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

Neil C. Thompson
MIT
neil_t@mit.edu

Douglas Hanley
University of Pittsburgh
doughanley@pitt.edu

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. In the months after uploading it, an average new Wikipedia article in Chemistry is read tens of thousands of times and causes changes to hundreds of related scientific journal articles. Patterns in these changes suggest that Wikipedia articles are used as review articles, summarizing an area of science and highlighting the research contributions to it. Consistent with this reference article view, we find causal evidence that when scientific articles are added as references to Wikipedia, those articles accrue more academic citations.

Our findings speak not only to the influence of Wikipedia, but more broadly to the influence of repositories of knowledge and the role that they play in science.

JEL Codes: O31, O33, O32

*The authors would like to thank Caroline Fry for excellent research assistance. They would also 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, and Dario Taraborelli at the Wikimedia Foundation for guidance.

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" are more than just summaries of science for the general public. Do popular treatises also influence researchers and their contributions to science? 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 by markets. They are thus good candidates for welfare-improving interventions by governments, organizations, and public-spirited individuals.

Given the potential for welfare improvement, it is heartening that many public scientific repositories exist. For example, governments fund repositories of physical objects, like seed banks (NCGRP 2005) and model organism repositories (MMRRC 2017). Governments also fund informational repositories, for example for the human genome (NIH 2017). Individuals and other organizations provide informational repositories as well. For example, StackOverflow.com is a widely used question-and-answer repository for knowledge about computer programming. Previous research on the effect of scientific repositories has shown that they can promote scientific activity (Furman and Stern, 2011).

Still, many important areas of scientific knowledge are not in public repositories. In particular, the most extensive repositories of scientific knowledge – academic journals – remain overwhelmingly restricted to those paying subscription fees. And even those with subscriptions may have difficulty understanding the information that journals contain because of technical jargon, poor writing, etc. But what if the key insights from journal articles were written up in easy-to-read articles available in a convenient public repository?

Wikipedia is one of the largest informational public goods providers in all of science. It is freely available, easily accessible, and is the 5th most visited website in the world (Alexa 2017). To put this

in context, it has 500 million unique visitors per month (Cohen, 2014). So, a significant fraction of humanity is using Wikipedia.

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 — or one article for every ∼120 scientific journal articles, as measured by listings in Web of Science (Wikipedia, 2019). 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 new 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 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 an excerpt from MIT citation guidelines makes 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

[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

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...you do not need to cite a source." It is quite plausible that a researcher, finding that a fact is 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 semantic 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 5,000 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 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 because it cannot rule out plausible alternatives, such as mutual causation. In this case, mutual causation would be when a scientific article (perhaps a major breakthrough) generates both a Wikipedia article as well as follow-on articles in the scientific literature. This would induce correlation between the words used in Wikipedia and the follow-on articles, but it would *not* indicate that it is Wikipedia that is shaping them. Since it seems obvious that this mutual causation must be going on, the interesting question is whether there is an additional impact that is *caused* by Wikipedia.

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 "treatment" articles should have a bigger effect on the scientific literature than the "control" articles. We find exactly that: the scientific content from the articles we upload to Wikipedia makes its way into the scientific literature more than the content from control articles that we don't upload...and these effects are large.

Wikipedia's influence is strongest for researchers in the 25% richest countries, although we also see effects in middle income countries. We also find that when a reference to a scientific article is

[2] *Both* sets of articles need to be written because the analysis is lexical and thus the wording of the control articles matters.

added to Wikipedia it generates more citations for that article. Taken together, these two findings suggest that Wikipedia readers that have access to journal articles (which is more likely in richer countries) use it as a guide to the scientific literature. The finding that adding a scientific reference to Wikipedia generates more citations is also a *causal* one because it also comes from an experiment (implicitly in our experiment adding articles to Wikipedia, we also have an experiment of adding references).

The finding that Wikipedia articles are not only summarizing the literature, but highlighting important underlying research is reminiscent of what one would expect from a scientific review article. This is consistent with the authors' experiences of Wikipedia and so we test this empirically. We find that the two are broadly similar. The addition of a Wikipedia article produces an effect that resembles that of publishing a review article, only weaker. Despite having articles that exert a weaker influence, Wikipedia has so many articles that it is (to the best of our knowledge) the *second largest repository of review articles in the world* — as measured either by the number of articles or by the influence that they are having on Science.

Finally, we consider public policy implications. We find that Wikipedia is an enormously cost-effective way to disseminate science.

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 insufficient to achieve welfare-maximizing outcomes because they fail to capture the spillover benefits to others. Under these conditions, and absent intervention by governments, organizations, or public-spirited individuals, there are fewer public goods than would be socially optimal.

A common way of resolving public goods problems is to make the resource 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, leading to better incentives for private provision at the cost of excluding some users from the market. But excluding users from *information goods* can carry substantial welfare losses if there is a "long-tail" of potential users whose individual value for the good is not high enough to justify paying the price charged, but who collectively represent a substantial fraction of total welfare benefit.

The challenge for informational public goods for the scientific literature is, however, worse than

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

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 prevent potential consumers from reading an article before purchase. But, being unable to read the articles, those consumers might find it hard to decide whether or not to purchase. This combination of a long tail and uncertain value can magnify the welfare loss. Aguiar and Waldfogel (2018) coined the term "random long-tail" for this phenomenon. They find that in entertainment, another information good context, uncertainty makes the welfare impact 10 times larger than with a deterministic long-tail.[4] This suggests that, in a closed-access model, the foregone welfare from researchers being unable to find the most-relevant article could be substantial.

The model of information sharing embodied in Wikipedia is a middle ground between the extremes of open-access and closed-access science. It provides a free, widely-accessible summary of the scientific findings with links to the underlying papers. At the same time, it omits the detailed empirics and derivations found in scientific articles. This mix of characteristics is reminiscent of review articles from the scientific literature. Because of this similarity, we hypothesize that the effect of adding a Wikipedia article may be similar to the effect of publishing an easily-accessible review article.

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.[5]

As of 2017, Wikipedia has 5.3 million articles. These were written and are edited by 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),

[4]Note: In their model, uncertainty comes from not knowing the quality of music before it is created. In our analogy, the uncertainty comes from unknown heterogeneous preferences over a particular good.

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

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.

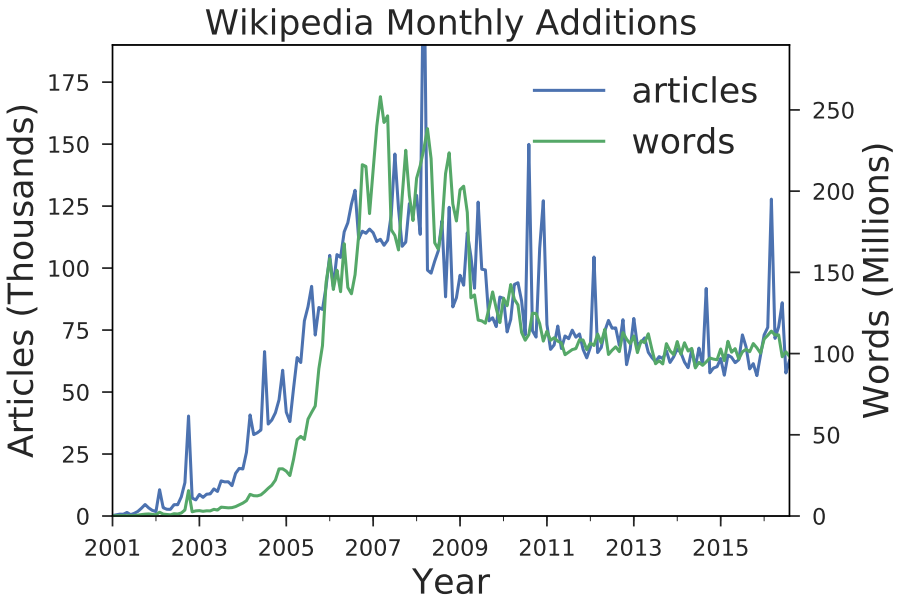

Figure 1: Words and articles added to Wikipedia since its inception

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 editing community actively brings in scientific information from academic journals (Jemielniak et al., 2019). They also enforce 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 erred at similar rates (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 is $\sim$ 45-80$\times$ bigger, with about 5.3 million articles totalling 1.8 billion words.

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 (0.5-1.0 million out of about 5 million).[6] As we discuss in more detail in section 8, these articles cover most undergraduate-level topics as well as some graduate-level topics. There exists substantial interest in the open-source community for continuing to deepen the scientific knowledge on Wikipedia (Shafee et al., 2017).

Wikipedia is widely read. As of 2014, it served a total of 18 billion page views to 500 million unique visitors each month. Wikipedia is 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). These results echo earlier, pre-internet findings that 60% of Wisconsin medical school faculty and 89% of North Carolina physicians had "learned of new scientific developments through the mass media" (Phillips et al., 1991).

Previous empirical work shows that exposure to particular pieces of research influences the work that scientists do and the citations that they make. Research by Biasi and Moser (2017) shows that making German textbooks more accessible to American researchers increased the number of new scientific articles, books, patents and PhDs in those areas. Work by Phillips et al. (1991) demonstrates the attention that mass media can bring to scientific articles by using a period when the New York Times went on strike as a natural experiment. During this period, Times science journalists still wrote up, *but did not publish*, the findings of new discoveries from the New England Journal of Medicine (NEJM). Phillips et al. find that the incremental effect of the New York Times publishing an article (not just selecting it) produces significantly more citations for the underlying research paper, including a 72.8% increase in first year citations.

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 lexical measure, where we can measure this effect directly in the words used by scientists.

[6] To determine which articles are considered part of each scientific field, 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 articles and using them to train a text-based Support Vector Machine classifier.

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 from 5,215 Elsevier journals, which we use to represent the state of the scientific literature. The third is data on citations to academic journal articles, 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 that are seldom seen by readers, which we also exclude. These include 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; redirect pages, which allow for multiple name variants of a single source page; pages associated with hosted media files such as images, audio, and video; and much more.

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 (or an IP address in the case of anonymous edits), 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[7] 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

[7]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 2^{10}=1024 gibibytes per tebibyte (a small change from the base 10 version of 1000 gigabytes per terabyte).

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 ("metrics").

