The Education-Innovation Gap*

Barbara Biasi† Song Ma‡

Education plays a key role in the production of human capital by equipping individuals with up-to-date knowledge and skills. What is the distance between the content of higher education and the frontier of knowledge? How does it vary across schools and students' background? How does it relate to students' economic outcomes? This paper answers this question with a new measure: the "education-innovation gap," defined as the distance between the material taught in college and university courses and the content of frontier research, patents, and employment postings. Using information on the text of three million syllabi across US institutions between 1992 and 2018, we calculate this gap as the ratio between the textual similarity of each syllabus with old and new academic publications, patents, and job ads. Armed with this new measure, we document three facts. First, the gap varies considerably within field, and it has steadily increased over time since 2000. Second, significant differences exist across schools serving different populations of students: The gap is smallest for Ivy-League and elite schools and for schools enrolling children of affluent parents, and largest for schools with the largest proportion of minority students. Third, the gap is correlated with student outcomes such as graduation rates, income, and rates of intergenerational mobility, especially for students in lower-tiered universities. This evidence suggests that the "education-innovation gap" is a useful dimension to examine for human capital accumulation, labor market success, innovation, and inequality.

JEL Classification: I23, I24, I26, J24, O33

Keywords: Education, Innovation, Textual Analysis, Inequality

*We thank seminar participants at Yale and at the Junior Entrepreneurial Finance and Innovation Workshop for helpful comments. Xugan Chen provided excellent research assistance. We thank Yale Center for Research Computing, Yale University Library, and Yale International Center for Finance for research support and Bledi Taska for helping us with the Burning Glass data. All errors are our own.

†Yale School of Management and NBER, barbara.biasi@yale.edu, +1 (203) 432-7868;

‡Yale School of Management, song.ma@yale.edu, +1 (203) 436-4687.

1. Introduction

Human capital is increasingly important for innovation and economic growth. In models of endogenous growth, human capital acts as a factor-augmenting term in the "knowledge" production function (Romer, 1990). Differences in human capital have been shown to explain differences in GDP across countries (Hanushek and Woessmann, 2012) and US states (Hanushek et al., 2017). Furthermore, as human capital determines individuals' labor market outcomes, the distribution of human capital in the population shapes the inequality and social mobility of any society (Goldin and Katz, 1996, 2008; Acemoglu and Autor, 2012).

When thinking about the production of human capital, economists have naturally focused on education. Initially, scholars have measured human capital using literacy rates (Romer, 1990) or educational attainment (Hanushek and Woessmann, 2012). More recent work, however, has placed emphasis on the type and content of education that individuals receive. Baumol (2005), for example, discusses the importance of "scientific" or technical knowledge in producing innovation. Altonji et al. (2012), among others, show how different types of college majors and school curricula lead to very different returns in the labor market (see Altonji et al., 2012, for a review). More recently, a growing body of evidence has revealed an important role of skills, partly acquired in school, in determining people's labor market success (Deming and Kahn, 2018; Hershbein and Kahn, 2018). An important feature of successful education is thus the ability to equip individuals with up-to-date knowledge and skills, facilitate innovation diffusion, which will ultimately lead to growth.

What is the distance between the content of higher education and the most up-to-date knowledge and skills? How does access to frontier knowledge vary across students with different backgrounds? How does the access to frontier knowledge relate with students' education choices, labor market success, inequality, and social mobility?

This paper addresses these questions by studying the "content" of higher education, using three millions of university course syllabi in nearly three decades. With these data, we build a novel measure: the education-innovation gap, i.e., the distance between the content of college and university courses and new vs old "technologies," such as academic research, innovation, and skills in high demand in the labor market. The intuition behind this measure

is as follows: A course will have a smaller gap if its content is more similar to newer academic papers, patents, or job ads than it is to older ones. The goal of this measure is to capture how close the content of a course is to the "technological" frontier, defined in terms of academic research, innovation, or labor market demand. Intuitively, a Computer Science course that teaches machine-learning techniques in 2018 is closer to the technological frontier than one which teaches more obsolete programming languages; at the same time, it should be more similar to academic articles published the previous year than to articles published 15 years back, and therefore have a smaller education-innovation gap.

We calculate the education-innovation gap for a sample of almost three million higher education courses taught in 2,500 US institutions and 62 different fields between 1992 and 2018. To do so we first collect information on the texts of these courses' syllabi, the titles and abstracts of twenty million academic articles published in top journals in each field since the 1970s, the abstracts of more than six million patents filed between 1975 and 2018, and the texts of nearly 200 million job advertisements posted online between 2010 and 2018. We then use textual analysis techniques to calculate the "distance", textual cosine similarity in specific, between the content of each syllabus and the content of each article, patent, or job ad released up to 15 years prior to the syllabus date.

Specifically, we use the full set of words ever listed on Wikipedia as a dictionary and represent each document (syllabus, article, patent, or job ad) as a numerical array of zeros and ones, depending on whether each Wikipedia word is present or not in that document. We also use a weighting technique to account for a) the length of each document and b) the frequency of each word in the English language. We then measure the distance of each syllabus from each one of papers, patents, or job ads by calculating the cosine similarity between their respective arrays. The cosine similarity between two documents will be larger, the more similar the language that they use.

Having computed measures of language distance between a syllabus and technologies of different ages, we construct the education-innovation gap of a given course as the ratio between the average cosine similarity of its syllabus with technologies released 15 years prior and the cosine similarity with technologies released one year prior. Intuitively, the gap will be larger the more similar the syllabus is to older technologies, relative to newer ones.

A natural way to validate this measure is to check whether syllabi with a larger gap directly reference bibliographic works (such as textbooks or articles assigned as course readings) that were published in earlier years. The correlation between the gap and the average age of referenced materials is large and statistically significant. This suggests that our measure performs well in capturing the "newness" of a given course.

Armed with this new measure, we document three main facts. First, we illustrate a sizable variation in the gap across syllabi within field and institution, both in the cross-section and over time. For example, in order to move a syllabus from the 75th percentile to the 25th percentile of the gap distribution one would have to add 350 new "knowledge words", defined as the top 10% most popular key words in the SCOPUS publication data in the previous year (or 62.5 percent compared with an average syllabus length of 480 skills). Furthermore, the gap changed substantially over time: It declined between 1992 and 2001, by the equivalent of an increase of 170 new knowledge words per syllabus, and it subsequently increased between 2007 and 2017, by an equivalent of 45 fewer knowledge words.

Second, we find considerable variation in the gap across institutions serving different types of students. The gap is significantly smaller in "Ivy Plus" institutions (Chetty et al., 2019), as well as in institutions disproportionately serving children of wealthier parents, compared with all other institutions. For example, in order to make the average syllabus in a highly-selective or selective institution equivalent to a syllabus of a course in the same field taught in an Ivy Plus institution, one would have to add 70 new knowledge words to it (15 percent from the baseline).

We also find the gap varies substantially across students with different backgrounds. For example, in order to make the average gap in schools where no students have parents in the top income percentile equal to the average gap in schools where more than 10 percent of all students have parents in the top percentile, one would have to add 220 new skills to a syllabus, or 46 percent. Furthermore, the gap is disproportionately larger for schools primarily enrolling Black and Hispanic students compared with schools where these students are a small minority.

Third, we find that the gap strongly predicts students' outcomes such as graduation rates, income, and intergenerational mobility (measured as the probability of reaching the top

quintile of the income distribution for students with parental income in the bottom quintile). For example, a 0.023 smaller gap (i.e., a gap that has 1/3 additional updated new skills compared with the mean) is associated with a 1 percent higher graduation rate, or 2.5 percent compared with the mean. The same difference in the gap is also correlated with a 2 percent higher income and with a a 22.86 percent larger probability that students with parents in the bottom income quintile reach the top quintile during adulthood.

Interestingly, these patterns are almost entirely driven by schools outside the elite. In fact, the correlation between the gap and either graduation rates, income, or mobility is negative for elite schools and for schools disproportionately serving wealthier students; it is instead negative and large for students in non-selective schools and in those serving a more disadvantaged student body. A possible interpretation of these findings is that the content of higher education is particularly crucial in determining the economic fortune of students who do not have access to the network and resources that schools at the top of the selectivity distribution are able to provide. Access to academic content closer to the technological frontier may make these students more likely to stay in school, graduate, and have better outcomes when they enter the labor market.

Taken together, our findings reveal large differences in students' access to up-to-date knowledge in higher education by school selectivity, parental income, and race. At the same time they suggest a possibly large role for these differences in explaining later-life outcomes. It should be stressed that these findings are purely correlational and should be interpreted as descriptive. Future research will explore possible causal determinants of the education-innovation gap, as well as a causal role for changes in the gap on students' outcomes.

This paper makes three main contributions. First, it is among the first to use the text of college and university syllabi as data and it proposes a new measure to to characterize the "innovative" content of higher education. Similarly to Kelly et al. (2018), who use cosine similarities between the text of patent documents to measure patent quality, and Gentzkow and Shapiro (2010), who use the language of newspaper articles to measure media slant, we use textual analysis techniques to characterize the content of each course and to link it to frontier technologies.

Second, this paper it is among the first to study the role of the "content" of education as a measure of knowledge and to relate it to schools' characteristics and students' outcomes. Earlier works have highlighted the role of educational attainment (Hanushek and Woessmann, 2012), majors and curricula (Altonji et al., 2012), college selectivity (Hoxby, 1998; Dale and Krueger, 2011), and skills (Deming and Kahn, 2018) for labor market outcomes and economic growth. Our analysis focuses instead on the specific concepts and topics covered in higher education, and to the extent to which these are up-do-date with respect to the innovation frontier and the labor market.

Third, this paper contributes to the literature on the production of innovation. Earlier works (Nelson and Phelps, 1966; Benhabib and Spiegel, 2005) have highlighted a role for human capital for technology diffusion. More recently, Jones (2009) has shown how innovators are increasingly required to update their knowledge and skills to keep up with a fast-paced world; failure to do so delays the correlation between researcher/inventor life-cycle productivity (Jones, 2010; Jones and Weinberg, 2011). Our paper sheds light on the extent to which higher education is able to provide future innovators with up-to-date content.

2. Data

In this section, we describe how we construct our analysis sample and collect historical information on university syllabi, academic publication, patent filings.

2.1. Collecting and Processing Syllabi Data

Our primary sample of university syllabi is obtained from the American Assembly's Open Syllabus Project (OSP). OSP collects millions of syllabi to support educational research and novel teaching and learning applications. The current dataset contains of seven million English-language syllabi of college and university courses from over 80 countries, dating back to the 1960s.[1]

OSP provides basic information for each syllabus, including the institution, date, field

[1]OSP data are obtained primarily by crawling and scraping publicly-accessible (university) websites and university archives. Those websites include official university systems, faculty personal websites that detail teaching materials, among others. Voluntary faculty and student contributions make up a small but significant portion of the collection.

and subject, in the form of a dataset. The metadata, i.e., the text of each syllabus, is also available; we use these data as textual information in our empirical analysis. Specifically, we focus our attention on the course description, i.e., the part of the syllabus which generally includes a short description of the basic structure of the course, the key concepts that are covered, and –in some cases– a class-by-class or week-by-week schedule that further details the content of the course and the teaching materials. On average, this section is 2150-word long.

A subset of the syllabi (62 percent) also includes information on the course catalog, i.e., a list of all required readings. This information is identified by parsing the section of the syllabus often labeled as "references", "recommended readings", "textbook", as well as by searching for in-text citations that are not explicitly listed. We further standardize each referenced item and extract its bibliographic information, most noticeably the year of publication and the textbook editions.

In order to link the syllabus data with other data, we perform several additional data processing steps. We use a fuzzy-matching algorithm to standardize all institution names; this allows us to link each syllabus with information on its institution (see below). We also identify the course code of each syllabus, which allows us to a) recover the course level (e.g., undergraduate, graduate, advanced), and b) construct a panel of courses over time. Lastly, we are in the process of identifying the names of instructors, which will allow us to link each syllabus with information on the person who designed it.

