 8, the average change in word frequency from our intervention was ~0.12%, implying an average change of 4.7 words per article.[24] Of course, as our quantile estimates suggest, this effect is actually skewed, with a small percentage of articles changed more, and many articles being unchanged.

On average each Wikipedia article in our data set had ~13,500 "post" comparators in the scientific literature and Elsevier has approximately 16% of the academic publishing market, so we'd expect ~85,000 articles to be impacted.[25] At an average of 4.7 words changed per article, this implies that ~400,000 words changed in total. Since each Wikipedia article cost $100 to write, this implies a cost per word of $0.00025 – roughly 4,000 words per dollar – or about 1/2000$^{\text{th}}$ of the cost of dissemination through traditional methods.

Our back-of-the-envelope analysis thus has stark conclusions: even with many conservative assumptions, dissemination through Wikipedia is ~2000x more cost-effective than traditional dissemination techniques. Thus, from a public policy perspective, funding the creation of content in public repositories of science like Wikipedia is compelling. We thus encourage governments, organizations, and publically-minded individuals to incorporate the creation of such articles into their activities and applaud those who are already advocating it (e.g. Shafee, 2017).

10 Conclusion

This paper documents the contribution of public informational good providers, like Wikipedia. Using a randomized control trial we show that the creation of a Wikipedia article leads scientists to use similar words in later scientific work, which we interpret as strong evidence that Wikipedia is promoting the dissemination of science. Because our work can go beyond correlation to establish causation, we can conclude that Wikipedia doesn't just *reflect* the state of the scientific literature, it helps *shape* it.

In general, the economics of public informational goods like Wikipedia strongly favor their underprovision: incentives are too low, free-riding is rampant, and Arrow's information paradox hinders market or contractual solutions. We therefore examine the case for public policy interventions in this area. We find that the dissemination of Science through Wikipedia is highly cost-effective compared to that associated with more formal channels, such as NIH grants.

In a very concrete sense, our paper shows that Darwin was right: "general and popular treatises are almost as important for the progress of science as the original work". But we can be more precise. We

is more important than general work. The number of papers is also rising over time, which would also cause this to be an underestimate. In contrast, it could be an over-estimate because it ignores citations accruing to old work or because (presumably) new works have substantial original content. While these factors could influence the result of our calculation, it is implausible that any of these changes would impact the clear policy conclusion that follows.

[24] This number is only a rough estimate since our 0.12% effect is calculated using a term frequency-inverse document frequency (TF-IDF) estimate. However we have done simulations to confirm that there is a reasonable equivalency between the two.

[25] For an overview of the nature and scope of Elsevier publishing see: https://www.elsevier.com/connect/elsevier-publishing-a-look-at-the-numbers-and-more.

show that Wikipedia has broad influence on the way that scientists discuss and contextualize their own work. Moreover, we show that it acts as an organizer of scientific knowledge, directing researchers to the underlying literature in a way that is akin to a review article in that field.

This paper shows that Wikipedia's contribution to Science is substantial. It disseminates an enormous amount of scientific knowledge, and scientists rely on it for their research. It is our hope that, by identifying this effect, our research will spur increased investment in the development of public resources, like Wikipedia, to the benefit of scientists and society at large.

References

Alexa. 2017. *The top 500 sites on the web.* Available online at https://www.alexa.com/topsites. Accessed Aug 22, 2017.

AOASG (Australian Open Access Strategy Group). 2017. "Open access medical content and the world's largest encyclopedia" online at: https://aoasg.org.au/2017/09/05/open-access-medical-content-and-the-worlds-largest-encyclopedia/. Accessed September, 2017.

Biasi, Barbara and Moser, Petra, Effects of Copyrights on Science: Evidence from the WWII Book Republication Program (February 28, 2017). Available at SSRN: https://ssrn.com/abstract=2542879 or http://dx.doi.org/10.2139/ssrn.2542879

Cameron, Colin and Douglas Miller. 2014. Robust Inferences for Dyadic Data. *Presentation at the Winter North American Meetings of the Econometrics Society*, Boston, Jan 5, 2015.

Clarivate Analytics. 2017. *Web of Science.* Online at www.webofknowledge.com. Accessed at various times in 2017.