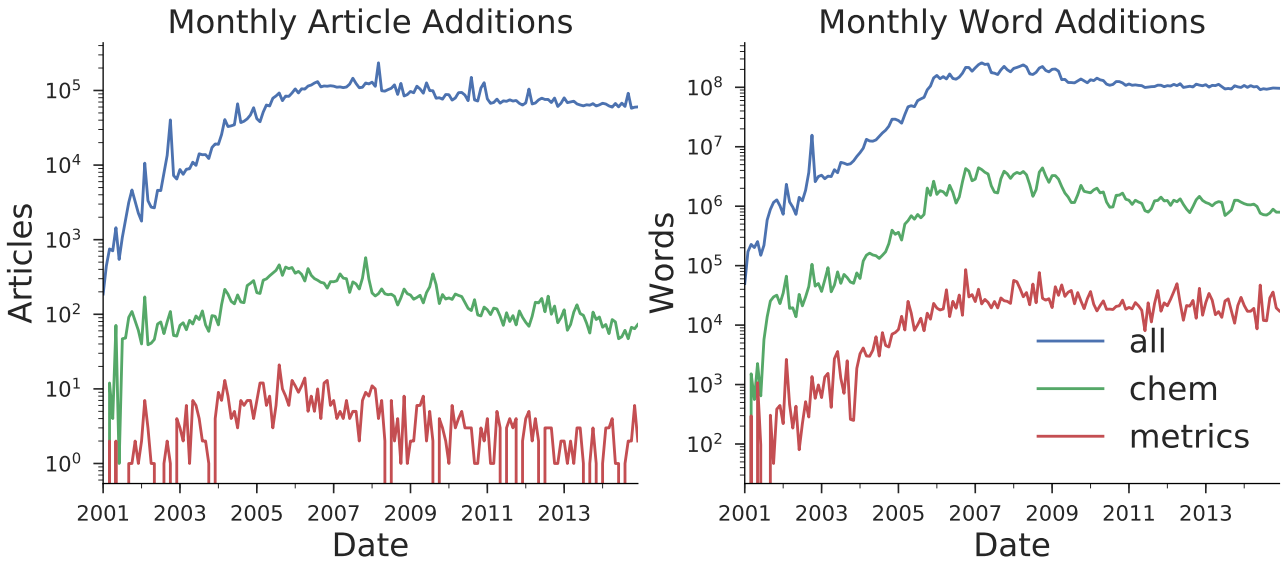

Figure 2: Monthly article and word additions to Wikipedia

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. These differences reflect the overall number of chemists and econometricians in society (authors' calculations).

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 articles begin as a "stub" — a short article, perhaps just a title and a 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.

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.

"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

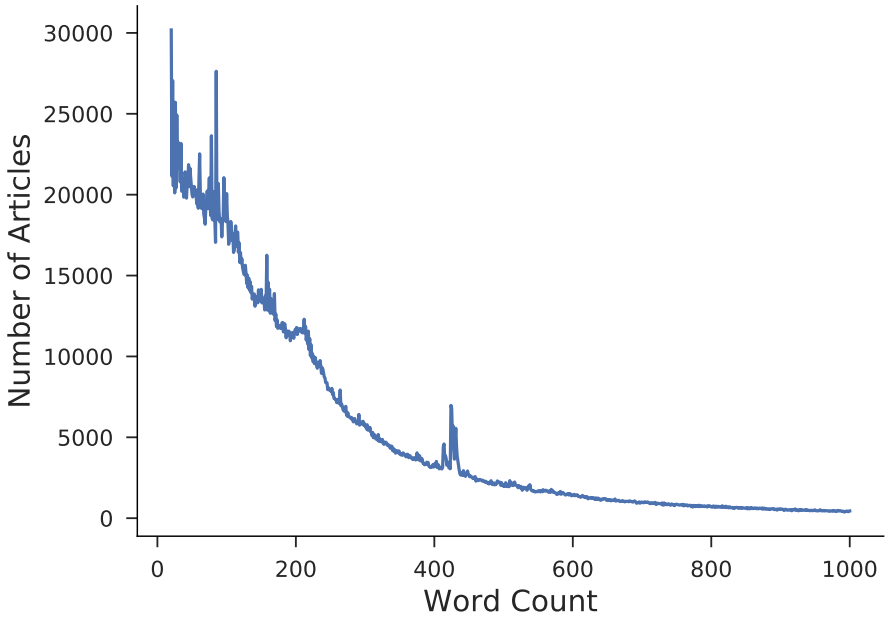

Figure 4: Size distribution of new articles longer than 20 words

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 earlier articles — again suggesting that these were on broader, more important topics.

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.

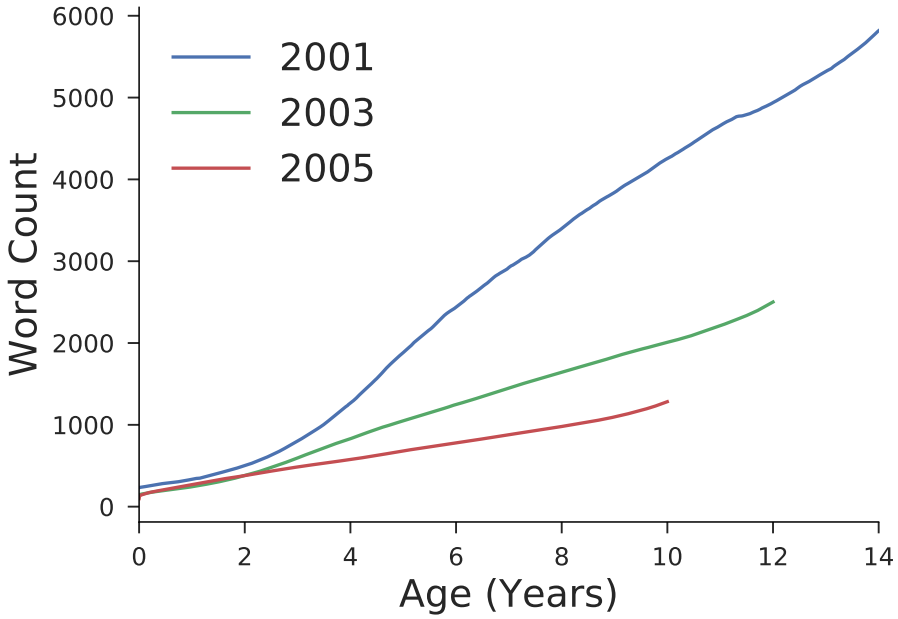

Figure 5: Average size of articles conditional on age (daily)

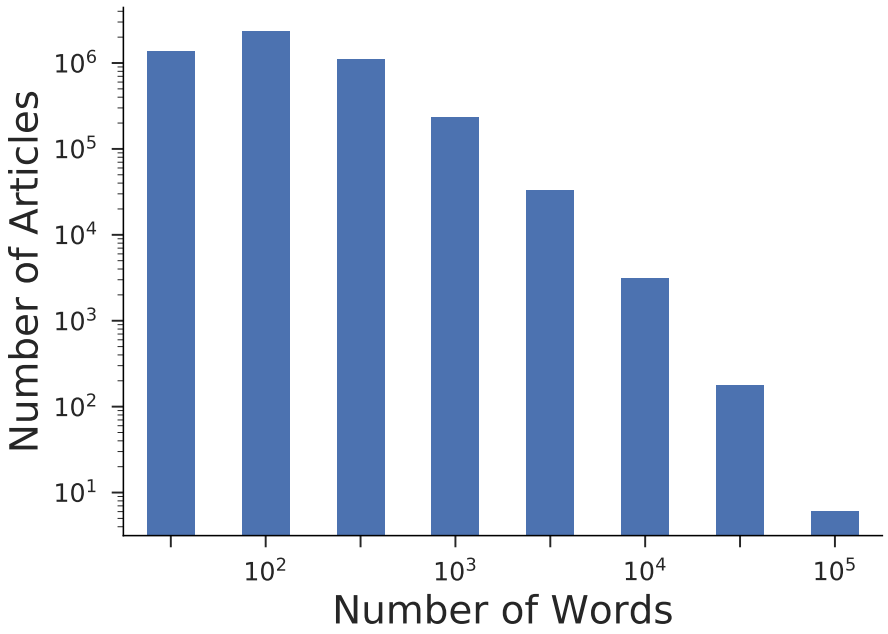

Figure 6: Distribution of article size for articles larger than 30 words

4.1.2 Word Coverage

The entire scientific literature corpus for Chemistry contains roughly 9.3M unique words. For our word-level analysis, we focus on those words that appear in more than 10 distinct documents while also not appearing in more than 90% of documents. This reduces the size of our vocabulary down

to ∼550K words. This pruning 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. Furthmore, by setting a maximum document frequency, we ignore effects coming from extremely common connective words such as "the" and "and." Even after culling these words from our vocabulary, our data still accounts for 78% of word usage in science and 79% of word usage in the Chemistry pages of Wikipedia. Our remaining set of ∼550K words still includes relatively common words such as "were" and "also," which contain little information. We address this in subsequent analysis using inverse document frequency weighting to ensure that our results are not driven by these words (as described in more detail below).

The vocabularies used in the scientific literature and in Wikipedia have substantial overlap. Of the words present in Wikipedia, 95% are represented in the scientific literature. In the other direction, roughly 55% of words in the scientific literature are seen in Wikipedia, which highlights that the scientific literature covers a more diverse array of topics. These word overlaps are already substantial, but weighting those comparisons by word usage frequency raises both to $\sim$ 99%. The following provides some context for the relative frequency of the words in our data:

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

Word	**Literature Rank**	**Wikipedia Rank**
Acid	29	88
Reaction	20	164
Graphene	1,618	4,983
Photovoltaic	6,186	8,172
Gravity	6,409	2,067

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.[8] 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 subcate-

[8] Surprisingly, this graph contains cycles. However, it can be trimmed 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 in which a sufficiently low PageRank node is the parent of a higher PageRank node. For our implementation we use a parent PageRank threshhold of 60% of the child's PageRank value, though the resulting category tree is not that sensitive to the exact cutoff value used. This eliminates only 1% of edges and renders the graph acyclic (a tree).

gories, then find all pages belonging to such categories.

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

To correct for these 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 a 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 ones.