Our analysis is restricted to syllabi from US institutions, representing 45 percent of the total corpus. Figure 1 (panel a) displays the number of syllabi available for each year. The earliest syllabi date back to 1975; the coverage dramatically improves over time, with 11,339 US syllabi in 1992, 69,594 in 2000, and 850,752 in 2017. To explore the patterns of selection into the sample, panel b) of Figure 1 outlines the field composition of syllabi over time. The composition remains quite stable across time, with the exception of an increase in the share of syllabi in STEM-related fields, equal to 20 percent in 1975, 28 percent in 1995, 30 percent in 2000, and 38 percent in 2018, and a decline in the share of humanities, with 45 percent of syllabi in 1975, 36 percent in 1995 and 2000, and 21 percent in 2018. To minimize the effect of selection into the sample over time, we focus our attention on 2.9 million syllabi of courses

taught between 1992 and 2018.

Syllabi appear evenly distributed across states. The number of syllabi per institution (in the period from 1992 to 2018) ranges from a minimum of 8 in Delaware to a maximum of 6,838 in Maryland (Figure AI, panel a), whereas the number of syllabi per student ranges from a minimum of 1.6 in Delaware to a maximum of 1,343 in Hawaii (Figure AI, panel b).

2.2. Academic Publications

We assemble a dataset of 20 million peer-reviewed articles in academic journals in the same fields of our syllabi using Elsevier's SCOPUS dataset, which lists all published articles since the journal's foundation along with their identifier (DOI), title, journal, author(s), publication date, and abstract.[2]

Since our ultimate goal is to capture the frontier of knowledge in each field at each point in time, we rank journals within field by Impact Factor and restrict our attention to journals that at least make into top ten in their respective field. This leaves us with 20 million articles in total and approximately 100,000 articles per publication year (Figure AII).

2.3. Patent Filings

We collect data on the text of more than six million patents from the USPTO patent search website, which records the text of all patents filed starting from 1976. The main textual information used in our analysis is the abstract field of all patents. Counts of patents per year are shown in Figure AII.[3]

2.4. Job Postings

The source of information on job postings is a database of 200 million employment vacancies provided by Burning Glass Technologies (hereafter BG), an employment and labor market analytics firm which collects and processes the near-universe of employment ads from nearly 40,000 online job boards and company web sites, removing duplicates. The BG data include education and experience requirements, detailed industry and occupation codes

[2]We access the SCOPUS data through the official API in April-August 2019.

[3]Our web crawler collected the text content of all patents (in HTML format) from http://patft.uspto.gov/netahtml/PTO/srchnum.htm, with patent number from 3850000 to 10279999).

(6-digit SOC), the location, the date the ad was posted, and –where available– firm identifiers. In addition, BG parses the actual text of each job vacancy and codes key words and phrases as additional job requirements, or "skills." This dataset has been used, among others, by Deming and Kahn (2018); Hershbein and Kahn (2018); Deming and Noray (2019). While these papers use BG's skill classification, we use the actual text of the ads as information in our analysis.

2.5. Information On US Higher Education Institutions

Our last piece of data contains institution-level information on all universities and colleges for which syllabi are available. Our primary source of data is the the Integrated Postsecondary Education Data System (IPEDS), maintained by the National Center for Education Statistics (NCES). Since 1993, IPEDS has collected data through a series of interrelated surveys from all postsecondary institutions; the completion of these surveys is mandatory for all institutions that participate in, or are applicants for participation in, any federal financial assistance program. Collected information includes institutional characteristics (such as name and address, control or affiliation, levels of awards offered, types of programs, tuition and fees), institutional prices, SAT and ACT scores of admitted students (where applicable), enrollment and completion rates. We further complement this data set with institution-level information on parental income, students' income, and students' intergenerational mobility derived from tax records, as assembled and aggregated by Chetty et al. (2019). Our final dataset includes syllabi from 4,353 institutions, including 2,941 institutions for which income and mobility data is available.

3. Measuring the Education-Innovation Gap

In this section, we describe the construction of the "Education-Innovation Gap". We describe the most important building block of this measure, textual similarities between textual documents, in Section 3.1. Section 3.2 defines the gap measure and describes its intuition. Section 3.4 provides useful validation tests. Section 3.5 discusses the methods of interpreting the economic magnitude of the measure.

3.1. Textual Similarities

The basic building block for our measure is the textual similarities between pairs of syllabus and innovation documents. The basic idea behind textual similarities is simple and intuitive. Any textual documents d can be represented in a form of term-frequency vector V_d of dimension W, W being the number of unique words in the document. In specific, for each word w, its term frequency in document d is defined as

$$TF_{dw} \equiv \frac{c_{dw}}{\sum_k c_{dk}}, \tag{1}$$

where c_dk is the number of times term k appears in document d. Intuitively, TF captures the relative importance of term w in document d adjusted for the length of the document. And the textual similarity of any pair of documents d and d' is $\frac{V_d}{\|V_d\|} \cdot \frac{V_{d'}}{\|V_{d'}\|}$. Essentially, we are capturing whether the two documents are close or far-away in the textual space.

However, a key consideration in devising a similarity metric for a pair of text documents is to appropriately weigh words by their importance. It is more informative if terms such as "machine learning" and "genome editing" enter more prominently into the similarity calculation than common words in syllabi like "assignments" and "office hours", or common words in scientific writing like "process" or "discussion".

We take care of this in two steps that are complementary to each other. The first one, we only focus on textual contents that are relevant to measuring knowledge, innovation, and labor skills, and exclude words that are likely just structural or grammatical components of language. We will use "dictionaries" to label those sets of useful words that are kept in the analysis, and we construct three dictionaries: (1) the dictionary of all unique words and terms that were ever entered as a key word in academic publications; (2) all words and terms that have an English Wikipedia webpage as of 2019; and (3) a list of nearly 20,000 knowledge points and skills compiled by Burning Glass Technologies.

Second, we use statistical methods to generate weighting matrix for the information of texts. In textual analysis, a leading approach to overweighting terms that are most diagnostic of a document's content is the "term-frequency-inverse-document-frequency (TFIDF)"

transformation of word counts:

$$TFIDF_{dw} = TF_{dw} \times IDF_w. \tag{2}$$

The second component added to the TF term, IDF is the inverse document frequency of term w, which is defined as

$$IDF_w \equiv \log(\frac{\text{\# of documents in sample}}{\text{\# of documents that include term w}}). \tag{3}$$

IDF measures the informativeness of term w by underweighing common words that appear in many documents, as these are less diagnostic of the content of any individual document.

The product of these two terms, $TFIDF$, captures the importance of a given word or phrase w in any given document d. A word tends to have a low $TFIDF$ if it appears infrequently in a document (due to low TF), or that if it appears in many documents (due to low IDF). A high value of $TFIDF_{dw}$ indicates that term w appears relatively frequently in document d but does not appear in too many other documents, thus conveying that word w is a distinguishing feature of document d's semantic content.

For our purposes, this traditional weighting scheme that uses the complete collection of documents to calculate the IDF component is not ideal because it ignores the temporal ordering technological evolution reflected in syllabi, publications, and patent filings. In particular, we are interested in the novelty of document d's text content given the history of innovation leading up to the development of d, without taking into account the future information beyond d. Consider for example Andrew Ng's famous Machine Learning course in the early 2000s (CS299 of Stanford University), which was among the first courses to devote full attention to the term "machine learning," a phrase used with great frequency particularly in the recent five years. Standard IDF would sharply de-emphasize this term in the $TFIDF$ vector representing the course because this phrase was used so intensively in later years. $TFIDF$ would therefore give a misleading portrayal of the course's core and path-breaking content.

To overcome this issue, we devise and analyze a modified version of the traditional $TFIDF$ approach. In particular, in place of (3), we instead construct a retrospective, or "point-in-time"

version of inverse document frequency. In specific, we define the "backward-IDF" of term w for document d, (denoted by $BIDF_{wp}$) as the log frequency of documents containing w in any document prior to document d's teaching or publication year. More specifically, backward-IDF is defined as:

$$BIDF_w \equiv \log(\frac{\#\text{ of documents in sample prior to } d}{\#\text{ of documents prior to } d \text{ that include term w}}). \tag{4}$$

This retrospective document frequency measure changes with time as a term becomes more or less widely used over time, giving a temporally appropriate weighting to each term.

In sum, we construct the similarity between the document pair $(d; d')$ as follows. First, for both patents we construct our modified-version of the $TFIDF$ for each term w in document $i \in d, d'$ as

$$TFBIDF_{w,d,t} = TF_{w,d} \times BIDF_{w,t}; \quad t \equiv \min(d, d') \tag{5}$$

and likewise for d'. These are arranged in a W-vector $TFBIDF_{d;t}$ where W is the size of the set union for terms in pair $(i; j)$. Next, each $TFBIDF$ vector is normalized to have unit length,

$$V_{d,t} = \frac{TFBIDF_{d,t}}{\|TFBIDF_{d,t}\|}. \tag{6}$$

Finally, we calculate the cosine similarity between the two normalized vectors:

$$\rho_{d,d'} = V_{d,t} \cdot V_{d',t}. \tag{7}$$

We build textual similarities between each syllabus and the collection of innovation documents, namely the publications, patents, and job postings, and these are constructed under different dictionaries (keywords, Wikipedia terms, and Burning Glass skills). Because $TFBIDF$ is non-negative, $\rho_{d,d'}$ lies in the interval $[0, 1]$. Documents that use the exact same set of words in the same proportion will have similarity of one, while documents with no overlapping terms have similarity of zero.

3.2. Quantify the Education-Innovation Gap

After constructing the textual similarities between any pair of syllabus and innovation document, the next step we try to quantify the gap between each course and the innovation frontier by aggregating those pair-wise similarities. Our main idea is that a syllabus that has a lower gap to innovation is one that is closer to more current innovation as opposed to prior innovation.

Our definition of the education-innovation gap thus combines both the distance to current innovation but also past innovation. We measure a syllabus's similarities to the current innovation as its textual similarities with all innovation documents appearing in some recent period, and it similarities with the past innovation as the average similarities with all innovation documents in some past period.

$$Similarity_t^{\tau}(d) = \sum_{i \in \Omega_{d,t,\tau}} \rho_{d,i}; \tag{8}$$

where $\rho_{d,i}$ is the pairwise similarity of syllabus d and an innovation document i defined in the previous section, and $\Omega_{d,t,\tau}$ denotes the set of "prior" innovation led in the τ calendar years prior to t years before d's course year. We will consider a backward-looking window of $\tau = 3$ years in our baseline quality measure, and $t = -10$ and -1 for faraway and recent innovation. That said, our results are insensitive to other window choices.

In this case, effectively we are creating two similarities between a syllabus and past innovation—one with more recent innovation and one further into the history. $Similarity_{-1}^{3}(d)$ is the average similarities between syllabus d and all innovation documents in year $-1, -2, -3$. $Similarity_{-10}^{3}(d)$ is the average similarities between syllabus d and all innovation documents in year $-12, -11, -10$. Our measure of the education-innovation gap is then defined in the following way:

$$Gap(d) \equiv \left(\frac{Similarity_{-1}(d)}{Similarity_{-10}(d)} \right)^{-1} \equiv \frac{Similarity_{-10}(d)}{Similarity_{-1}(d)}. \tag{9}$$

Intuitively, a syllabus has a lower education-innovation gap if its contents are more similar to newer technological materials captured by a high value of $Similarity_{-1}(d)$; or equivalently, if

its contents are less similar to older knowledge.

At the first glance, it might be a tempting to simply define that education-innovation gap using $Similarity_{-1}(d)$, which is the similarity to the current innovative knowledge. But this simple approach suffer several weaknesses compared to the scaled version in Eq. (9). The key motivation behind the scaling is to mitigate the noisy heterogeneities of levels of cosine similarities across different field or even across syllabus due to styles or data processing techniques. For example, different syllabi have different levels of details put into the documents, or OCR-generated documents may have lower accuracy than web-scraped ones, both of which will lead to different cosine similarity measures in levels. Effectively, we are taking out a "fixed effect" by scaling $Similarity_{-1}(d)$ using $Similarity_{-10}(d)$, thus we are mitigating the effects from unobserved heterogeneities.