D&B Hoovers. 2017. *Addgene Inc. Revenue and Financial Data.* Online at http://www.hoovers.com/company-information/cs/revenue-financial.addgene_inc.b801fd5c8243ca53.html.

Elsevier. Various Years. *ConSyn database.* Accessed from 2012-2017.

Furman, Jeffrey, and Scott Stern. 2011. "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Knowledge Production" *American Economic Review* 101(5): 1933-1963

Gallus, Jana. Forthcoming. Fostering public good contributions with symbolic awards: A large-scale natural field experiment at Wikipedia. *Management Science.*

Giles, Jim. 2005. "Internet encyclopedias go head to head" *Nature*, vol 438, pp 900-901.

Greenstein, Shane, and Feng Zhu. "Is Wikipedia Biased?" American Economic Review: Papers and Proceedings 102, no. 3 (May 2012): 343–348.

Hughes, Benjamin, Indra Joshi, Hugh Lemonde, and Jonathan Wareham. 2009. "Junior physician's use of Web 2.0 for information seeking and medical education: A qualitative study" *International Journal of Medical Informatics*, Vol 78, Issue 10, October, pp 645-655.

Lightman, Bernard. 2007. *Victorian Popularizers of Science: Designing Nature for New Audiences.* University of Chicago Press.

MIT. 2017. "Citing Electronic Sources" in *Academic Integrity at MIT: A Handbook for Students.* Available online at https://integrity.mit.edu/handbook/citing-your-sources/citing-electronic-sources.

MMRRC – Mutant Mouse Resource & Research Centers. 2017. Official website, available at https://www.mmrrc.org/.

NCGRP – National Center for Genetic Resources Preservation. 2005. *Annual Report.* United States Department of Agriculture.

NIH – National Institutes of Health. 2017. National Human Genome Research Institute Official Website. Available online at https://www.genome.gov/10001772/.

Morgan, Diana. 1990. *American Type Culture Collection Seeks To Expand Research Effort.* The Scientist magazine, available online at http://www.the-scientist.com/?articles.view/articleNo/11292/title/American-Type-Culture-Collection-Seeks-To-Expand-Research-Effort/.

Princeton University. 2017. "When to Cite Sources" in *Academic Integrity at Princeton.* Available online at https://www.princeton.edu/pr/pub/integrity/pages/cite/.

Radford, Tim. 2008. *The book that changed the world.* The Guardian newspaper, Feb 8 2008. Available at https://www.theguardian.com/science/2008/feb/09/darwin.bestseller.

Shafee, Thomas. 2017. "Wikipedia-integrated publishing: A comparison of two successful models", working paper.

Shafee, Thomas, Daniel Mietchen, and Andrew I. Su. 2017. "Academics can help shape Wikipedia" *Science*, Vol 357, Issue 6351, pp. 557-558.

Shafee, Thomas, Gwinyai Masukume, Lisa Kipersztok, Diptanshu Das, Mikael Haggstrom, and James Heilman. 2016. "Evolution of Wikipedia's medical content: past, present and future" *Journal of Epidemiology & Community Health*, Published Online First: 28 August 2017. doi: 10.1136/jech-2016-208601.

Times Higher Education. 2011. "Citation averages, 2000-2010, by fields and years" *Times Higher Education*, March 31, 2011, available online at https://www.timeshighereducation.com/news/citation-averages-2000-2010-by-fields-and-years/415643.article.

U.S. Bureau of Labor Statistics. 2016. "Occupational Outlook Handbook." Online at https://www.bls.gov/ooh/.

U.S. Department of Health & Human Services. 2017. "NIH Research Portfolio Online Reporting Tools." Online at urlhttps://report.nih.gov/fundingfacts/fundingfacts.aspx.

Wikipedia. Various Years. Full edit history. https://en.wikipedia.org/wiki/Wikipedia:Database_download.

Wuchty, Stefan, Benjamin Jones, and Brian Uzzi. 2007. *The Increasing Dominance of Teams in Production of Knowledge.* Science 316, 1036 DOI: 10.1126/science.1136099.