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. Since we are interested in the interaction of the scientific literature with Wikipedia, we use only data from the year 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 (e.g. figures or charts) or 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 information 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 the technical content in Wikipedia and the technical content 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 7 and 8 we return to the question of causality.

5.1 Semantic Similarity

To evaluate whether content in the scientific literature is similar to that in Wikipedia we use cosine similarity, a "Vector Space Model" where the words from each document are used to form a vector and then the two vectors are compared. The elements in our vectors are individual words. Some other analyses have extended these to, for example, two-word phrases ('bi-grams'), but such models are not tractible with data as large as ours. Even if such models were tractible, however, other work shows that they would be unlikely to improve performance (discussed below).

Vector space models are widely used to analyze the semantic similarities and differences between documents. For example, these models are used in search engines, document retrieval, document clustering, document classification, essay grading, document segmentation, question answering and call routing (Turney and Pantel, 2010). Particularly relevant for our context, previous work shows that cosine similarity works well for analyzing technical and scientific discoveries. In particular, Younge and Kuhn (2016) analyze how inventors describe the scientific or technical underpinnings of their invention in patent specifications. They show that measuring cosine similarity is highly predictive of the human evaluations of invention similarity by (a) patent attorneys, (b) crowd-sourced workers, (c) inventors in that technical area, and (d) patent classifiers. Thus, for documents in general and for the comparison of scientific content in particular, there is strong evidence that that analysis of words via a cosine similarity measure reflects differences in the underlying semantic content of the documents.

5.1.1 Word Co-occurrence in Documents

To analyze content similarity, we take advantage of their arrangement into documents in both corpora (Wikipedia and the scientific literature). 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 fraction of words are represented in any given document.[9]

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

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

where $\|v_i\| = \sqrt{v_i \cdot v_i}$. This satisfies the natural properties that: (1) $d(v, v) = 1$, (2) $d(v_1, v_2) = 0$ when v_1 and v_2 have non-overlapping bases, and (3) $d(v_1, v_2) \in [0, 1]$ because, as in our setting, word frequency vectors are non-negative.

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 weights given to rarer (and presumably more important) words. In particular, this scheme weights tokens by the log of the inverse of the fraction of all documents that the token appears in. This is a standard metric used in text analysis problems. Previous work has shown that cosine measures with tf-idf perform well for comparisons of similarly technical and scientific work (in patents) (Shahmirzadi et al, 2018). They further show that for our type of data, cosine similarity with tf-idf performs comparably, and sometimes better, than other modeling techniques (e.g. embedded models, such as topic models or neural models) or other variants on our approach (e.g. including longer phrases, a.k.a. "n-grams").

Most articles in the scientific literature have some similarity to any given Wikipedia article, but not that much. Figure 7 shows this empirically, plotting the average similarity between all pairs of Wikipedia and scientific articles in our Chemistry sample. Virtually all scientific article-Wikipedia pairs have similarities between 0% and 10%.

Estimating the effect of Wikipedia on Science from our observational data requires two elements: an observed change in correlation between Wikipedia and the scientific literature before and after publication, and a counterfactual about how content would have changed absent the Wikipedia article. We consider the latter first.

5.1.2 Content Drift

There is a natural ebb and flow to the content in scientific writing over time. This can be due to the advent of genuinely new concepts or discoveries, such as the "CRISPR" gene-editing technique in biology, or the natural shifting of the attention of the scientific community from one topic to another. This content drift means that, as time passes, the literature will become less and less similar to what

[9]Some approaches reduce words with the same root (such as the singular and plural form of a certain word) to a single representative word. This process is known as destemming. We don't employ this technique in our case due to concerns about obscuring what are sometimes critical differences in chemical terminology that are expressed using suffixes.

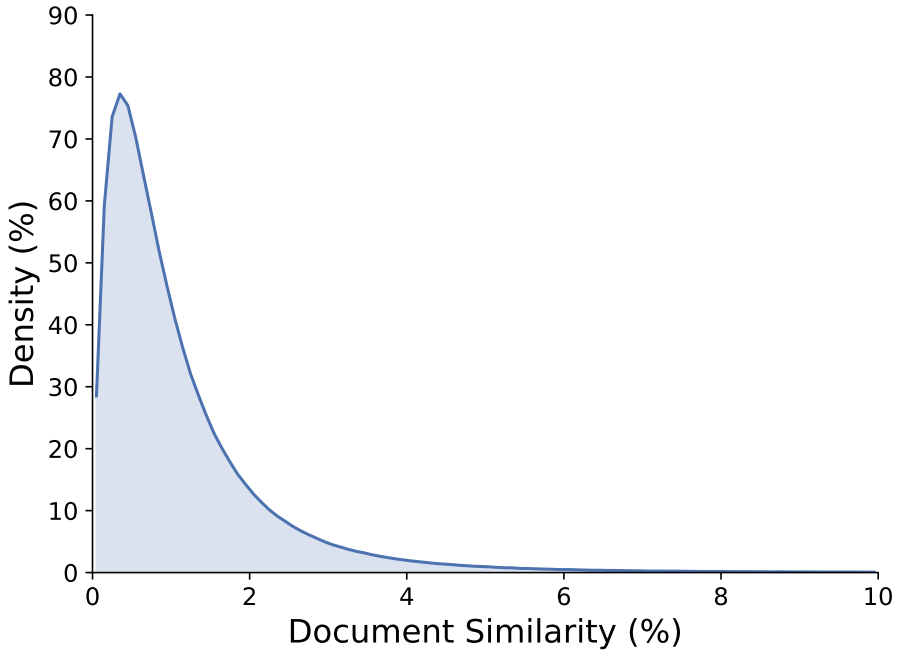

Figure 7: Density of similarity between Wikipedia and scientific articles (all pairs shown)

came before it. Content drift is thus the counterfactual for our analysis of what the literature would have done absent the addition of our Wikipedia articles. Figure 8 illustrates how this affects similarities: decreasing the number of high-similarity document pairs and thereby increasing the number of low-similarity ones.[10] In section 8.1 we show empirically that this is the correct characterization content drift.

Because the historical creation of Wikipedia articles was not random, our observational analysis lacks a natural counterfactual group showing the extent of content drift. Our experiment, however, solves this problem by having a randomized control group.[11]

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 $Similarity_{ijt}$. This notation can include any Wikipedia-science article pair, even those where the scientific article was published before the Wikipedia article. Thus, let us also denote by $After_{ijt}$ the binary indicator of whether scientific article j was written after Wikipedia article i.

[10] This version of drift tracked word usage frequencies in science over time by grouping words into finely spaced frequency bins and calculating their Markov transition matrix from f_t to f_{t+1}.

[11] The experimental results are useful, but not perfect, as a counterfactual for the observational analysis since there are differences between the two sets of articles (e.g. on average our experimental sample has more advanced Chemistry than the observational sample).

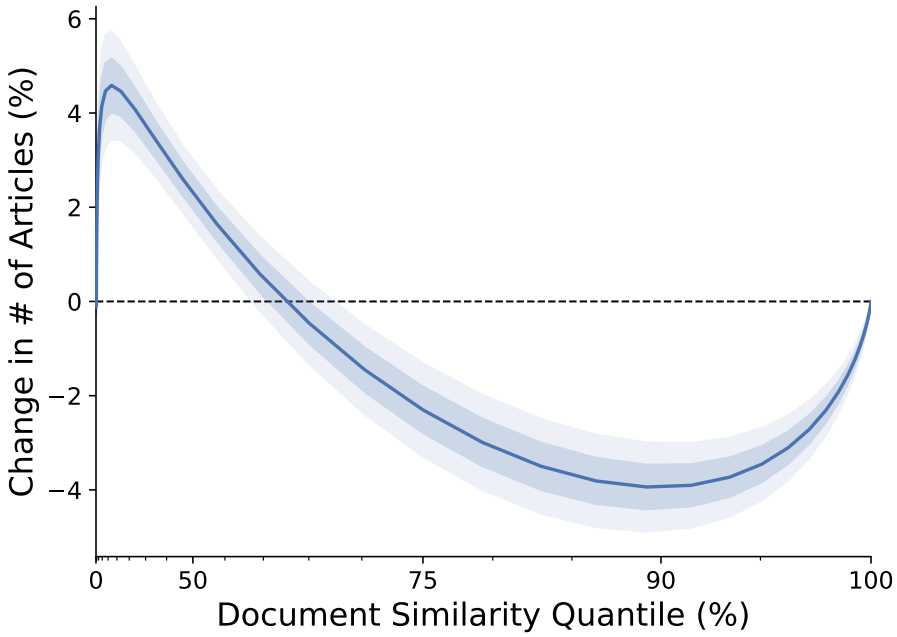

Figure 8: Effect of content drift on document similarity

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

$$Similarity_{ijt} \sim \alpha + \tau * After_{ijt}$$

This is essentially a difference in means that compares document similarity before and after the Wikipedia article is created. By adding this estimate of the raw treatment effect, τ, to our synthetic counterfactual of natural content drift in science, δ, we get our observational estimator for the net effect 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 be more similar to the article than was the science published beforehand.

We allow for a three month lag after a Wikipedia article is first "created", and use the wording at the end of those three months as the "new" article language. This approach reflects a common article-creation process 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.[12]

We look for effects of the creation of a Wikipedia article on science in two time windows: one six-month window preceding it and one six-month window after it. Implicitly, the aforementioned three month delay also accounts for publication lags in science. Importantly, the publication lags in Chemistry are much shorter than those in Economics. For example, the average time that an article is with a Chemistry journal at Elsevier, from when it is first submitted to when it is published online is 10.6 weeks (author calculations based on Elsevier, 2019). Figure 9 explains our measurement timeline segmentation:

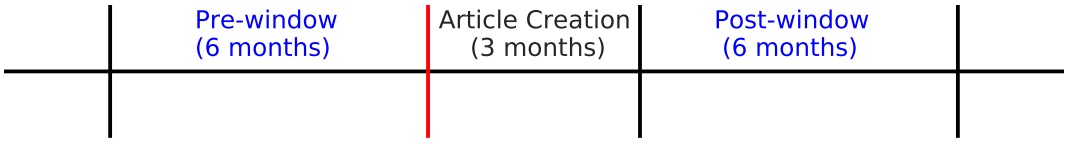

Figure 9: Measurement timeline

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 in the content between Wikipedia and the scientific articles. In other words, the content in Wikipedia and the scientific articles has become more similar.