3.3. Notes on Non-STEM Majors

Even though we are using terms like "technology" and "innovation," the content and meaning of those terms should be taken more broadly. Essentially, based on our methodology, new popular views in history, new writing and journalism principles and skills, new models and quantitative methodologies in economics and political sciences, are all included in the technology definition as long as they are reflected in academic publications or job postings.

3.4. Validation

Next, we would like to validate our measure and show that the gap measure indeed captures the lag between the knowledge that is taught and the knowledge that is at the frontier. To do so, we make use of the references information extracted from the syllabi. We construct a measure of average gap between the reference publication year and the year of the course taught. Intuitively, courses that teach older materials, possibly with the exception of some classic literature, should have a larger gap measure.

In Figure 2 we present the correlation between the gap measure and the average reference age of the same syllabus. For the education-innovation gap measure, we cut the sample into 25 different bins from the lowest to the highest, and plot the average catalog age in each bin. The relationship is nearly a perfect linear line, validating our measure. This graph also

partially mitigate the concern that textual information might differ due to professors' ability and willingness to use more trendy wording in describing the same underlying materials. The selection of references and reading list is less exposed to this concern.

But our measure is still advantageous compared to the reference-based measure: we cover more syllabus—Note that only 62% of the syllabi have such reference information; use more information that is not necessarily covered in books.

3.5. Interpreting the Economic Magnitude

Before going into the analysis, the next question to address is to create a framework to understand the economic magnitude of the education-innovation gap. Ultimately, one would like to be able to map a 0.01 change in the education-innovation gap into more readily quantifiable differences in the content of a course.

Given that the textual analysis nature of our methodology, it is difficult to get an answer to this problem in an analytical form. Our approach is to implement a simulation-based method to uncover the relation between the content (and miss of content) and changes in the gap measure. In specific, we first randomly choose 100,000 syllabi from the sample. Our simulation tries to compute—how many up-to-date knowledge terms to we need to add into these syllabi in order to narrow the gap. To do so, we randomly insert top new knowledge terms in the syllabi, which will naturally bring down the gap, and measure the change of the education-innovation gap as a response to the addition of these. When injecting new terms, we choose from the top 10% popular terms in the SCOPUS keywords list in year $t-1$. For each picked term, we add at a frequency at of mediam term frequency of a knowledge term in that specific syllabus. For example, for a Computer Science syllabus, whose average term frequency of CS-related terms is 30 times, we will choose to add top knowledge terms one at a time, 30 times per term. Mathematically, we switch one of the "0" components in the word frequency vector of the syllabus to non-zero.

We then recalculate the gap measure for this simulated syllabus, after add one, two, ... until 100 new knowledge terms, and this allows us to plot a correlation, for each syllabus, between the number of new terms added and the decrease of the education innovation gap. After averaging this correlation for each of the 100,000 sampled syllabi, we form a way to

interpret the magnitude of the gap measure.

This baseline graph is shown in Figure 3. Essentially, this graph means that adding ten more trendy technology terms is associated with a 0.0025 shrink of the education-innovation gap. A syllabus need to add more than 50 new technological terms to narrow the gap by 0.01. Note that on average, each syllabi includes roughly 480 key technology words, which means that to decrease the gap by 0.01, the course needs to cover roughly 10.5% more up-to-date material. This chart will be referenced back when we start performing analysis for interpretation purposes.

4. The Education-Innovation Gap Within Fields and Over Time

We begin our description of the education-innovation gap by investigating how it varies across and within fields and over time. Figure 4 (dashed line) shows the distribution of the gap across fields and years. The figure highlights a significant amount of variation in the gap across syllabi. The gap is equal to 0.964 on average; syllabi on the 25th percentile of this distribution have a gap equal to 0.927, whereas syllabi on the 75th percentile have a gap equal to 0.997. To quantify the extent of the difference between these two numbers, we follow the mapping in Figure 2 between changes in the gap and the addition of new knowledge words to a gap. The figure implies that in order to move a gap from the 75th to the 25th percentile of the distribution we would have to add 350 new knowledge words to it (73 percent compared with an average of 480 knowledge words in a syllabus).

Part of this variation in the gap, however, could be driven by differences across fields. Figure 4 (solid line) shows a distribution of gap residuals across fields, controlling for field fixed effects. Syllabi on the 25th and 75th percentile of this distribution have a gap equal to 0.929 and 0.996, respectively, which implies that in order to move a gap from the 75th to the 25th percentile within the same field we would have to add approximately 300 new knowledge words to it. This implies that, even within fields, the variation appears substantial.

4.1. Trends in The Gap Over Time

To more systematically investigate the evolution of the gap over time across and within fields, in Figure 5 (*baseline* series) we plot the change in the average gap between 1992 and

2018, relative to the 1992 level. Two things emerge from this figure. First, the gap declined substantially between 1992 and 2001, and it plateaued between 2001 and 2007 at a level 0.03 smaller than 1992 (equivalent to 170 additional knowledge words). Second, the gap started to rapidly increase between 2007 and 2017, becoming 0.008 higher than in 1992 (equivalent to 45 fewer additional knowledge words).

To investigate whether this time pattern is driven by changes in the sample of syllabi over time, we also re-estimate this trend by controlling for field fixed effects (Figure 5,*within field*). The estimates are essentially unchanged, indicating that these trends are not driven by compositional changes in our sample.

5. The Education-Innovation Gap Across Schools

In this section we explore how the education-innovation gap varies at the university level, to document differences in access to university courses with smaller education-innovation gaps among students. Given that students with different demographic characteristics tend to enroll in different types of schools (Chetty et al., 2019), our focus is on relationship between the gap and a set of school characteristics: Selectivity tiers, parental income of enrolled students, and the racial and ethnic makeup of the student population. The goal of this exercise is to test whether students from low-income families have differential access to education with up-to-date content.

5.1. School Tier

We begin our analysis documenting how the gap varies across school tiers. Schools are classified into tiers based on their selectivity, control (public/private), and the degrees offered. Following Chetty et al. (2019), we use six tiers: "Ivy Plus," which include Ivy League schools plus Chicago, Duke, MIT, and Stanford; "Elite," which include 65 schools in Tier 1 of Barron's 2009 ranking of selectivity (excluding the Ivy-Plus); "Highly selective/selective" public and private, which include schools in Barron's Tiers 2 to 5; "Non-selective," which includes Barron's Tier 9 and all four-year colleges not included in the Barron's selectivity index; and "Two-years," which includes all schools offering at most an Associate degree.

To compare the gap across different school tiers, we estimate the parameters of the following equation via OLS:

$$\begin{aligned} Gap_i = &\beta^{Ivy} Ivy_{s(i)} + \beta^{Elite} Elite_{s(i)} + \beta^{SelPub} SelPub_{s(i)} + \beta^{SelPriv} SelPriv_{s(i)} \\ &+ \beta^{Non-sel} NonSel_{s(i)} + \beta^{TwoYr} TwoYr_{s(i)} + \phi_{f(i)} + \tau_{t(i)} + \varepsilon_i \end{aligned} \tag{10}$$

where Gap_i measures the gap between syllabus i, taught in school $s(i)$ in year $t(i)$, and all academic publications. The indicator variables $Ivy_{s(i)}$, $Elite_{s(i)}$, $SelPub_{s(i)}$, $SelPriv_{s(i)}$, $NonSel_{s(i)}$, and $TwoYr_{s(i)}$ equal one if school $s(i)$ is in the Ivy Plus, Elite, Highly selective/selective public or private, Non-selective, or Two-years tier, respectively. A vector of field fixed effects $\phi_{f(i)}$ controls for field-specific differences across fields, to account for the fact that schools in different tiers could be more likely to offer courses in certain areas. Lastly, a vector of year fixed effects $\tau_{t(i)}$ controls flexibly for aggregate trends in the gap over time.

Point estimates of the coefficients on the indicator variables in equation (10) represent conditional averages for schools in each field, and are shown, along with confidence intervals, in Figure 6. Two striking pattern emerge. First, the education-innovation gap for courses offered by Ivy Plus schools is significantly smaller than that of schools in any other tier, and equal to 0.951 on average. Second, no significant difference exists in the gap of courses offered by schools in the other tiers, including elite. Elite schools have a gap of 0.961; highly selective/selective public and private and two-years schools all have a larger 0.964 gap, not significantly distinguishable from the gap of elite schools. Non-selective schools have a smaller 0.957 gap, but indistinguishable from that of other, non-Ivy Plus schools.

To interpret the economic magnitude of these differences we refer back to Figure 3. A 0.012 difference in the gap between Ivy Plus schools and other schools implies that, in order to bring other schools to the level of Ivy Plus, one would have to add 70 new knowledge words to each syllabus, which corresponds to a 14.6 percent increase compared with an average of 480. In other words, students in schools outside of the Ivy Plus are exposed to 14.6 percent fewer knowledge words per course on average.

The evidence in Figure 6 is confirmed by the estimates shown in Table 1, where we re-estimate equation (10) grouping highly selective/selective public and private schools together and using non-selective and two-years schools as the reference tier. In column 1 we omit field

fixed effects: relative to an average of 0.97 for non-selective and two-years schools, the gap is 0.012 smaller for Ivy Plus (or 70 new words, significant at 1 percent), 0.005 smaller for other elite schools (27 new words, significant at 10 percent), and not statistically different for other highly selective and selective schools . This estimate is also robust to controlling for field and field-by-year fixed effects (Table 1, columns 2 and 3).

5.2. Parental Income

5.2.1. Median Income. We study how the gap varies across schools serving students from different socio-economic backgrounds. We begin by using median parental income at the school level as a proxy for family background. A simple binned scatterplot of the gap and median parental income at the school level shows a negative relationship between these two variables, with a slope coefficient equal to -0.0005 (significant at the 5 percent level, Figure 7). This implies that a $10,000 higher average parental income is associated with the average syllabus in the school containing 25 new knowledge words, or 5 percent compared with an average of 480.

To better investigate how the gap varies across students with different backgrounds, we divide schools in five bins depending on their percentile in the distribution of median parental income across all schools, grouping together those in the bottom 25 percent, 25-50 percent, 50-75 percent, 75-95 percent, and top 5 percent. We then estimate a specification similar to equation (10), using indicators for bins of parental income in lieu of indicators for school tiers.

Point estimates and confidence intervals of the coefficients in this specification, shown in Figure 8, show no significant differences in the gap across schools with different median parental income, except for schools in the top five percent of the distribution. For these schools, the gap is equal to 0.955, compared with 0.964 for schools in the 75-95 percent, 0.962 for schools in the 50-75 percent, and 0.963 for schools in the 25-50 and bottom 25 percent respectively. These estimates imply that, in order to make other schools comparable to schools in the top five percent, one would have to add approximately 50 new knowledge words to the average syllabus, or 10 percent.

These patterns are confirmed by the estimates in Table 2 (columns 1-3), where we use

schools in the bottom 25 percent as the reference group. Schools in the top 5 percent have a 0.008 percent smaller gap (10 percent more knowledge words), and no significant differences exist among the other groups. These estimates are robust to controlling for field and year fixed effects (column 2) and field-by-year fixed effects (column 3).

5.2.2. Share of Parents in Top Income Percentile. To capture an even more extreme measure of inequality in access to high-quality education, we also investigate whether the gap varies across institutions serving different shares of students with parental income in the top 1 percent of the distribution. Figure 9 reveals a strong, negative correlation between the gap and the share of students whose parents have income in the top 1 percent.

As before, we further investigate this relationship by comparing the gap across schools with different percentages of students with parents in the top income percentile. Estimates and confidence intervals, shown in Figure 10, reveals a striking pattern: schools where no students have parents in the top percentile have a much larger gap compared with the remaining schools, and equal to 0.998 (compared with a gap of 0.957 for schools with 10 percent or more students with parents in the top quintile, or a 4.3 percent difference). This pattern is confirmed by the estimates in Table 2 (columns 3-6). These estimates imply that, to make the gap schools where no students have parents in the top 1 percent equal to the gap in all other schools, one would have to add more than 220 new knowledge words to it, or 46 percent.

5.3. Students' Race/Ethnicity

Lastly, we investigate whether schools serving a larger portion of minority students have significantly different gaps. We define minority students as those who are either Black and Hispanic. A binned scatterplot of the gap with the share of minority students in each school reveals a positive relationship between these two variables, although imprecise (with a slope coefficient equal to 0.058 and a T-statistic of 1.15, Figure 11).