A Econometrics Article Experiment

The Econometrics wave of the experiment was run in November 2015, with 45 articles randomized into treatment and control. Some examples of topics covered included:

Table A.1: Examples of New Wikipedia Articles Created

Econometrics
• Two-Step M-Estimators
• Smoothed Maximum Score estimation
• Truncated Normal Hurdle Model

We stratified the 45 articles by article-author, and then did complete randomization within those, yielding 50% (to the nearest integer) in treatment and 50% in control. To ensure that randomization produced covariate balance, we compared the following characteristics of the treatment and control groups:

- # words in the article
- # links in the article
- # figures in the article
- # academic references cited in the article
- # non-academic references cited in the article
- # of web-of-science articles written on that topic
- # of google hits when searching for that topic

The following tests show the balance using both a t-test (comparing differences in means) and a Kolmogorov-Smirnov test (comparing for differences in distribution):

Table A.2: Covariate Balance for Econometrics Sample

	Treatment (mean)	Control (mean)	T-test (p-value)	KS-test (p-value)
# words	439	452	0.64	0.96
# links	11.4	10.6	0.71	0.97
# academic refs	2.4	2.7	0.56	0.99
# non-academic refs	2.3	2.6	0.50	0.95
# web-of-science articles	16,782	12,582	0.82	0.99
# google hits (million)	33.6	23.9	0.55	0.51

As Table A.2 shows, the covariate balance is excellent, for both sets of articles, particularly on variables that pertain to the article itself (# words, # links, # academic / non-academic references).

Unfortunately, the econometrics articles received only ∼100 views per month, less than 3% of those received by the Chemistry articles. With so few views, only a tiny fraction of the authors of econometrics articles in the scientific literature could have viewed them, and thus we would expect our experiment to be underpowered. This will also magnify the difference between the intent-to-treat and treatment-on-the-treated estimators, since so few of the authors of the "treatment" articles would have viewed the Wikipedia articles. Thus our estimates should also be closer to zero.

This is indeed what we see, with the large error bars reflecting our lack of power

A similar story can be seen in the regression analysis, though we do see that for the lowest quantile, there is a significant effect.

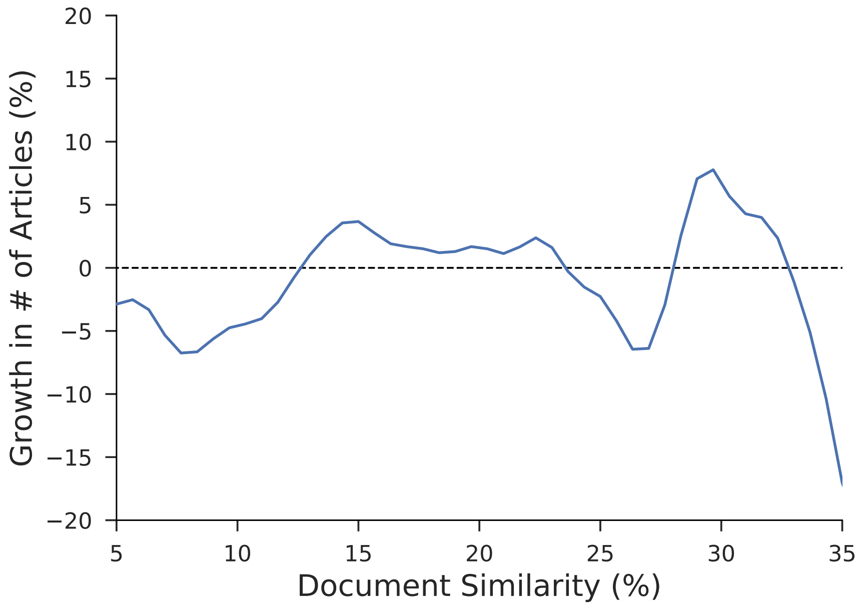

Figure A.1: Treatment Effect Estimates for Econometrics

	Similarity (OLS)	Similarity (q=25%)	Similarity (q=50%)	Similarity (q=75%)
Intercept	15.7633*** (0.3480)	11.8991*** (0.3392)	14.8131*** (0.3558)	18.5073*** (0.3765)
Treated	0.2633 (0.5148)	0.1425 (0.4822)	0.2797 (0.5163)	0.4245 (0.5400)
After	−0.2774*** (0.0319)	−0.1805*** (0.0388)	−0.1065*** (0.0290)	−0.1601*** (0.0367)
Treated x After	0.0696 (0.0524)	0.1374*** (0.0501)	0.0638 (0.0516)	0.0527 (0.0822)
N	152547	152547	152547	152547
R^2	0.0010			
Adjusted R^2	0.0010			
F Statistic	53.1126			
Note:				*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Figure A.2: Experimental Regression Results for Econometrics

B Word Frequency Approach to the Observational Chemistry Analysis

Instead of analyzing our Chemistry data at the document level, we could look for frequency changes at the word level. This section describes these results.