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, showing the strong relationship between the relative frequencies of words in the two corpora, but also the variance around it.

[12] An example of a newly-created article with almost no content can be seen at https://en.wikipedia.org/w/index.php?title=Paracamelus&oldid=750127001, which displays a version from November 2016.

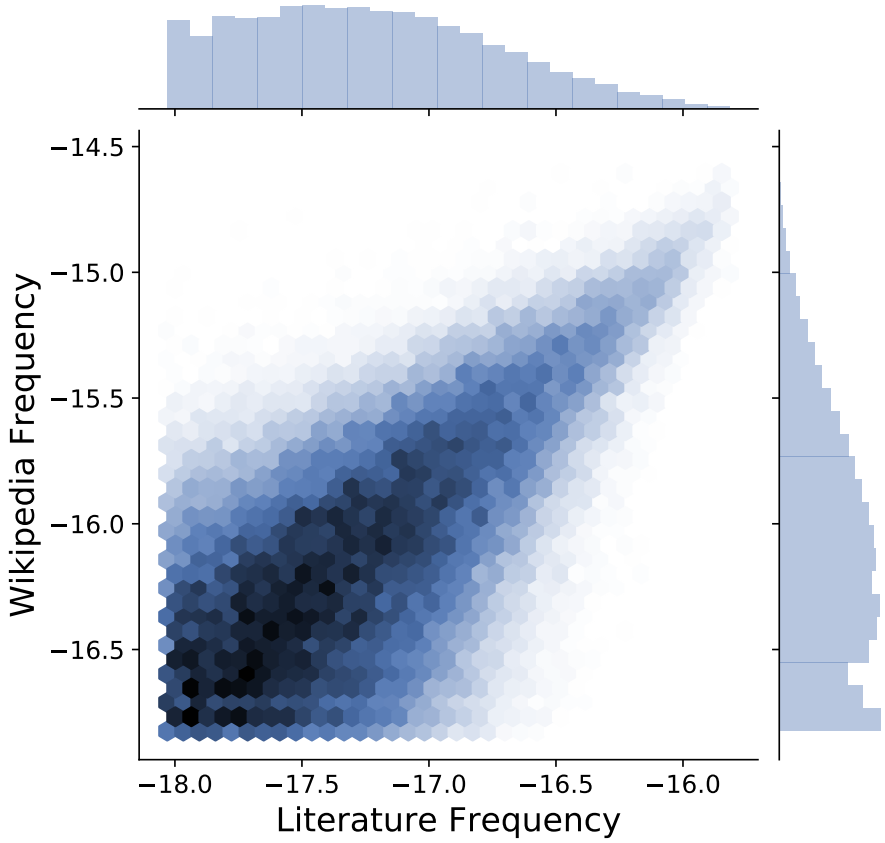

Figure 10: Token frequencies in Wikipedia and Science (above median tokens)

6.2 Event studies

Figure 11 shows some examples of the frequency of words being added to science (new articles) and to Wikipedia (edits). 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 correlations in usage between Wikipedia and Science, but others like "reaction" whose trends seem mostly unrelated.

Naturally, not all words appear in Wikipedia at founding. Most tokens start to be used between 2005 and 2007. In terms of trends, it is quite common to see a large amount of editing activity near 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.

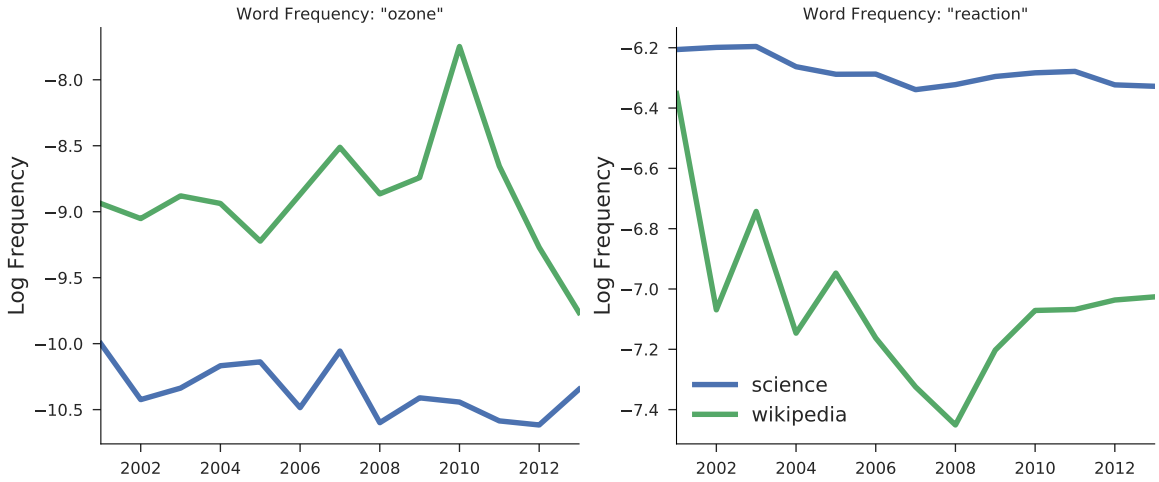

Figure 11: Frequencies of Chemistry words in Wikipedia edits and new journal articles

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. Table 2 shows the results from the regression in 5.1.3.

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

	Similarity (OLS)	Similarity (q=25%)	Similarity (q=50%)	Similarity (q=75%)
Intercept	1.1488^{***} (0.0050)	0.4125^{***} (0.0022)	0.7868^{***} (0.0037)	1.4329^{***} (0.0068)
After	0.0081^{***} (0.0022)	0.0058^{***} (0.0007)	0.0078^{***} (0.0012)	0.0089^{***} (0.0022)
N	4448939	4448939	4448939	4448939
R^2	0.00			
Adjusted R^2	0.0000			
F Statistic	42.56			

Note: Standard errors are dyadically clustered (Cameron and Miller, 2014). Significance: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

The positive and highly statistically significant coefficient on "After" in the regression confirms that articles published afterwards are indeed more similar. The quantile regressions, also shown in Table 2, confirm that these effects are not limited to the addition of a few highly-impactful Wikipedia articles, but instead represent a broad effect across most new articles.

Having observed the average effect, we now examine the distributional changes — that is, which scientific journal articles are affected. We expect journal articles on highly similar topics (and thus

with higher cosine similarity scores) to be most affected. Not surprisingly, Figure 12 shows that this is exactly what is happening, with an increase of $\sim 1\%$ in the number of articles with a similarity in the 95^{th} quantile and a $\sim 3.5\%$ increase in those with similarity in the $\sim 99^{\text{th}}$ quantile.

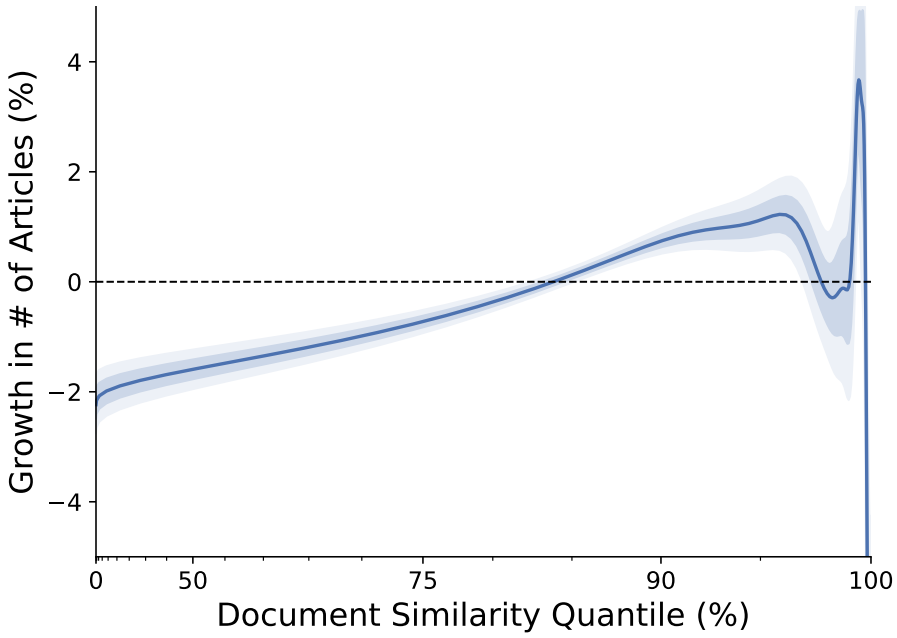

Figure 12: Proportional change in density of similarity between pre and post windows, not accounting for content drift

The estimate from Table 2 represents only the raw effect, of 0.008pp. This net effect needs to include the counterfactual content drift. Borrowing our experimental estimate for this counterfactual yields a net correlational effect of $\omega = 0.008\text{pp} - (-0.026\text{pp}) = 0.034\text{pp}$.

In appendix B, we show that analyzing these results using simple word-frequency (rather than cosine similarity) also shows this correlation between adding a Wikipedia article and increased use of those ideas in the scientific literature. This result holds whether this is analyzed on the extensive or intensive margin (i.e. both from more scientific articles using the words from the Wikipedia articles, and from those that use the words using them more).

To 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. The effect of a review article is calculated by re-running the observational analysis, but substituting contemporaneous review articles instead of Wikipedia articles as the treatment.[13] Quantitatively, we get a point estimate for the raw effect of a review article of 0.037pp (not statistically signficant, regression table shown in Appendix

[13]We observe review articles through a tag that Elsevier has in their records. We look at only the review articles published at the time of the Wikipedia articles to ensure that they are being tested against the same scientific literature, and hence are comparable analyses.

C), or a net effect, accounting for drift, of 0.063pp$=$ 0.037pp$-(-0.026\text{pp})$. If we accept this noisy point estimate, it suggests that a Wikipedia article's effect is roughly half as large as that of a review article.

The correlations in this section show that, averaged across roughly 18,000 new Chemistry articles in Wikipedia from 2001 to 2015, the creation of a Wikipedia article is followed by a statistically significant movement of the content in the scientific literature towards it.

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. We did this by having new Wikipedia articles on scientific topics written by PhD students from top universities who were studying those fields. Then, half the articles were randomized to be uploaded to Wikipedia, while the other half were not uploaded.[14] We then considered the differential impact that these articles 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 statistically 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 on science using textbooks and course syllabi from leading universities

[14] Holding back the control articles was only necessary until the end of the experiment.