To better explore how access to university courses with smaller gaps varies across students of different races and ethnicities, we divide schools in five bins depending on their share of minority students and we estimate a specification similar to Eq. (10), using indicators

for bins of parental income as independent variables. Estimates and 95 percent confidence intervals of the coefficients, shown in Figure 12, show no significant differences in access to a lower gap by race or ethnicity, with one exception: Schools with a population of students who are more than 80 percent minority. For these schools, the gap is equal to 0.976, compared with 0.964 for schools with 40-80 percent minority, 0.961 for schools with 20-40 percent, and 0.963 and 0.962 for schools with 10-20 percent and less than 10 percent respectively. These estimates imply that, in order to make the gap in schools with more than 80 percent minority students equal to the gap in all other schools, we would have to add 70 new knowledge words to the average syllabus, or 480 percent.

These patterns are confirmed by the estimates in Table 3, where we use schools with less than 10 percent of students who are minority as the reference group. Schools with more than 80 percent minority students have a 0.015 larger gap (or 1.6 percent compared with an average of 0.96 for the reference group), and no significant differences exist among the other groups. These estimates are robust to controlling for field and year fixed effects (column 2) and field-by-year fixed effects (column 3).

6. Student Outcomes and the Education-Innovation Gap

The findings in the previous section suggest significant differences in access to university courses with smaller gaps across schools in different tiers of selectivity and serving different populations of students. A natural question is whether these differences in access bear some relationship with future student outcomes. We focus our attention on graduation rates, income, and rates of intergenerational mobility.

6.1. Gap and Graduation Rates

Figure 13 shows a binned scatterplot of school-level graduation rates, measured in 2000 and 2013, and the education-innovation gap in the same years, controlling for field-by-year fixed effects. The relationship between these two variables is negative and significant: a 0.1 lower gap, equivalent to 60 additional knowledge words, is associated with a 4.1 percentage points larger graduation rate, or 8 percent compared with an average of 43 percent in 2013.

Next, we study whether this relationship varies across school tiers. To do so, we estimate

the following equation:

$$\begin{aligned} Outcome_{s(i)t} =& \beta^{Ivy} Gap_{it} * Ivy_{s(i)} + \beta^{Elite} Gap_{it} * Elite_{s(i)} + \beta^{SelPub} Gap_{it} * SelPub_{s(i)} \\ &+ \beta^{SelPriv} Gap_{it} * SelPriv_{s(i)} + \beta^{Non-sel} Gap_{it} * NonSel_{s(i)} \\ &+ \beta^{TwoYr} Gap_{it} * TwoYr_{s(i)} + \phi_{f(i)} + \tau_{t(i)} + \varepsilon_i \end{aligned} \tag{11}$$

where $Outcome_{st}$ is the student's outcome variable, i.e., $Grad_{st}$ the graduation rate of school s in year t in this subsection. In this specification, the parameters β capture the correlation between the gap and graduation rates for schools in each tier, conditional on field ($\phi_{f(i)}$) and year ($\tau_{t(i)}$) fixed effects.

Figure 14 shows point estimates and 95-percent confidence intervals of the parameters in equation (11). For Ivy Plus and elite schools, a positive correlation emerges between the gap and graduation rates. In particular, a 0.1 smaller gap (60 additional words) is associated with a 2.8 and 2.1 percentage points lower graduation rate in Ivy Plus and elite schools, respectively. In the remaining schools, on the other hand, this correlation is negative and significant; a 0.1 smaller gap is associated with a 1.4, 0.9, 2.1, and 4.9 percentage points higher graduation rate in highly-selective and selective public, private, non-selective, and two-years schools respectively. These patterns are confirmed by the estimates in Table 4 (column 3).

In Figure 15 we investigate whether the relationship between the gap and schools' graduation rates varies across percentiles of median parental income. The correlation is positive and significant for schools with parental income in the top quartile: a 0.1 smaller gap is associated with a 3.4 percentage points lower graduation rates for schools in the top 5 percent and with a 1.8 percentage points lower rate for schools in the 75th to 95th percentile. The correlation becomes progressively more negative as schools' median parental income declines; a 0.1 smaller gap is associated with a 2.4 percentage points higher graduation rate in schools in the 25 to 50 percentile and with a 3.8 percentage points higher rate for schools with parental income in the bottom quartile. These findings are summarized in Table 4 (column 4).

A positive correlation between the gap and graduation rates in top-tier schools and in

schools serving more affluent students might appear surprising at a first glance. A possible explanation for this finding is that smaller gaps expose students in more selective institutions and from affluent backgrounds to more innovative, albeit less academic, materials and content. This exposure might make them more likely to not graduate and pursue more entrepreneurial ventures instead. It is important, however, to notice how this association completely reverts in lower-tiered schools; here, on the other hand, a smaller gap is associated with better graduation chances, and especially so for schools at the bottom of the selectivity distribution.

6.2. Gap and Students' Income

Next, we turn our attention to a simple economic outcome: the median income of students who graduate from each school. Figure 16 shows a binned scatterplot of median student income, measured in 2014 for students who graduated between 2002 and 2007, and the education-innovation gap relative to the year 2004. The relationship between these two variables is negative and significant: a 0.1 lower gap is associated with a $1,850 higher income.

As before, we explore whether this relationship varies across school tiers by re-estimating Eq. (11) with the natural logarithm of students' income as the dependent variable. The point estimates and confidence intervals of the parameters in this equation, shown in Figure 17, indicate again that the correlation between the gap and students' income is positive for top-tier schools; a 0.1 smaller gap is associated with a 5.5 and 2.1 percent lower income in Ivy Plus and elite schools, respectively. As before, this correlation is instead negative and significant in all other schools; a 0.1 smaller gap is associated with a 1.5, 1.9, 6.0, and 6.3 percent higher income for students in highly-selective and selective public, private, non-selective, and two-years schools respectively. These patterns are confirmed by the estimates in Table 5 (column 3).

In Figure 18 we explore whether the relationship between the gap and students' incomes varies across percentiles of median parental income. The correlation is positive and significant for schools with parental income in the top quartile: a 0.1 smaller gap is associated with a 1.5 percent lower income for schools in the top 5 percent and with a 0.8 lower income for schools in the 75th to 95th percentile. The correlation becomes progressively more negative as schools' median parental income declines; a 0.1 smaller gap is associated with a 3.5 percent

higher income in schools in the 25 to 50 percentile and with a 5.5 percent higher income for schools with parental income in the bottom quartile. These findings are summarized in Table 5 (column 4).

6.3. Gap and Intergenerational Income Mobility

Lastly, we study whether a school's education-innovation gap is related to the economic opportunities of the low-income students it serves. To do so, we explore the correlation between the gap and a simple measure of intergenerational mobility: The probability that a student with parental income in the bottom quintile reaches the top quintile during adulthood. Figure 19 shows a binned scatterplot of this measure of mobility, measured using 2014 incomes for students who graduated between 2002 and 2007, and the education-innovation gap relative to the year 2004. The relationship between these two variables is negative and significant: a 0.1 lower gap is associated with a 2 percentage points lower probability, or 10 percent compared with an average probability of 0.2.

Estimating this correlation separately by school tiers varies across school tiers reveal that the correlation is positive for top-tier schools; a 0.1 smaller gap is associated with a 1.7 and 0.8 percentage points higher mobility in Ivy Plus and elite schools, respectively (Figure 20). Again, the correlation becomes progressively more negative as we move down the selectivity ranking. A 0.1 smaller gap is associated with a 1.1, 1.3, 2.9, and 3.2 percentage points higher mobility for students in highly-selective and selective public, private, non-selective, and two-years schools respectively. These patterns are confirmed by the estimates in Table 6 (column 3).

In Figure 21 we explore whether the relationship between the gap and students' incomes varies across percentiles of median parental income. The correlation is positive and significant for schools with parental income in the top 5 percent, and it becomes progressively more negative as schools' median parental income declines. These findings are summarized in Table 6 (column 4).

7. Conclusion

This paper defines the "education-innovation gap" as the gap between course materials in universities and the frontier knowledge. We measure this gap using information on the text of 3 million university syllabi taught in nearly three decades and 20 million academic publications. We construct the gap by comparing the textual similarity between each gap and each article published in a top 10 journal in each field and year, since each journal's foundation. Syllabi have a larger education-innovation gap if they are more similar to older knowledge, compared to newer one. A significant amount of variation in the gap exist across university courses, and both across and within fields. In addition, the gap has increased over time in the last decade. We document large variation in access to small-gap education across schools and students with different backgrounds. For example, students in schools outside of the Ivy Plus and elite, in schools with lower average parental income, and in schools serving more minority students all have access to courses with a significantly larger gap. We also find that the gap is related to students' outcomes such as graduation rates, income, and mobility, even controlling for other school characteristics. Overall, we argue that the education-innovation gap is can be an important tool to study the formation of human capital, skilled labor, innovation, and inequality.

References

Acemoglu, Daron, and David Autor, 2012, What does human capital do? a review of goldin and katz's the race between education and technology, Journal of Economic Literature 50, 426.

Altonji, Joseph G, Erica Blom, and Costas Meghir, 2012, Heterogeneity in human capital investments: High school curriculum, college major, and careers, Annu. Rev. Econ. 4, 185–223.

Baumol, William J, 2005, Education for innovation: Entrepreneurial breakthroughs versus corporate incremental improvements, Innovation policy and the economy 5, 33–56.

Benhabib, Jess, and Mark M Spiegel, 2005, Human capital and technology diffusion, Handbook of economic growth 1, 935–966.

Chetty, Raj, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan, 2019, Income segregation and intergenerational mobility across colleges in the united states, NBER Working Paper .

Dale, Stacy, and Alan B Krueger, 2011, Estimating the return to college selectivity over the career using administrative earnings data, NBER Working Paper .

Deming, David, and Lisa B Kahn, 2018, Skill requirements across firms and labor markets: Evidence from job postings for professionals, Journal of Labor Economics 36, S337–S369.

Deming, David J, and Kadeem L Noray, 2019, Stem careers and the changing skill requirements of work, NBER Working Paper .

Gentzkow, Matthew, and Jesse M Shapiro, 2010, What drives media slant? evidence from us daily newspapers, Econometrica 78, 35–71.

Goldin, Claudia, and Lawrence F Katz, 1996, Technology, skill, and the wage structure: insights from the past, The American Economic Review 86, 252–257.

Goldin, Claudia Dale, and Lawrence F Katz, 2008, The Race Between Education and Technology (Harvard University Press).

Hanushek, Eric A, Jens Ruhose, and Ludger Woessmann, 2017, Knowledge capital and aggregate income differences: Development accounting for us states, American Economic Journal: Macroeconomics 9, 184–224.

Hanushek, Eric A, and Ludger Woessmann, 2012, Do better schools lead to more growth? cognitive skills, economic outcomes, and causation, Journal of economic growth 17, 267–321.

Hershbein, Brad, and Lisa B Kahn, 2018, Do recessions accelerate routine-biased technological change? evidence from vacancy postings, American Economic Review 108, 1737–72.

Hoxby, Caroline M, 1998, The return to attending a more selective college: 1960 to the present, Unpublished manuscript, Department of Economics, Harvard University, Cambridge, MA .

Jones, Benjamin F, 2009, The burden of knowledge and the death of the renaissance man: is innovation getting harder?, Review of Economic Studies 76, 283–317.

Jones, Benjamin F, 2010, Age and great invention, The Review of Economics and Statistics 92, 1–14.

Jones, Benjamin F, and Bruce A Weinberg, 2011, Age dynamics in scientific creativity, Proceedings of the National Academy of Sciences 108, 18910–18914.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy, 2018, Measuring technological innovation over the long run, NBER Working Paper .

Nelson, Richard R, and Edmund S Phelps, 1966, Investment in humans, technological diffusion, and economic growth, American Economic Review 56, 69–75.

Romer, Paul M, 1990, Endogenous technological change, Journal of political Economy 98, S71–S102.