B.1 Word Frequency

The main object of interest in the text data is the evolution of the usage frequency of various words. In particular, since we seek to uncover the relationship between Wikipedia and the scientific literature, we analyze the parallel evolution of these frequencies in the respective corpora.

In order to make the frequency series arising from these two corpora comparable, we focus on contemporaneous activity. That is, on the scientific literature side, we make the natural choice of looking at the stream of published articles, and for the Wikipedia side, we consider the stream of new words entering the encyclopedia through edits, rather than the state of the text at a particular point in time.[26] Figure A.3 shows schematically the hypothesized effect that a Wikipedia article could have on shaping the scientific literature.[27] Not pictured, but also important, would be effects that reverse the causality or come from common causes (such as development in science).

Let $f_{i,t}^{wiki}$ and $f_{i,t}^{sci}$ denote the relative log frequencies for word (token) i at time t in the Wikipedia and science corpora.[28] Throughout the analysis, we use relative word frequency to denote the absolute frequency of the word divided by the total frequency in the entire Wikipedia corpus in that period.

Further, for $k \in \{wiki, sci\}$, define adjacent frequency differences as

$$\Delta f_{i,t}^{k} = f_{i,t}^{k} - f_{i,t-1}^{k}$$

Similarly, let the indicator $I(f_{i,t}^{k} > 0)$ denote whether token i appears in corpus k at time t.

We focus on a simple model of word frequencies, where Wikipedia increases total usage of the word above what would be seen absent it.[29] Here word frequencies follow the difference equation

$$f_{i,t+1}^{sci} = f_{i,t}^{sci} + \alpha(\overline{f}_{i}^{sci} - f_{i,t}^{sci}) + \beta f_{i,t}^{wiki} \tag{1}$$

where $\overline{f}_{i}^{sci}$ is the frequency of that token at its "natural potential," that is, the prevalence that it would achieve over the long-run (absent Wikipedia).

[26] Note that this means that the cumulative added word counts will not always exactly reflect the current size of an article as words can be deleted as part of editing, but this more accurately tracks activity in all its forms.

[27] The frequency of terms in science is assumed to be growing because Wikipedia articles are often written about emerging areas.

[28] Henceforth we will often use this term, from the natural language processing field, to refer to the term we are searching for.

[29] One could also imagine more complicated models where Wikipedia just accelerates progress towards the long-term steady-state. Reliably distinguishing such models would likely take longer observational data (to verify the reaching of the equilibria) and thus we do not consider them here.