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

3. Commission subject-matter 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 our journal-level analysis was also broad-based.[15] Some examples of the graduate articles that were identified as missing from Wikipedia and which we targeted for article creation included: Synthesis of Hydrastine, Multiple Michael/Aldol Reaction, and 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 content similarity between these Wikipedia articles and articles in the scientific literature using cosine similarity. Figure 13 shows the distribution of similarities for the experimental articles.

[15]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.

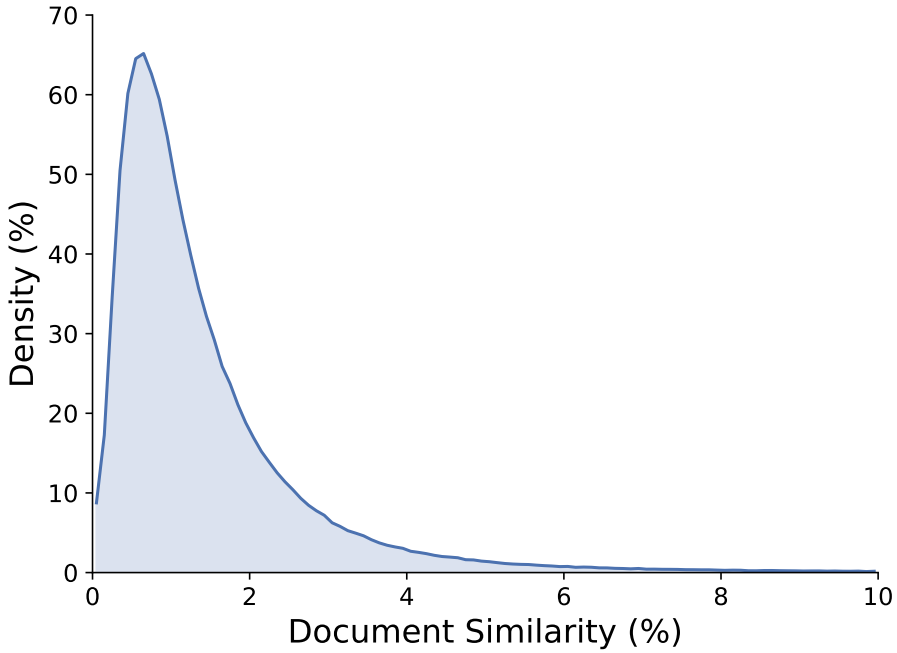

Figure 13: Density of similarity metric for all pairs of experimental Wikipedia and scientific articles

7.2 Article Stratification and Randomization

To maximize the statistical power of the experiment, we stratified the set of at-risk 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, and (5) number of non-academic references cited in the article.

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):

As Table 4 shows, the covariate balance is excellent across both sets of tests. This mitigates concerns of selection effects biasing our results. Our articles lengths are also consistent with those of average new articles outside the experiment, ∼250 words, as discussed in Section 4.1.1.

Table 4: 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

7.3 Implementation

The treatment articles were uploaded to Wikipedia in January 2015.[16] All the articles were initially uploaded as unique pages. After this point, the self-governing, open-source nature of Wikipedia became important. 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 editor 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 $\rightarrow$ scientists reading the articles $\rightarrow$ 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 content drift, the presence

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

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 to 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 difference in differences, comparing the similarity of articles in the scientific literature after the Wikipedia articles, as compared to beforehand (first difference), 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 14. Note that this figure plots changes across document similarity quantiles of the distribution shown in figure 13, *not* document similarity itself.

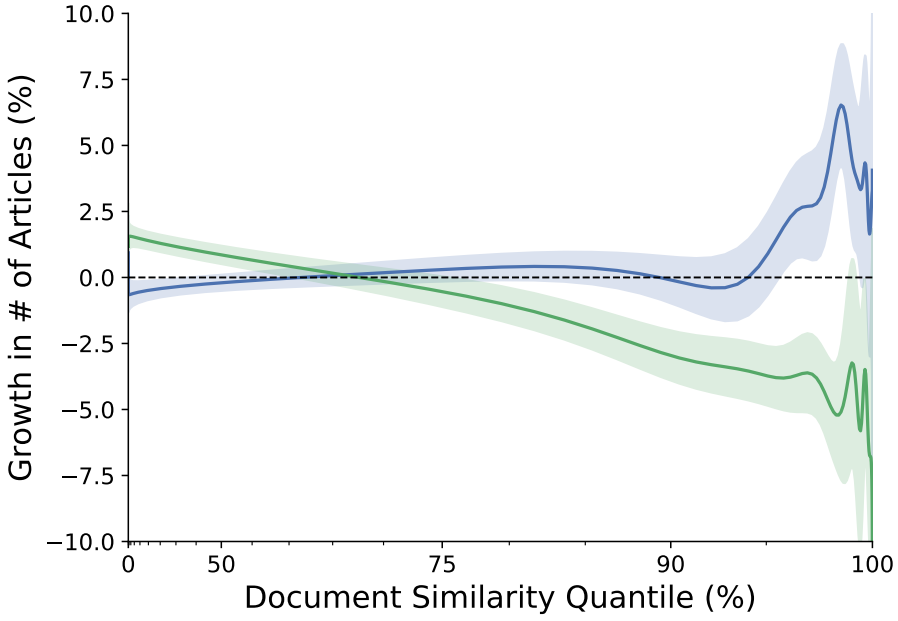

Figure 14: Treatment and Control proportional density differentials

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 natural content drift in the scientific literature – i.e. the evolution over time in the topics covered in science and words used to describe them. This is the same overall pattern we saw in Figure 8. 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 the ∼98% quantile suggests an increase of ∼7% in the number of articles at that level of similarity. Thus, Figure 14 already makes it clear that the addition of a new Wikipedia article *causes* an increase in the number of scientific literature articles that are similar to it.

The net effect can be seen by looking at the full difference-in-differences (i.e. by comparing the treatment effect to the counterfactual effect shown by the control articles). Figure 15 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 it would be surprising if our articles had effects on semantically-distant areas of Chemistry.

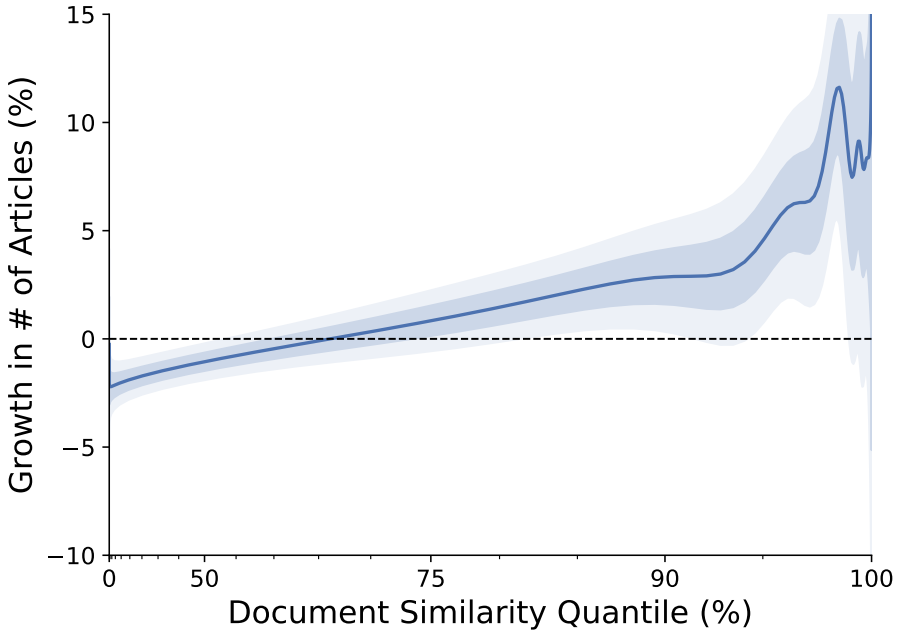

Figure 15: Effect of treatment on scientific article similarity to Wikipedia (the shaded areas indicate 1 and 2 standard errors, respectively)

Having presented our results visually, we now present them in a 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 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.

The coefficient estimate for the "Treated" indicator in Column 1 in Table 5 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

Table 5: Experimental Results

	Similarity (OLS)	Similarity (q=25%)	Similarity (q=50%)	Similarity (q=75%)
Intercept	1.4823^{***}	0.6480^{***}	1.1005^{***}	1.8561^{***}
	(0.0932)	(0.0568)	(0.0841)	(0.1205)
Treated	0.0068	-0.0410	-0.0867	-0.1301
	(0.1583)	(0.0711)	(0.1053)	(0.1719)
After	-0.0255^{**}	-0.0092^{***}	-0.0176^{***}	-0.0332^{***}
	(0.0110)	(0.0027)	(0.0037)	(0.0085)
Treated x After	0.0330^{***}	0.0131^{***}	0.0234^{***}	0.0479^{***}
	(0.0102)	(0.0044)	(0.0077)	(0.0170)
N	850604	850604	850604	850604
R^2	0.0001			
Adjusted R^2	0.0001			
F Statistic	26.07			

Note: Standard errors for OLS are dyadically clustered (Cameron and Miller, 2014) and for quantile regressions are bootstrapped. Significance: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

randomization worked. The coefficient on "After" shows the content drift of -0.026pp**, i.e. towards lower similarity. Our coefficient of interest, "Treated x After," shows that adding a Wikipedia article on a topic moves the scientific literature that follows it by 0.033pp***. The quantile regressions again confirm that this average effect is not just from a few highly-impactful Wikipedia articles, but instead that most new Wikipedia articles have (statistically significant) impacts.

A back-of-the envelope calculation can help make our estimated effects more concrete by considering the the number of articles affected. If we assume that the impact is focused in the top 4% of the most-similar journal articles where Figure 15 shows that our estimates are statistically significantly different from zero, we see an increase of $\sim$ 7%. This implies that 41 Elsevier journal articles in Chemistry were affected. If we then scale this up to account for Elsevier's share of the journal market (16%, as of 2015 (Reller, 2016)), then we would estimate each Wikipedia article is influencing $\sim$ 250 scientific articles (to some extent).