Table 1. Gap and School Tiers. Dependent Variable is Gap in Cosine Similarities Between Syllabi and Publications at $t-15$ vs $t-1$

	(1)	(2)	(3)
School tier			
Ivy Plus	-0.012***	-0.011***	-0.011***
	(0.003)	(0.002)	(0.002)
Other elite	-0.005*	-0.001	-0.001
	(0.002)	(0.002)	(0.002)
Highly selective/selective	-0.001	0.002	0.001
	(0.002)	(0.001)	(0.001)
Field FE	No	Yes	No
Syll. year FE	Yes	Yes	No
Field*year FE	No	No	Yes
Mean of dep. var. for non-selective	0.97	0.97	0.97
N	1758736	1758736	1758710
# Inst-by-field	29842	29842	29835

Note: The table shows OLS estimates of the coefficients for indicators for school tiers (*Ivy Plus, Other elite, Highly selective/selective*) on the gap between syllabi and publications. The omitted category is *Non-selective and two-years* institutions. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained pooling data for the years 1992 to 2018. Column 1 controls for year fixed effects, column 2 controls for field and year fixed effects, and column 3 controls for field-by-year fixed effects. Standard errors in parentheses are clustered at the school-by-field level. Information on school tiers is taken from Chetty et al. (2019). $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

Table 2. Gap and Schools' Median Parental Income. Dependent Variable is Gap in Cosine Similarities Between Syllabi and Publications at $t-15$ vs $t-1$

	(1)	(2)	(3)	(4)	(5)	(6)
Median income						
25-50%	0.001	-0.000	-0.000			
	(0.003)	(0.002)	(0.002)			
50-75%	-0.005	-0.002	-0.002			
	(0.003)	(0.002)	(0.002)			
75-95%	-0.000	0.001	0.001			
	(0.003)	(0.002)	(0.002)			
top 5%	-0.008***	-0.008***	-0.008***			
	(0.003)	(0.002)	(0.002)			
% parents in top 1%						
<0.5%				-0.030***	-0.027***	-0.026***
				(0.006)	(0.005)	(0.005)
0.5-1%				-0.029***	-0.023***	-0.022***
				(0.007)	(0.005)	(0.005)
1-5%				-0.031***	-0.025***	-0.025***
				(0.006)	(0.005)	(0.005)
5-10%				-0.029***	-0.024***	-0.023***
				(0.006)	(0.005)	(0.005)
>10%				-0.036***	-0.031***	-0.031***
				(0.006)	(0.005)	(0.005)
Field FE	No	Yes	No	No	Yes	No
Syll. year FE	Yes	Yes	No	Yes	Yes	No
Field*year FE	No	No	Yes	No	No	Yes
Dep. var. mean, median <25%	0.97	0.97	0.97			
Dep. var. mean, share in top1 = 0				0.99	0.99	0.99
N	1758736	1758736	1758710	1758736	1758736	1758710
# Inst-by-field	29842	29842	29835	29842	29842	29835

Note: The table shows OLS estimates of the coefficients for indicators for percentiles of median parental income in each school (columns 1-3) and for the share of parents in the top percentile of the income distribution (columns 4-6) on the gap between syllabi and publications. The omitted category is *bottom 25%* in columns 1-3, and = 0 in columns 4-6. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Percentiles of parental incomes are calculated using the distribution of median parental incomes across all schools. Estimates are obtained pooling data for the years 1992 to 2018. Column 1 controls for year fixed effects, column 2 controls for field and year fixed effects, and column 3 controls for field-by-year fixed effects. Standard errors in parentheses are clustered at the school-by-field level. Information on parents' income is taken from Chetty et al. (2019). * ≤ 0.1, ** ≤ 0.05, *** ≤ 0.01.

Table 3. Gap and Students' Race/Ethnicity. Dependent Variable is Gap in Cosine Similarities Between Syllabi and Publications at $t-15$ vs $t-1$

	(1)	(2)	(3)
10-20% minority	0.001	0.001	0.001
	(0.002)	(0.001)	(0.001)
20-40% minority	-0.002	-0.001	-0.001
	(0.002)	(0.002)	(0.001)
40-60% minority	0.003	0.001	0.001
	(0.003)	(0.003)	(0.003)
60-80% minority	0.000	0.001	0.001
	(0.002)	(0.002)	(0.002)
>80% minority	0.015***	0.013***	0.013***
	(0.003)	(0.002)	(0.002)
Field FE	No	Yes	No
Syll. year FE	Yes	Yes	No
Field*year FE	No	No	Yes
Mean of dep. var. for bottom 10%	0.96	0.96	0.96
N	1757893	1757893	1757867
# Inst-by-field	29739	29739	29732

Note: The table shows OLS estimates of the coefficients for indicators for intervals of the share of students who are either Black or Hispanic in each school. The omitted category is schools with $<$ *10%* Black or Hispanic students. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained pooling data for the years 1992 to 2018. Column 1 controls for year fixed effects, column 2 controls for field and year fixed effects, and column 3 controls for field-by-year fixed effects. Standard errors in parentheses are clustered at the school-by-field level. Information on students' race and ethnicity at the school level is taken from Chetty et al. (2019). * ≤ 0.1, ** ≤ 0.05, *** ≤ 0.01.

Table 4. Graduation Rate and Gap with Publications. Dependent Variable is Average School Graduation Rate in 2002 and 2013

	Dep. Var: Graduation rate			
	(1)	(2)	(3)	(4)
Gap	-0.411***	-0.413***		
	(0.062)	(0.059)		
Gap * Non selective			-0.484***	
			(0.035)	
Gap * Highly selective/selective			-0.131***	
			(0.036)	
Gap * Other elite			0.206***	
			(0.037)	
Gap * Ivy Plus			0.276***	
			(0.037)	
Gap * <25pctile				-0.378***
				(0.033)
Gap * 25-50th pctile				-0.244***
				(0.036)
Gap * 50-75th pctile				-0.027
				(0.034)
Gap * 75-95th pctile				0.180***
				(0.036)
Gap * ¿5th pctile				0.341***
				(0.036)
Field FE	Yes	No	Yes	Yes
Syll. year FE	Yes	No	Yes	Yes
Field*year FE	No	Yes	No	No
Mean of dep. var.	0.43	0.43	0.43	0.43
N	164379	164378	164379	164379
# Inst-by-field	12728	12728	12728	12728

Note: The table shows OLS estimates of specifications where the dependent variable is a school's graduation rate, measured in 2002 and in 2013. The variable *Gap* is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. The variables *Ivy Plus, Other elite, Highly selective/selective*, and *Non selective* denote school tiers. The variables *bottom 25%, 25-50%, 50-75%, 75-95%*, and *top 5%* denote percentiles of parental incomes, calculated using the distribution of median parental incomes across all schools. Estimates are obtained pooling data for the years 2002 and 2013. Columns 1, 3, and 4 controls for field and year fixed effects, while column 2 controls for field-by-year fixed effects. Standard errors in parentheses are clustered at the school-by-field level. School-level information is taken from Chetty et al. (2019). * ≤ 0.1, ** ≤ 0.05, *** ≤ 0.01.

Table 5. Median Income and Gap with Publications. Dependent Variable is Median Student Income in 2011, By School

	Dep. Var: ln(Income)			
	(1)	(2)	(3)	(4)
Gap	-0.446***	-0.429***		
	(0.091)	(0.067)		
Gap * Non selective			-0.636***	
			(0.030)	
Gap * Highly selective/selective			-0.176***	
			(0.031)	
Gap * Other elite			0.205***	
			(0.033)	
Gap * Ivy Plus			0.491***	
			(0.032)	
Gap * ¡25pctile				-0.378***
				(0.033)
Gap * 25-50th pctile				-0.244***
				(0.036)
Gap * 50-75th pctile				-0.027
				(0.034)
Gap * 75-95th pctile				0.180***
				(0.036)
Gap * ¿5th pctile				0.341***
				(0.036)
Field FE	No	No	Yes	Yes
N	32498	32498	32498	164379
# Inst-by-field	7949	7949	7949	12728

Note: The table shows OLS estimates of specifications where the dependent variable is the average income of students who graduated from each school between 2002 and 2007, measured in 2014. The variable *Gap* is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. The variables *Ivy Plus, Other elite, Highly selective/selective*, and *Non selective* denote school tiers. The variables *bottom 25%, 25-50%, 50-75%, 75-95%*, and *top 5%* denote percentiles of parental incomes, calculated using the distribution of median parental incomes across all schools. Estimates are obtained using syllabus data for 2004. Columns 1, 3, and 4 controls for field and year fixed effects, while column 2 controls for field-by-year fixed effects. Standard errors in parentheses are clustered at the school-by-field level. School-level information is taken from Chetty et al. (2019). $^{*} \leq 0.1$, $^{**} \leq 0.05$, $^{***} \leq 0.01$.

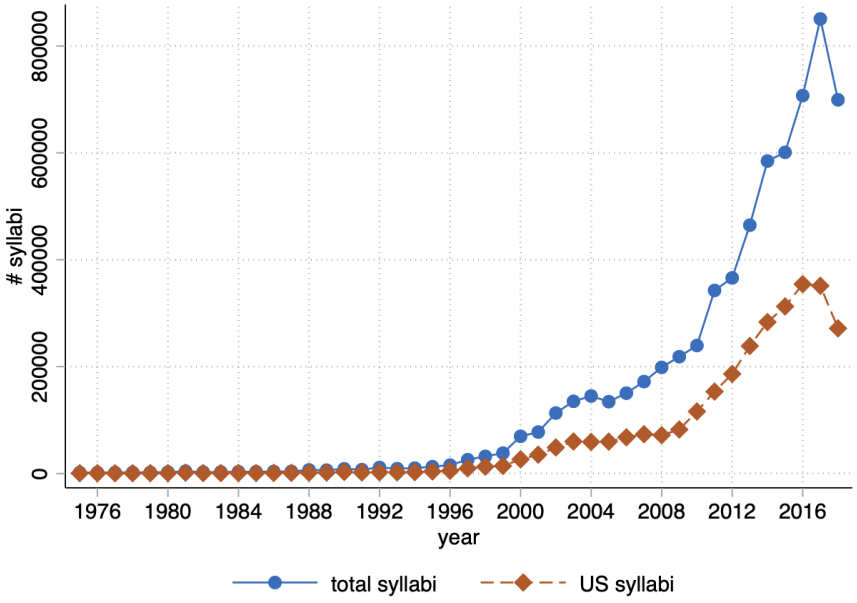
Table 6. Intergenerational Mobility and Gap with Publications. Dependent Variable is P(Child in top income quintile—Parent in bottom quintile) in 2011, By School

	Dep. Var: Mobility rate			
	(1)	(2)	(3)	(4)
Gap	-0.230***	-0.229***		
	(0.047)	(0.033)		
Gap * Non selective			-0.317***	
			(0.020)	
Gap * Highly selective/selective			-0.121***	
			(0.020)	
Gap * Other elite			0.081***	
			(0.021)	
Gap * Ivy Plus			0.169***	
			(0.021)	
Gap * <25pctile				-0.287***
				(0.020)
Gap * 25-50th pctile				-0.192***
				(0.022)
Gap * 50-75th pctile				-0.072***
				(0.020)
Gap * 75-95th pctile				0.025
				(0.021)
Gap * >5th pctile				0.148***
				(0.021)
Field FE	No	No	Yes	Yes
Syll. year FE	No	No	Yes	Yes
Field*year FE	No	Yes	No	No
Mean of dep. var.	0.30	0.30	0.30	0.30
N	32498	32498	32498	32498
# Inst-by-field	7949	7949	7949	7949