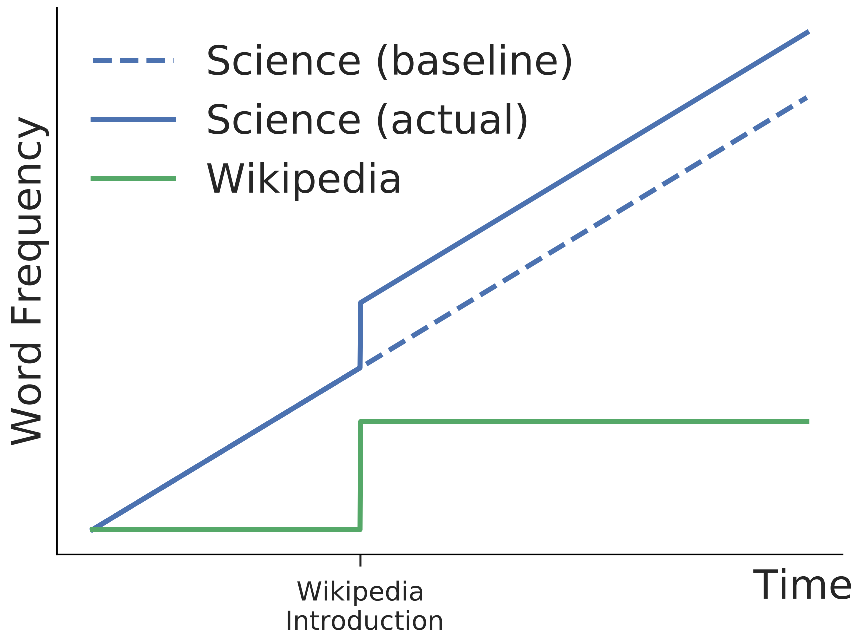

Figure A.3: Schematic of the how Wikipedia might shape science

Differencing the specification given in Equation 1, we find

$$\Delta f^{sci}_{i,t+1} = (1-\alpha)\Delta f^{sci}_{i,t} + \beta \Delta f^{wiki}_{i,t}$$

Thus our prime objects of interest in a regression model will be the pure slope persistence of science word frequencies $1-\alpha$ and the effect of Wikipedia occurrence β.

Motivated by the above model, we perform a number of regression analyses on the word frequency time series, both in levels and in first differences. In addition, we control for levels when looking at first differences to allow for potential deviations from this simple model in how the word diffusion and adoption process works.

B.2 Results

We find strong support for linkages between the evolution of word frequencies in Wikipedia and the scientific literature. We focus on the field of chemistry so as to make these results relatable to the observational and experimental analyses at the document level. Using the notation from above, we estimate the following equation

$$f^{sci}_{i,t+1} \sim f^{sci}_{i,t} + f^{wiki}_{i,t}$$

As Table A.3 shows, in levels, that a 100% increase in Wikipedia frequency is associated with a 25% increase in science frequency.

A more nuanced analysis recognizes that word frequencies often follow adoption curve like dynamics. Hence the importance of Wikipedia may not be so much on the particular level but on the rate at which

Table A.3: Levels on levels regression of science on Wikipedia

	Log Science Frequency (t+1)
Intercept	-0.4862^{***} (0.0271)
Log Wikipedia Frequency (t)	0.2536^{***} (0.0016)
Log Science Frequency (t)	0.7354^{***} (0.0010)
N	446167
R^2	0.7233
Adjusted R^2	0.7233
F Statistic	583169.3028
Note:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

new ideas are spread, as proxied by their word usage. We thus estimate the following equation.[30]

$$\Delta f_{i,t+1}^{sci} \sim \Delta f_{i,t}^{sci} + f_{i,t}^{sci} + f_{i,t}^{wiki}$$

Of particular interest is the coefficient on the Wikipedia frequency level. Table A.4 summarizes the regression results. Here we can see that a 100% increase in Wikipedia frequency is associated with a 17% increase in the growth rate of the science frequency (recall that all frequencies are in logs).

Since all regressions in log frequencies must condition on positivity, they only pick up the intensive margin of usage. To see the extensive margin, we also run a binary regression on frequency positivity, i.e. whether the word was used in the corpora in that time period.

Not surprisingly, we see that the presence of words in each corpus is also correlated.

[30] Importantly, although it has an intuitive appeal, this equation is NOT unbiased because any idiosyncratic error associated with $f_{i,t}^{sci}$ is present on both sides of the regression, and thus will induce some correlation. However, because of the signs of the terms, this correlation works *against* us observing an effect.

Table A.4: Differences in differences and levels regression of science on Wikipedia

	Δ Log Science Frequency (t+1)
Intercept	-0.1133^{***}
	(0.0257)
Log Wikipedia Frequency (t)	0.1685^{***}
	(0.0015)
Log Science Frequency (t)	-0.1722^{***}
	(0.0010)
Δ Log Science Frequency (t)	-0.3651^{***}
	(0.0014)
N	415641
R^2	0.2676
Adjusted R^2	0.2676
F Statistic	50610.2688
Note:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table A.5: Regression of the existence of tokens in science on Wikipedia

	Science Frequency > 0 (t+1)
Intercept	0.1922^{***}
	(0.0003)
Wikipedia Frequency > 0 (t)	0.2255^{***}
	(0.0006)
Science Frequency > 0 (t)	0.4276^{***}
	(0.0005)
N	2899125
R^2	0.2786
Adjusted R^2	0.2786
F Statistic	559757.5848
Note:	$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$