Our results indicate that Wikipedia articles *causally* affect the content of scientific articles and our back-of-the-envelope estimates suggests that these effect sizes are meaningful and that they happen quickly. While this might seem surprising, it is actually in-line with previous results about broadening the exposure of Science. Recall that Phillips et al. (1991) found that the publication of an average New York Times article about New England Journal of Medicine research led to the underlying article getting 73% more citations in its first year. And that effect comes from a single article being written

on a single day in 1978 when the New York Times circulation was roughly one million subscribers (NYT 1979) and there were no online ways of accessing the article. Even if 1% of New York Times subscribers read a particular article, more people were reading our Wikipedia articles than were reading their newspaper story. Moreover, because the Wikipedia readers actively sought out our articles, they are probably more likely to be influenced by them.

Is it good that Wikipedia influences Science in this way? For authors, revealed preference suggests the answer is "yes." By choosing to use the Wikipedia information, authors are indicating that they deem the Wikipedia information to be "better" in aggregate. This could be for many reasons, for example because the information is of higher quality, because it is more easily accessible, or because it is easier to understand.

But author preferences are only one part of what we might care about. 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. In economic parlance, there might be spillovers from an author using Wikipedia on other researchers. For example, if Wikipedia were less reliable than the source the author would have used otherwise, then erroneous information might be passed on to readers via the author's article. Conversely, if Wikipedia had better information (or were replacing a situation where no alternate source would be used), then there could be positive spillovers on readers.

To better understand the impacts of the Wikipedia articles, we consider a series of questions that shed light on the mechanisms that underpin our results and the distribution of their effects:

- Which parts of scientific articles are being changed (e.g. Methods, Results, etc.)?
- Does using Wikipedia as a source help or hurt the quality of the resulting scientific journal articles?
- Do particular groups (e.g. those without access to traditional journal articles) benefit disproportionately?

8.2 Which parts of scientific articles are being changed?

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 sections of the paper. In Chemistry, the structure of articles is quite standardized. The vast majority follow the layout: abstract, introduction, methods, results, conclusion (or some minor variation thereof).

Table 6: Effect of Wikipedia by Article Section

	Abstract	Introduction	Methods	Results	Conclusion
Intercept	0.7443^{***}	1.2928^{***}	1.0021^{***}	1.3316^{***}	0.8807^{***}
	(0.0404)	(0.0873)	(0.0727)	(0.0898)	(0.0538)
Treated	0.0255	−0.0147	−0.0389	0.0387	−0.0066
	(0.0785)	(0.1472)	(0.1254)	(0.1493)	(0.0898)
After	0.0224^{***}	0.0044	0.0047	-0.0209^{**}	-0.0162^{**}
	(0.0096)	(0.0103)	(0.0073)	(0.0113)	(0.0089)
Treated x After	0.0040	0.0275^{**}	0.0123^{*}	0.0232^{**}	0.0153^{**}
	(0.0122)	(0.0120)	(0.0092)	(0.0107)	(0.0084)
N	511262	696944	791027	619246	706677
R^2	0.0002	0.0000	0.0001	0.0003	0.0003
Adjusted R^2	0.0002	0.0000	0.0001	0.0003	0.0003
F Statistic	40.55	4.19	15.31	58.20	66.31

Note: Standard errors for all sections are dyadically clustered (Cameron and Miller, 2014). Significance: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

The results for this analysis are displayed in Table 6. Here we can see that there is a statistically significant effect in all sections except the abstract. The size and statistical significance is weakest in the Methods section and strongest in the Introduction. This suggests that our Wikipedia articles are having their largest effect on the contextualization of science and the connections that the authors are making to the rest of the field. This latter finding is consistent with our hypothesis that a Wikipedia article is essentially an easily-accessible review article. The point estimate for changes to the methodology section is only about half as large as those for other sections and only weakly statistically significant. This *might* indicate that scientists are less disposed to shape their experiments based on the content of Wikipedia than they are to contextualize it. Alternatively, however, it could just indicate that our window for our estimation (3 to 9 months after the Wikipedia article is first created) is less likely to pick up methodology changes, which may be made long-before a scientific article is published. A nice solution to resolve this ambiguity would be to use a longer observation window. Unfortunately the natural evolution of content on Wikipedia makes this difficult. Put another way, our intent-to-treat estimator loses statistical power rapidly as the underlying text diverges from our uploaded version. As such, we cannot distinguish between these interpretations.

8.3 Does using Wikipedia as a source help or hurt the quality of the resulting scientific journal articles?

Evaluating the scientific quality of a journal article, even apart from this experiment, is tricky. As such, we do not attempt to answer the question fully. Nevertheless, we can ask two questions that shed partial light on this question, and which are of independent interest to academics: do articles that use Wikipedia as a source get fewer citations, and does Wikipedia help direct scientists to good primary sources?

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.[17] We find no evidence that Wikipedia-influenced articles accrue either more or less citations than non-Wikipedia influenced ones.

We can analyze in more detail whether Wikipedia helps direct scientists to primary sources. We hypothesize that Wikipedia acts like a review article, in that it discusses topics broadly and accessibly, referencing primary sources as the scientific support. If correct, and Wikipedia helps scholars identify high-quality primary sources, then this function alone would make Wikipedia an important public good for science, as "Wikipedia is the 6^{th} highest referrer of DOI links (the unique hyperlinks assigned to academic articles)" (AOASG, 2017).[18]

We measure the ability of Wikipedia to direct scientists to the underlying primary source using the references we added as part of our experiment. Each of the Wikipedia articles that we created for our experiment averaged $2-3$ academic references. When we randomized our articles in the experiment, we implicitly also randomized the academic references being added to Wikipedia. That is, there is implicitly a second experiment that we ran on the effect of randomly adding academic references citations to Wikipedia. To test this effect, we look at average monthly citations to the primary sources listed in both the treatment and control articles in the 2 year windows before and after our Wikipedia intervention.

Table 7 reports multiple specifications for estimating these effects. As the outcome variable is count (integer) data, we employ both Poisson regression and Negative Binomial regression in addition to OLS. In specifications 1 and 3 we see that scientific articles referenced in the treatment Wikipedia

[17] This hurts our statistical power and thus makes it less likely for us to find a non-zero effect.

[18] According to Wass (2019) at Crossref, the top ten DOI referrers are: (1) webofknowledge.com; (2) baidu.com; (3) serialssolutions.com; (4) scopus.com; (5) exlibrisgroup.com; (6) wikipedia.org; (7) google.com; (8) uni-trier.de; (9) ebsco.com; (10) google.co.uk.

articles get $e^{0.6490} - 1 \approx 91\%$ more citations than those referenced in the control Wikipedia articles, and this difference is highly significant. In specifications 2, 4, and 6, rather than using a yes-no treatment indicator, we use the number of pageviews for the Wikipedia article as a measure of the intensity of the treatment. By definition, Wikipedia articles in the control group were not posted, and thus have no pageviews.[19] The estimate for specification 1 shows that getting 100% more (i.e. double) the pageviews increases citations by $2^{0.1762} - 1 \approx 13\%$.

Table 7: Effect of Wikipedia on Article Citations (Log)

	(1) Poisson	(2) Poisson	(3) Neg Binom	(4) Neg Binom	(5) OLS	(6) OLS
Treatment	0.6490^{***}		0.6490^{***}		0.2131	
	(0.0433)		(0.1995)		(0.2391)	
Log Views		0.1762^{***}		0.1508^{***}		0.0690^{*}
		(0.0066)		(0.0325)		(0.0386)
Intercept	2.7726^{***}	2.5224^{***}	2.7726^{***}	2.6181^{***}	1.4803^{***}	1.3975^{***}
	(0.0365)	(0.0362)	(0.1504)	(0.1389)	(0.1791)	(0.1630)
N	107	107	107	107	107	107
R^2					0.0075	0.0296
F Statistic					0.7945	3.1978

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.

Collectively, our results are unanimous in finding a positive citation effect and show the expected effect: that more views of a Wikipedia page lead to more citations.

8.4 Distributional Effects

One might imagine that public repositories of knowledge would be particularly valuable to those with fewer other sources of knowledge, for example developing country scientists without access to scientific journals. Conversely, scientists might benefit less from Wikipedia articles if they cannot access the journal articles that Wikipedia references. 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

[19] For any case in which we analyze $\log(x)$ for a variable that can take zero as a value, we use $\log(1+x)$.

country of the scientific authors (assuming that those with lower GDP per capita will have less access). Our results are presented in Table 8, where we present one regression for each quintile of the GDP per capita distribution, except the two lowest which we pool because the sample sizes are much smaller.

Table 8: Experimental Results by the Quintile of Country GDP per capita)

	Low (<40%)	Mid (40-60%)	High (60-80%)	Top (>80%)
Intercept	1.7113^{***} (0.1364)	1.4176^{***} (0.0880)	1.6461^{***} (0.1061)	1.5200^{***} (0.0973)
Treated	0.0385 (0.2165)	0.0066 (0.1531)	0.0034 (0.1792)	0.0057 (0.1609)
After	0.0324 (0.0487)	-0.0174 (0.0179)	-0.0066 (0.0296)	-0.0615^{***} (0.0170)
Treated x After	-0.0141 (0.0270)	0.0179^{*} (0.0115)	0.0445^{**} (0.0257)	0.0622^{***} (0.0144)
N	57529	269531	121398	357906
R^2	0.0001	0.0000	0.0001	0.0003
Adjusted R^2	0.0000	0.0000	0.0001	0.0003
F Statistic	1.91	3.88	5.34	41.20

Note: Standard errors for OLS are dyadically clustered (Cameron and Miller, 2014) and for quantile regressions are bootstrapped. Significance: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Interestingly, we find, that wealthier countries benefit much more than poorer countries. This is presented graphically in Figure 16 (error bars are $\pm$ one standard error).

These results suggest that the biggest benefits accrue to the highest-GDP countries, but that there are also substantial benefits for some developing-world scientists, although not for those in the poorest countries. There are multiple plausible explanations for this finding. We view the most likely one as being that usage of Wikipedia is lower in low GDP-per-capita parts of the world (Zachte, 2018), and thus the weaker effect comes from reduced exposure to our articles . Another explanation, as suggested earlier, could be that access to academic journals is necessary to get the full benefit of a Wikipedia article. Finally, 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 graduate-level content at developed-world universities and, for example, Jones (2011) has shown that there significant differences in the topics and level of specialization of classes between developed and developing world universities. We are not able to distinguish between these effects, so we just note them as potential hypotheses.