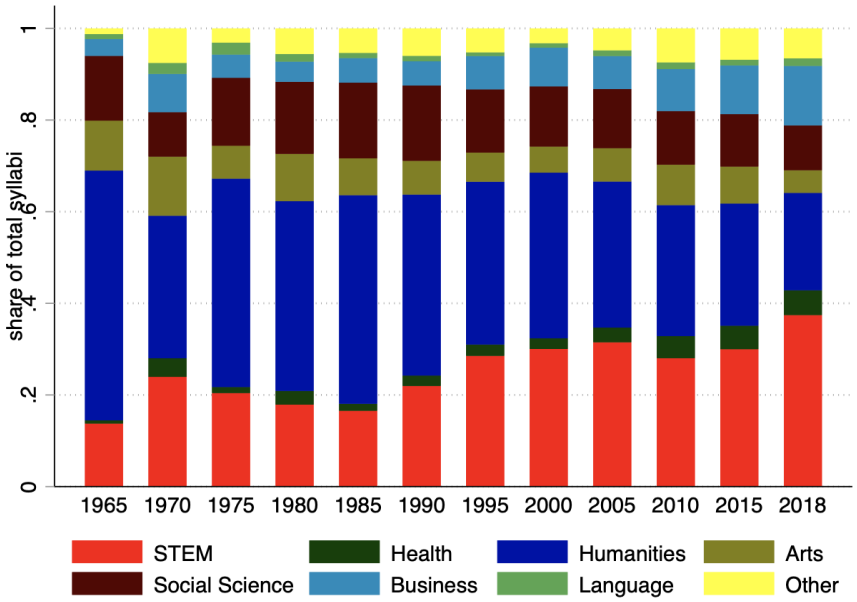
Note: The table shows OLS estimates of specifications where the dependent variable is a measure of intergenerational mobility for students who graduated in each school between 2002 and 2007, defined as the probability that a student with parents in the top 20 percent of the national income distribution reaches the top 20 percent of the income distribution and measured in 2014. The variable *Gap* is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. The variables *Ivy Plus, Other elite, Highly selective/selective*, and *Non selective* denote school tiers. The variables *bottom 25%, 25-50%, 50-75%, 75-95%*, and *top 5%* denote percentiles of parental incomes, calculated using the distribution of median parental incomes across all schools. Estimates are obtained using syllabus data for 2004. Columns 1, 3, and 4 controls for field and year fixed effects, while column 2 controls for field-by-year fixed effects. Standard errors in parentheses are clustered at the school-by-field level. School-level information is taken from Chetty et al. (2019). * ≤ 0.1, ** ≤ 0.05, *** ≤ 0.01.

Figure 1. Data Coverage

Panel a) Number of Syllabi Over time

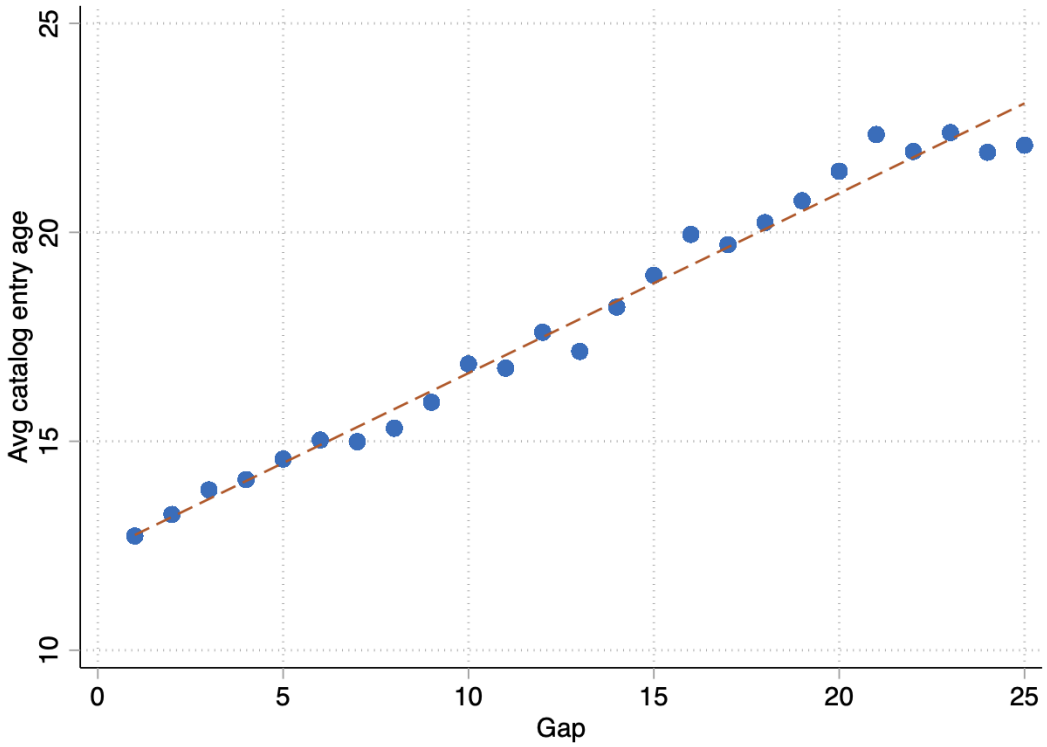


Panel b) Field Composition



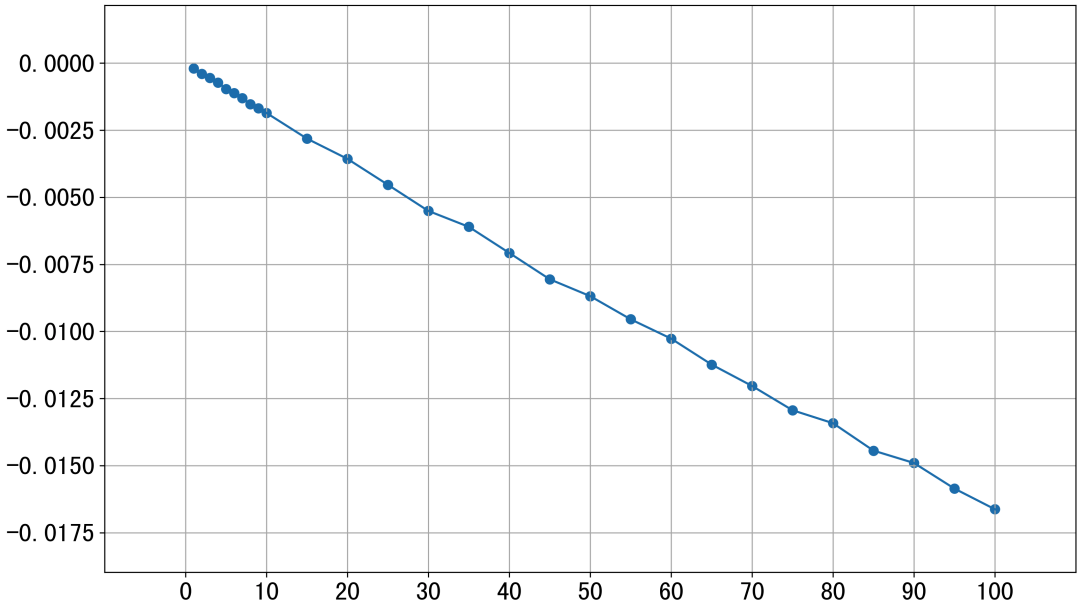
Note: The top panel shows the number of syllabi relative to each academic year, for all institutions and for US institutions. The bottom panel shows the field composition of syllabi in each five-year period.

Figure 2. Validation



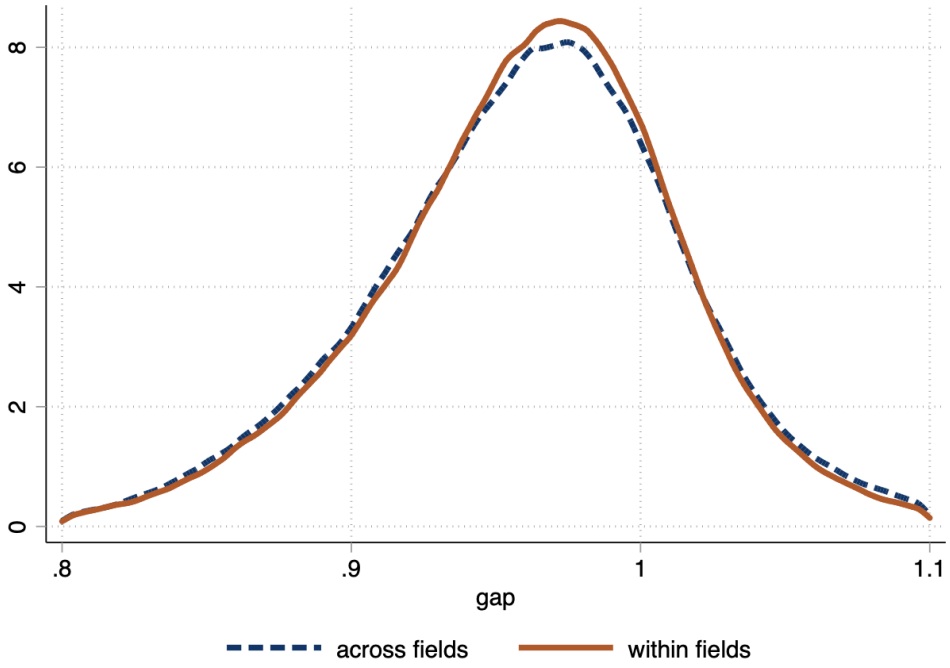
Note: This paper presents the correlation between the gap measure and the average reference age of the same syllabus. To do so, we make use of the references information extracted from the syllabi. We construct a measure of average gap between the reference publication year and the year of the course taught. For the education-innovation gap measure, we cut the sample into 25 different bins from the lowest to the highest, and plot the average catalog age in each bin.

Figure 3. Economic Magnitude



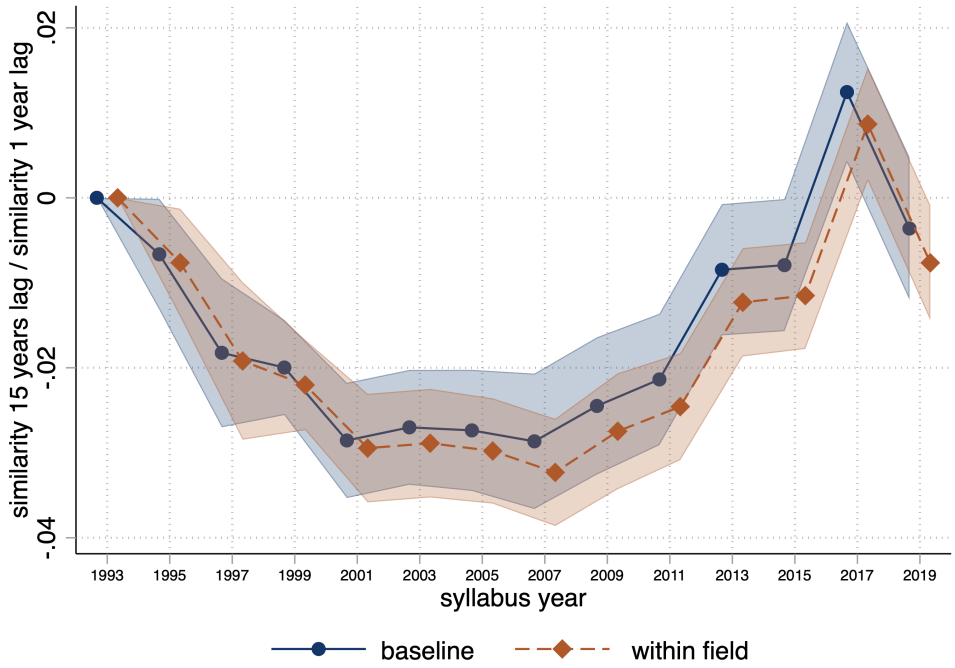
Note: This figure links the differences in education-innovation gap with economically interpretable quantities of skills in a syllabus. Specifically, we implement a simulation-based method to umcover the relation between the content (and miss of content) and changes in the gap measure. We randomly choose 100,000 syllabi from the sample, and randomly inject new technological terms in the syllabi and measure the change of the education-innovation gap induced by this change. When injecting new technologies, we choose to add one the the top 10% popular terms in technology dictionaries in year $t-1$, and we do so at a frequency of the median word count. For example, for a Computer Science syllabus, whose average term frequency of CS-related terms is 30 times, we will choose to add technologies one at a time, 30 times per term. Mathematically, we switch one of the "0" components in the word frequency vector of the syllabus to non-zero. We then recalculate the gap measure for this simulated syllabus, after add one, two, ... until 100 new technological terminologies, and this allows us to plot a correlation, for each syllabus, between the number of new terms added and the decrease of the education innovation gap.

Figure 4. Education-Innovation Gap, By Field - Selected Fields



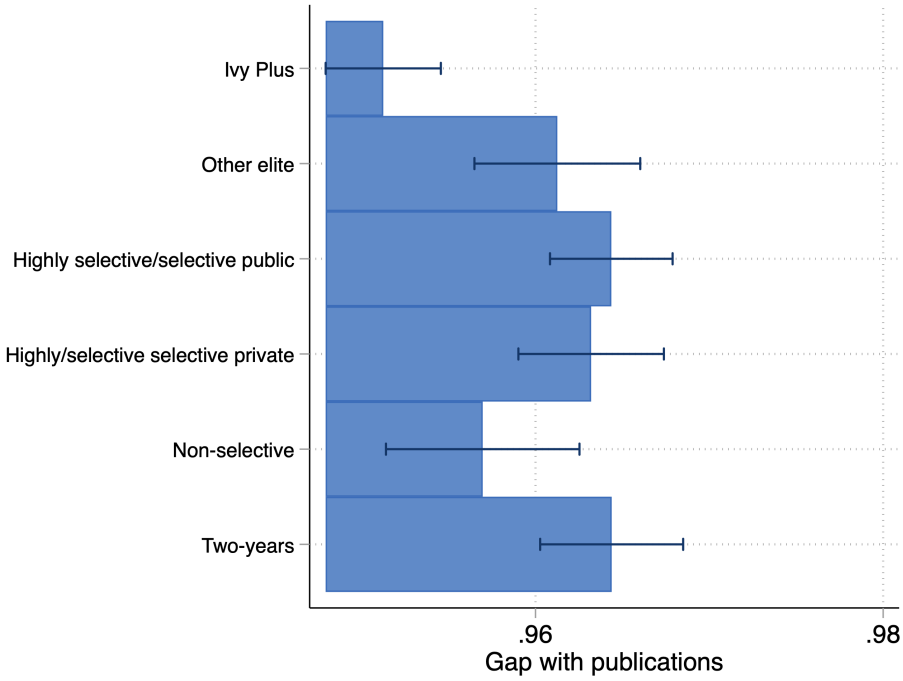
Notes: The figure shows the distribution of the gap, overall (dashed line) and within field (solid line). The within-field gap is obtained as the residual from a regression of the gap on field fixed effects.

Figure 5. Education-Innovation Gap, Over Time



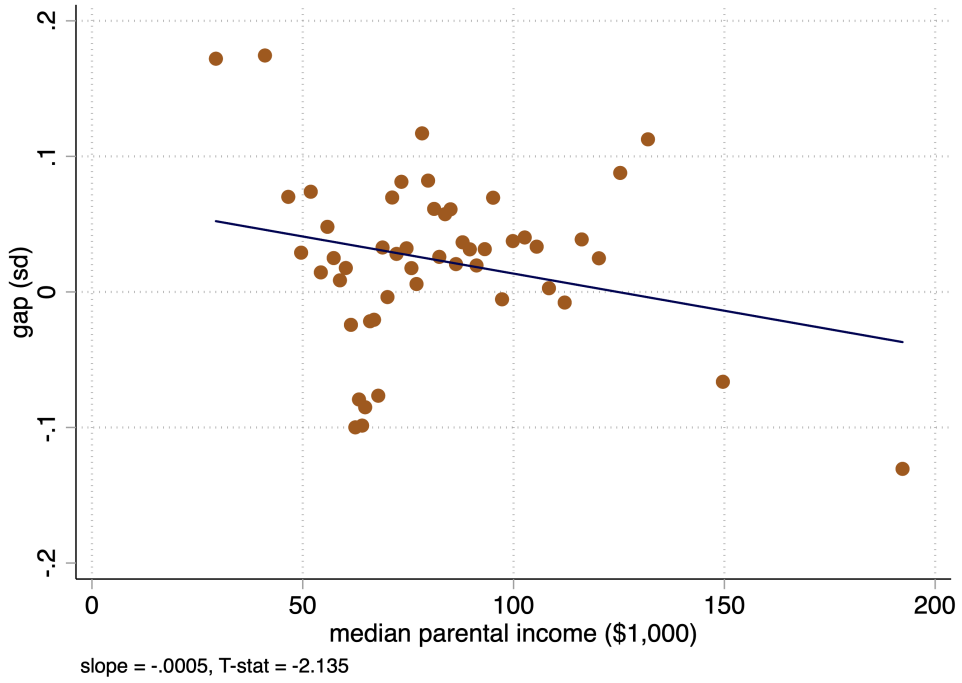
Notes: The figure shows the average gap between syllabi and publications by field, between 1992 and 2018 and relative to 1992. In the *within field* series we control for field fixed effects.

Figure 6. Gap with Publications, By School Tier



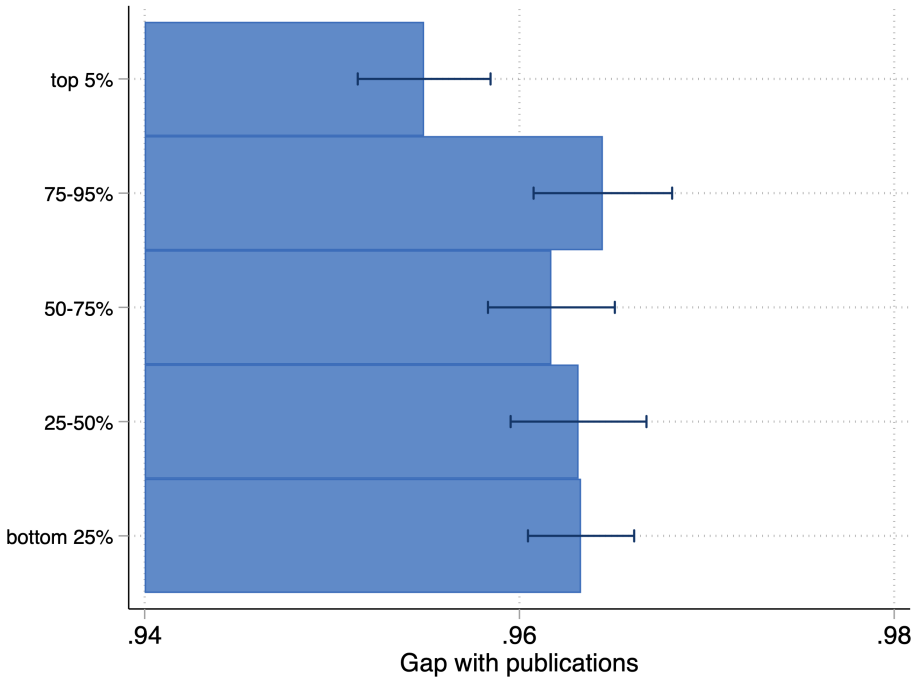
Notes: The figure shows averages and 95-percent confidence intervals of the gap between syllabi and publications by school tier. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained pooling data for the years 1992 to 2018, and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on school tiers is taken from Chetty et al. (2019).

Figure 7. Gap with Publications And Schools' Median Parental Income



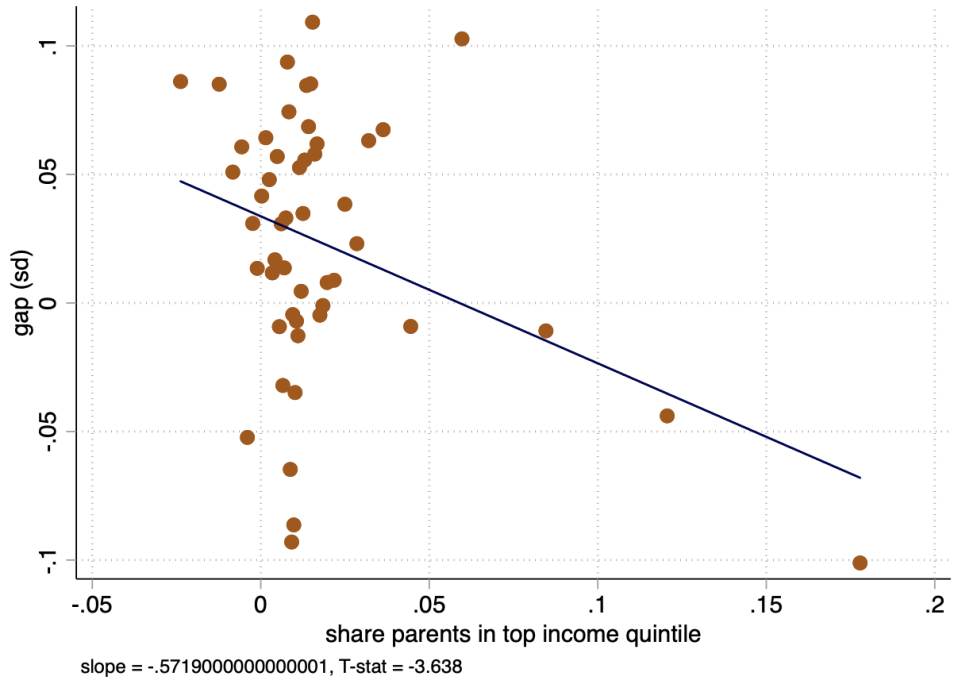
Notes: The figure shows a binned scatterplot of the gap between syllabi and publications and parental income in the school. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Percentiles are calculated using the distribution of median parental incomes across all schools. Estimates are obtained pooling data for the years 1992 to 2018, and controlling for field and syllabus year fixed effects. Information on median parental income at the school level is taken from Chetty et al. (2019).

Figure 8. Gap with Publications, By Schools' Median Parental Income



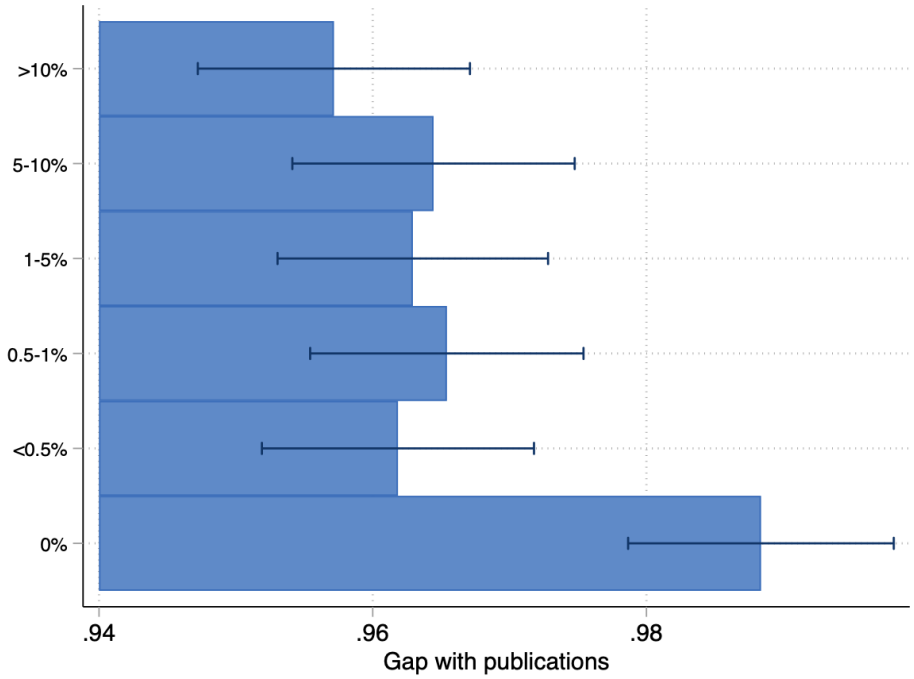
Notes: The figure shows averages and 95-percent confidence intervals of the gap between syllabi and publications by percentile of median parental income in the school. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Percentiles are calculated using the distribution of median parental incomes across all schools. Estimates are obtained pooling data for the years 1992 to 2018, and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on median parental income at the school level is taken from Chetty et al. (2019).