We also examine if our estimated effects vary by journal quality. To do this, we disaggregate the

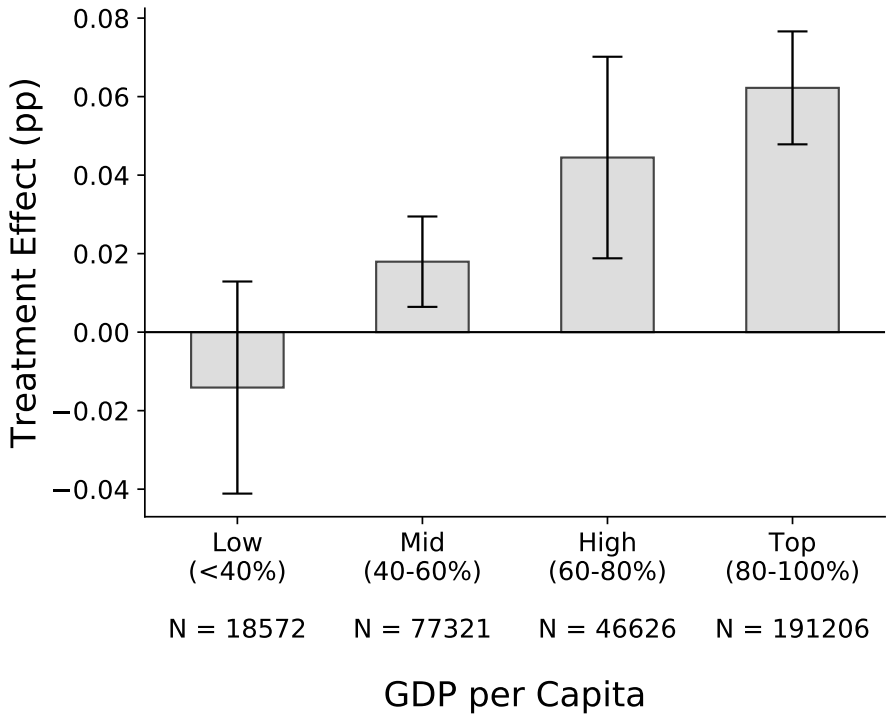

Figure 16: Effect of Wikipedia across Countries with varying access to other resources

treatment effect into four quartiles, dividing 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.

9 Discussion

When Darwin extoled the importance of "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.

A back-of-the-envelope calculation reveals the scale of Wikipedia as a repository of review articles. Across the Chemistry journals in our sample, from 2001 to 2014, there were an average of 553 review articles per year. Scaled up by Elsevier's share of the market, this implies $\sim$ 3400 review articles per year across comparable Chemistry journals. In that same period, Wikipedia averaged 1900 new Chemistry articles per year (although this somewhat understates Wikipedia's role since review articles

in the academic literature are seldom updated as time passes, whereas Wikipedia's 27000 Chemistry articles are actively edited and deepened, as we saw in section 5). Thus, depending on how one counts the evolving Wikipedia articles, Wikipedia is either the largest or second-largest repository of up-to-date review articles articles in the world. Moreover, even though Wikipedia articles have a weaker effect on the literature than review articles, there are so many of them that Wikipedia is highly likely to also be the first- or second-most *influential* repository[20].

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 in a separate repository. Two existing examples of this type of effort include the Stanford Encyclopedia of Philosophy (Zalta 2019) and the Encyclopedia of Database Systems (Liu and Ozsu, 2019). 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 order-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 each Wikipedia article that we created for this

[20]Two additional comparisons reinforce the size of Wikipedia as a repository of reviews. The Stanford Encyclopedia of Philisophy has 1600 articles (Zalta, 2019). The Cochrane Database of Systematic Reviews, which published the "third highest number of citable items of the journals in Medicine, General & Internal" has 1764 citable items (Cochrane Library, Undated).

experiment we paid students $100. Assuming one Wikipedia article (or equivalent contribution) per research paper, the implicit tax on research would be $(\frac{\$100}{\$220,000}) = 0.05\%$. This is a small enough tax to be feasible, but is it more effective than other methods of disseminating knowledge?

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). If a paper generates N citations[21], 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![22,23] And thus we conclude something which is fairly intuitive: an average paper cannot influence more than $\left(N * \frac{1}{N}\right) = 1$ other papers. This implies that the cost of influencing another paper through grant funding is, at best, $\frac{\$220,000*1\%}{1} = \$2200$ [24]

In contrast, we found in Section 8.1 that the causal impact of adding a Wikipedia article was to influence ~ 250 other papers [25] and that the causal effect was a change in cosine similarity of 0.0330 on a base 1.48. Given these assumptions, the cost of disseminating knowledge through Wikipedia would be $\frac{\$100}{250*\frac{0.0330}{1.48}} = \$18$.

Our back-of-the-envelope analysis thus has stark conclusions: even with many conservative as-

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

[22] 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.

[23] Of course, multiple pieces of research could influence the same follow-on work, 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.

[24] This number could be an over- or under-estimate for many reasons. It could be an underestimate because NIH-funded work 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. For example, even if we clearly overstated the impact of a paper by giving it 100% credit for each citation made to it, this would be $\frac{\$220,000*1\%}{19} = \$116$.

[25] Wikipedia articles are not bound by the one-to-one ratio since there are many more scientific articles than Wikipedia articles.

sumptions, dissemination through Wikipedia is $\sim 120\times$ more cost-effective than traditional dissemination techniques. This is driven almost entirely by the extraordinary reach of Wikipedia, and thus, from a public policy perspective, funding the creation of content for accessible 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).

We hope that our findings also help overturn the approach that many academics and universities take to Wikipedia: urging people not to use it. Our findings show that this has not worked, and that scientists themselves (and surely students as well) use Wikipedia as a source and that it influences their work. Rather than trying to dissuade people from using Wikipedia for fear of it being untrustworthy, we encourage academics to embrace Wikipedia and make it more trustworthy.

10 Conclusion

This paper analyzes the impact of public repositories of scientific knowledge. Using a randomized control trial, we show that the creation of a Wikipedia science article leads to changes in hundreds of follow-on articles in the scientific literature — providing strong evidence that Wikipedia is an important source for disseminating knowledge. Because our work goes 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 intervention. 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 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 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. Because of Wikipedia's enormous scope, this almost assuredly means that it is one of the most important sources of scientific review articles in the world.

This paper shows that Wikipedia's contribution to Science is substantial. It disseminates an enor-

mous 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

Aguiar, Luis and Joel Waldfogel, "Quality Predictability and the Welfare Benefits from New Products: Evidence from the Digitization of Recorded Music," Journal of Political Economy 126, no. 2 (April 2018): 492-524.

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.

Cochrane Library. Undated. "Cochrane Database of Systematic Reviews: 2017 Journal Impact Factor and Usage Report." Available online at https://community.cochrane.org/sites/default/files/uploads/inline-files/CDSR-2017JournalImpactFactorandUsagereport.pdf.

Cohen, Noam. 2014. "Wikipedia vs. the Small Screen." *New York Times*: Feb 9, 2014. Available online at: https://www.nytimes.com/2014/02/10/technology/wikipedia-vs-the-small-screen.html?_r=0.

Clarivate Analytics. 2017. *Web of Science*. Online at http://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.

Dolado, J.J., Felgueroso, F. and Almunia, M. SERIEs (2012) 3: 367. https://doi.org/10.1007/s13209-011-0065-4.

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

Elsevier. 2019. Fast Publication data pages. Available online at urlhttps://www.elsevier.com/physical-sciences-and-engineering/chemistry/journals/fast-publication-in-physical-and-theoretical-chemistry, urlhttps://www.elsevier.com/physical-sciences-and-engineering/chemistry/journals/fast-publication-in-organic-and-

inorganic-chemistry, urlhttps://www.elsevier.com/physical-sciences-and-engineering/chemistry/journals/fast-publication-in-analytical-chemistry. Accessed January 2019.

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.

Jemileniak, Dariusz, Gwinyai Masukume, and Maciej Wilamowski. 2019. "The Most Influential Medical Journals According to Wikipedia: Quantitative Analysis." Journal of Medical Internet Research, vol 21, issue 1, p1.

Jones, Benjamin F. 2011. "The knowledge trap: human capital and development reconsidered." NBER Working paper 14138.

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

Liu, Ling and M. Tamer Ozsu. 2019. *Encyclopedia of Database Systems*. Springer. Available online at https://link.springer.com/referencework/10.1007/978-1-4614-8265-9

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/.

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/.

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/.

NYT. 1979. "Effects of the '78 Newspaper Strike on Sales and Ads." *New York Times*, Feb 5, 1979.

Phillips, D.P., Kanter, E.J., Bednarczyk, B. and Tastad, P.L. 1991. "Importance of the lay press in the transmission of medical knowledge to the scientific community." *The New England Journal of Medicine*, 325(16), pp.1180-1183.

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.

Reller, Tom. 2016. "Elsevier publishing - a look at the numbers, and more." Online at urlhttps://www.elsevier.com/connect/elsevier-publishing-a-look-at-the-numbers-and-more, accessed Jan 2019.

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.

Shahmirzadi, Omid, Adam Lugowski and Kenneth Younge. 2018. "Text Similarity in Vector Space Models: A Comparative Study." arXiv working paper. Available at https://arxiv.org/abs/1810.00664v1

Zalta, Edward. 2019. *Stanford Encyclopedia of Philosophy*. Available at https://plato.stanford.edu

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.

Turney, P.D. and Pantel, P. 2010. "From frequency to meaning: Vector space models of semantics." *Journal of Artificial Intelligence Research*, 37, pp.141-188.

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.

Van Noorden, Richard. 2014. "Global scientific output doubles every nine years."

Nature News Blog, May 7 2014, available at http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html.

Waldfogel, Joel. 2017. "The Random Long Tail and the Golden Age of Television," Innovation Policy and the Economy 17: 1-25. https://doi.org/10.1086/688842

Wass, Joe. 2019. "Where do DOI clicks come from." CrossRef.org blog, available online at https://www.crossref.org/blog/where-do-doi-clicks-come-from/

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

Wikipedia. 2019. "Web of Science." Accessed Feb 15, 2019. Available online at https://en.wikipedia.org/wiki/Web_of_Science.

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.

Younge, Kenneth A. and Kuhn, Jeffrey M. 2016. "Patent-to-Patent Similarity: A Vector Space Model." Working Paper. DOI: 10.2139/ssrn.2709238.

Zachte, Erik. 2018. "Pageviews per capital to any Wikipedia in September 2018." Available online at https://stats.wikimedia.org/wikimedia/animations/wivivi/wivivi.html, accessed March 24, 2019.

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: Two-Step M-Estimators, Smoothed Maximum Score estimation, and 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

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.1: 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

As Table A.1 shows, the covariate balance is excellent.

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, although interestingly we still observe statistically significant effects.

Table A.2: Experimental Results - Econometrics

	Similarity (OLS)	Similarity (q=25%)	Similarity (q=50%)	Similarity (q=75%)
Intercept	2.6735^{***} (0.1163)	1.3089^{***} (0.0811)	2.0404^{***} (0.1168)	3.2068^{***} (0.1661)
Treated	0.0895 (0.1620)	0.0369 (0.1111)	0.0830 (0.1564)	0.1474 (0.2214)
After	-0.1051^{**} (0.0557)	-0.0404^{***} (0.0094)	-0.0299^{***} (0.0082)	-0.0558^{***} (0.0179)
Treated x After	0.0221^{***} (0.0084)	0.0259^{**} (0.0149)	0.0146 (0.0195)	0.0199 (0.0280)
N	149588	149588	149588	149588
R^2	0.0009			
Adjusted R^2	0.0008			
F Statistic	43.33			

Note: Standard errors for OLS are dyadically clustered (Cameron and Miller, 2014) and for quantile regressions are bootstrapped. Significance: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

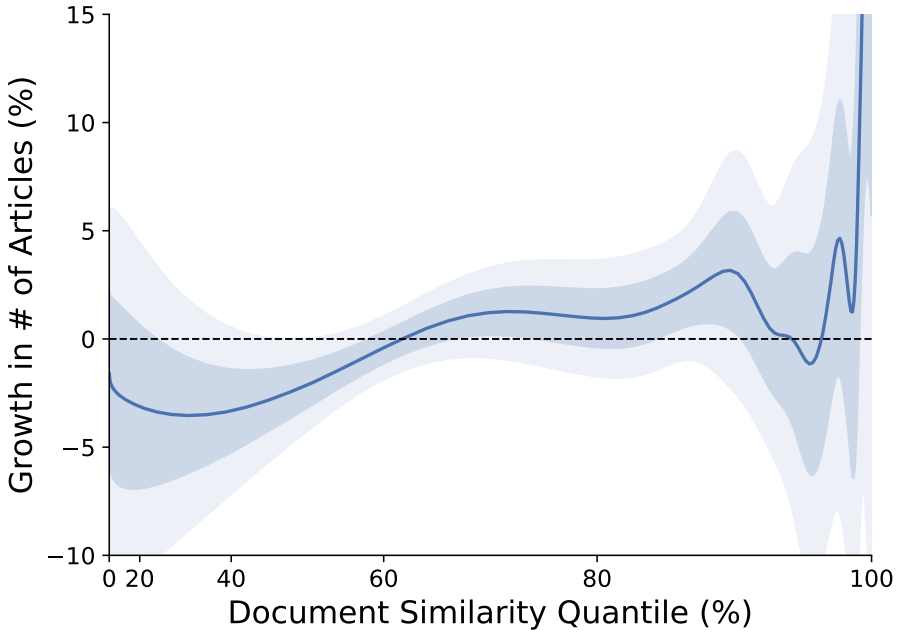

Figure A.1: Treatment Effect Estimates for Econometrics

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.2 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

[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.

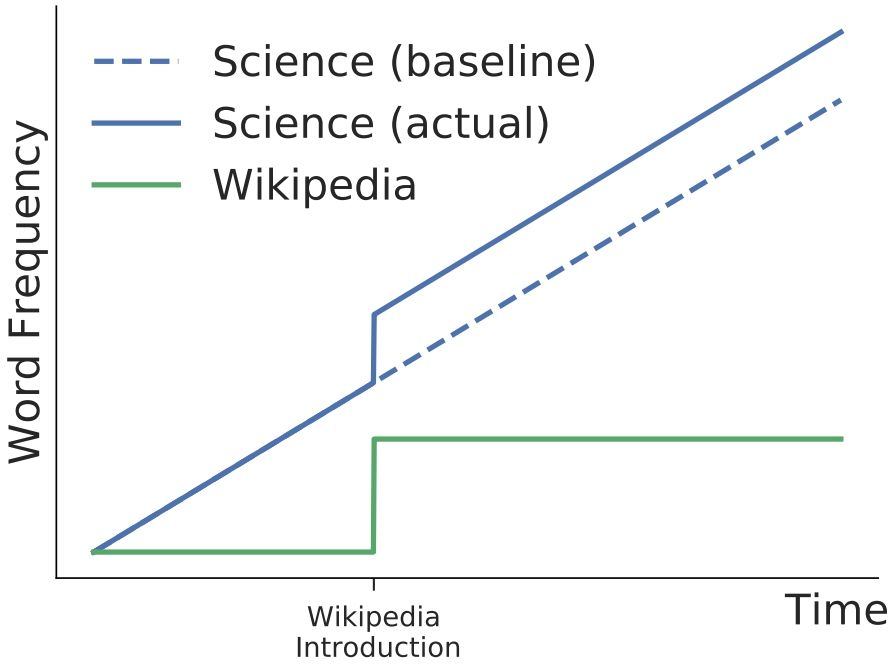

Figure A.2: Schematic of the how Wikipedia might shape science

word above what would be seen absent it.[29] Here word frequencies follow the difference equation

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

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

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.

[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.

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 $2^{0.0797} - 1 \approx 5.7\%$ increase in science frequency.

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

	Log Science Frequency (t+1)
Intercept	-0.3958^{***}
	(0.0074)
Log Wikipedia Frequency (t)	0.0797^{***}
	(0.0007)
Log Science Frequency (t)	0.8949^{***}
	(0.0006)
N	587715
R^2	0.8812
Adjusted R^2	0.8812
F Statistic	2179164
Note:	$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

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

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

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 $2^{0.0462} - 1 \approx 3.3\%$ 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 inten-

[30] Importantly, although it has an intuitive appeal, this equation is NOT unbiased because any idiosyncratic error associated with $f^{sci}_{i,t}$ 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	-0.1629^{***}
	(0.0067)
Log Wikipedia Frequency (t)	0.0462^{***}
	(0.0006)
Log Science Frequency (t)	-0.0558^{***}
	(0.0006)
Δ Log Science Frequency (t)	-0.4355^{***}
	(0.0012)
N	587715
R^2	0.2331
Adjusted R^2	0.2331
F Statistic	59535
Note:	$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

sive 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.

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

	Science Frequency > 0 (t+1)
Intercept	0.1139^{***}
	(0.0002)
Wikipedia Frequency > 0 (t)	0.1865^{***}
	(0.0004)
Science Frequency > 0 (t)	0.4843^{***}
	(0.0003)
N	7337902
R^2	0.3198
Adjusted R^2	0.3198
F Statistic	1725097
Note:	$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

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

C Regression Results for Review Articles

Table A.6: Review Article Results

	Similarity (OLS)	Similarity (q=25%)	Similarity (q=50%)	Similarity (q=75%)
Intercept	2.9588*** (0.0877)	1.3966*** (0.0551)	2.2821*** (0.0665)	3.6209*** (0.0991)
Review	0.2691** (0.1259)	0.1988** (0.0929)	0.2699*** (0.1156)	0.3509** (0.1726)
After	0.0087 (0.0344)	−0.0103 (0.0227)	−0.0146 (0.0297)	−0.0006 (0.0429)
Review x After	0.0370 (0.0531)	0.0475 (0.0419)	0.0536 (0.0489)	0.0437 (0.0673)
N	7512866	7512866	7512866	7512866
R^2	0.0025			
Adjusted R^2	0.0025			
F Statistic	6254.40			

Note: Standard errors for OLS are dyadically clustered (Cameron and Miller, 2014) and for quantile regressions are bootstrapped. Significance: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

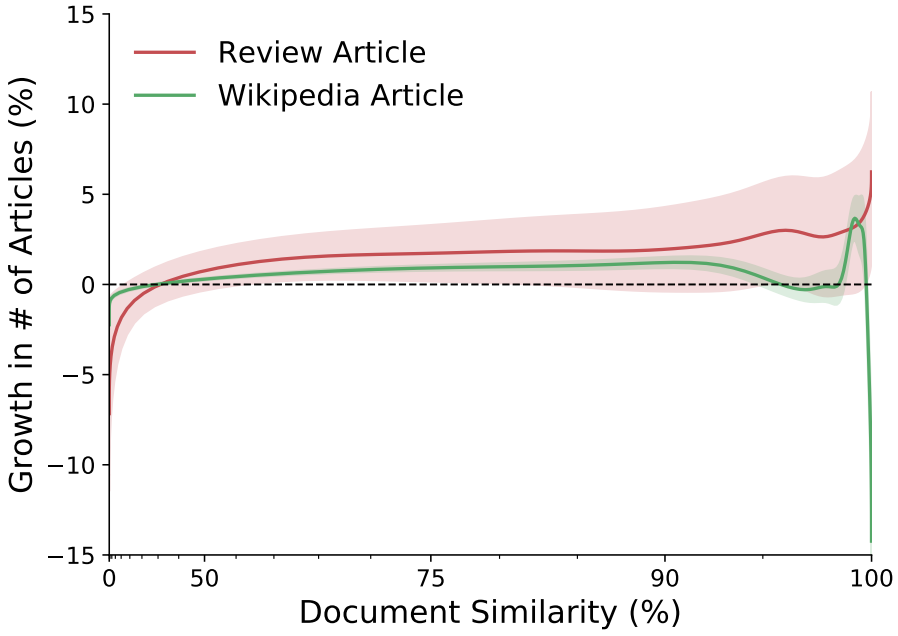

Figure A.3: Effect of new Review Articles or new Wikipedia Articles on the scientific literature (observational analysis)