Figure 9. Gap with Publications And Schools' Share of Students With Parental Income In The Top 1 Percent



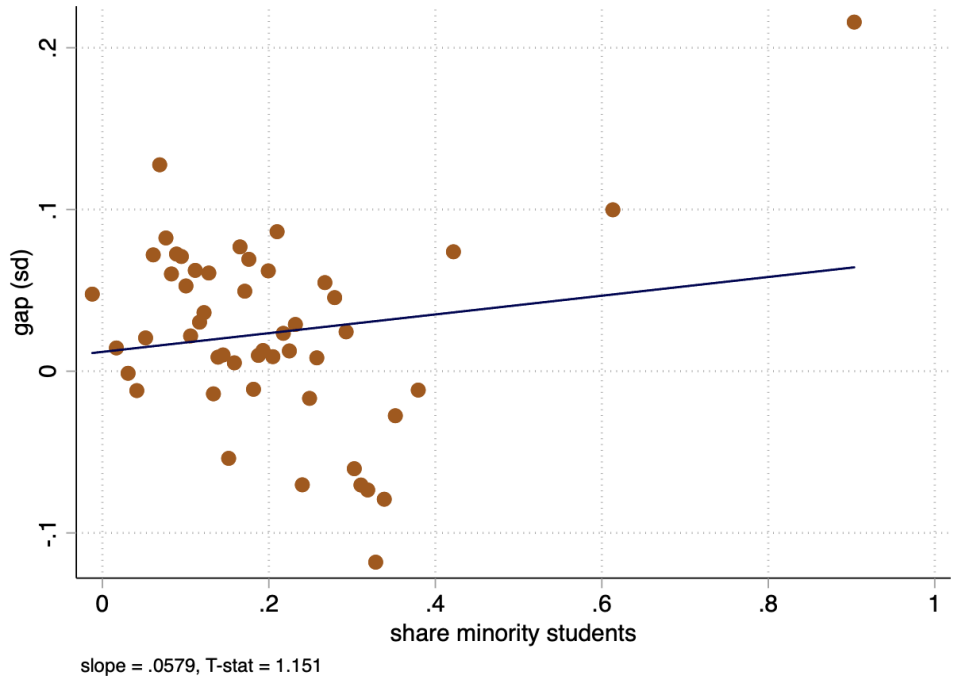
Notes: The figure shows a binned scatterplot of the gap between syllabi and publications and the share students with parents in the top income percentile in the school. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Percentiles are based on the national income distribution. Estimates are obtained pooling data for the years 1992 to 2018, and controlling for field and syllabus year fixed effects. Information on parental income at the school level is taken from Chetty et al. (2019).

Figure 10. Gap with Publications, By Schools' Share of Students With Parental Income In The Top 1 Percent



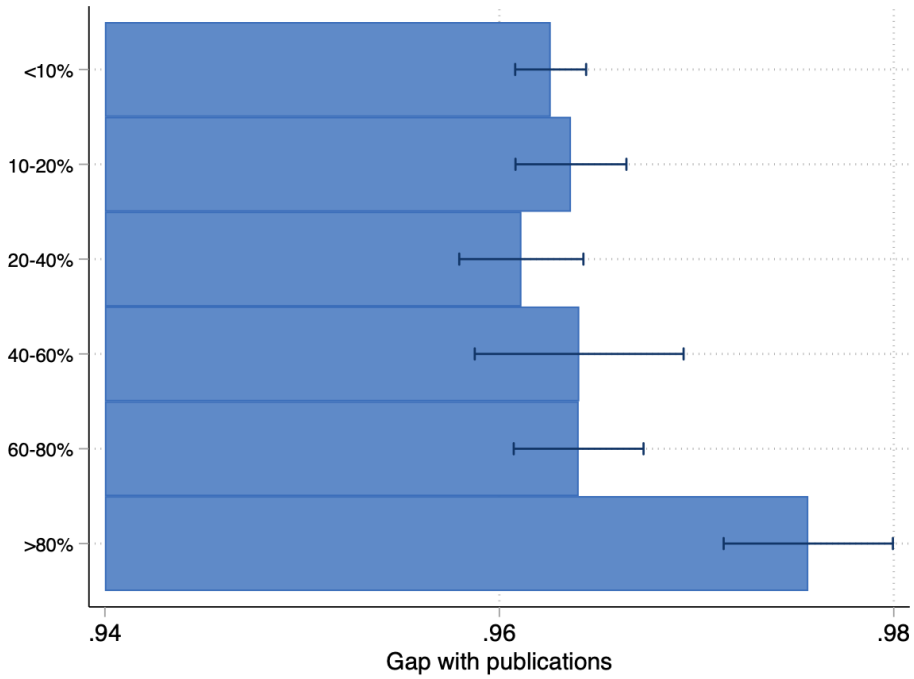
Notes: The figure shows averages and 95-percent confidence intervals of the gap between syllabi and publications by share students with parents in the top income percentile in the school. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Percentiles are based on the national income distribution. Estimates are obtained pooling data for the years 1992 to 2018, and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on median parental income at the school level is taken from Chetty et al. (2019).

Figure 11. Gap with Publications And Share of Minority Students



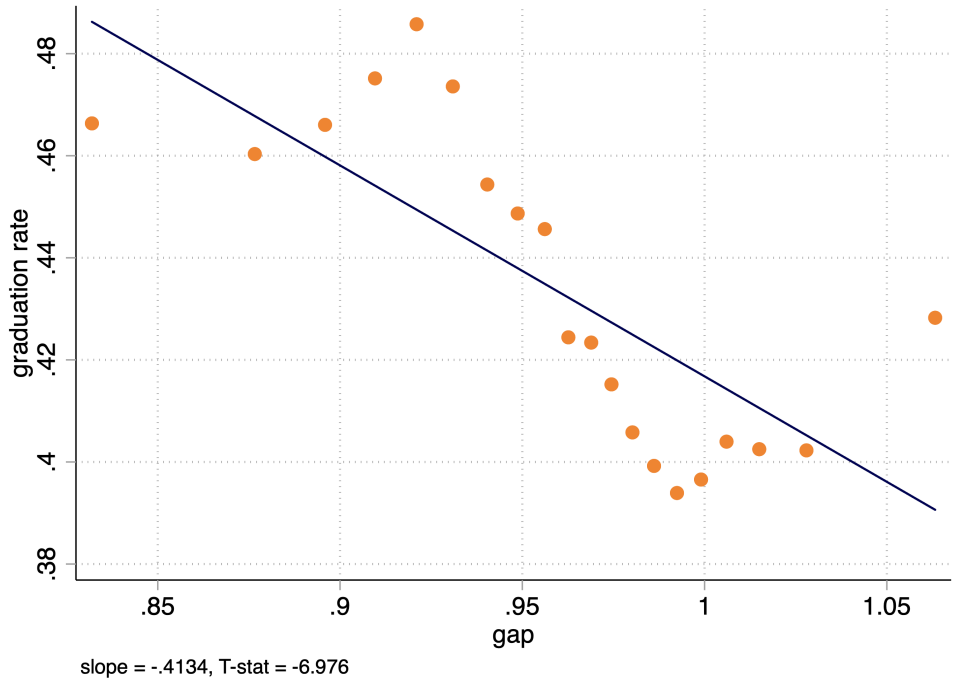
Notes: The figure shows a binned scatterplot of the gap between syllabi and publications and the share of students who are either Black or Hispanic. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained pooling data for the years 1992 to 2018, and controlling for field and syllabus year fixed effects. Information on students' race and ethnicity at the school level is taken from Chetty et al. (2019).

Figure 12. Gap with Publications, By Schools' Share of Minority Students



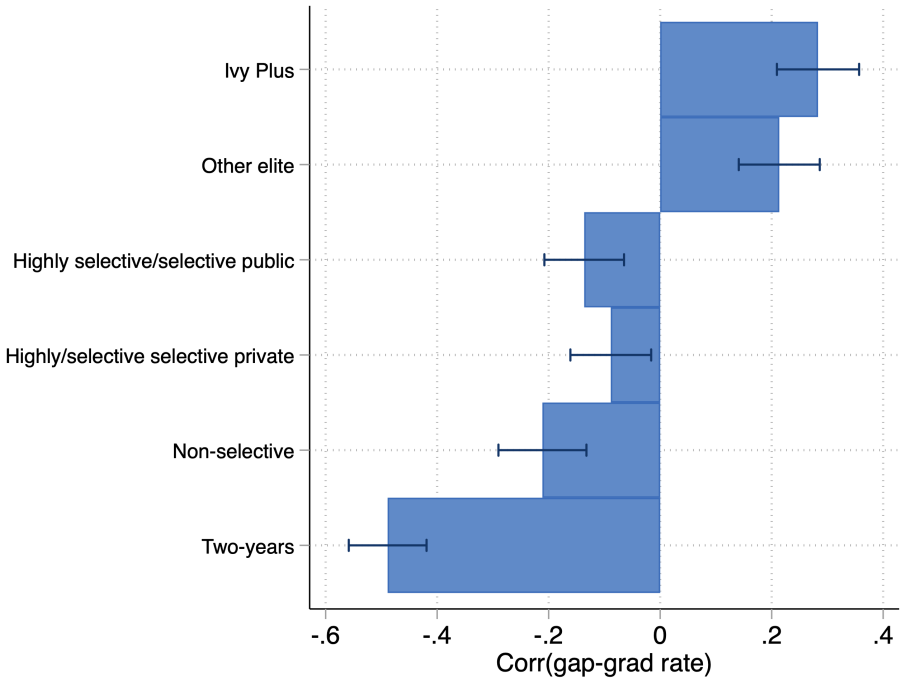
Notes: The figure shows averages and 95-percent confidence intervals of the gap between syllabi and publications by share of students who are either Black or Hispanic. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained pooling data for the years 1992 to 2018, and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on students' race and ethnicity at the school level is taken from Chetty et al. (2019).

Figure 13. Gap with Publications and Graduation Rate



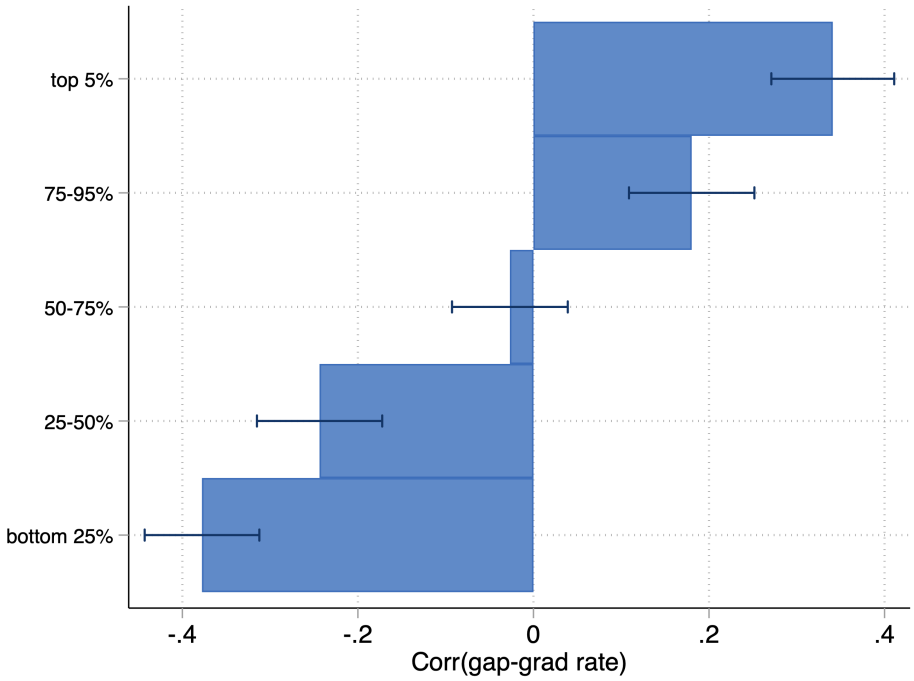
Notes: The figure shows a binned scatterplot of graduation rates in 2002 and 2013 and the gap between syllabi and publications. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained pooling data for the years 2002 and 2013, and controlling for field and syllabus year fixed effects. Information on graduation rates at the school level is taken from Chetty et al. (2019).

Figure 14. Gap with Publications and Graduation Rate, By School Tier



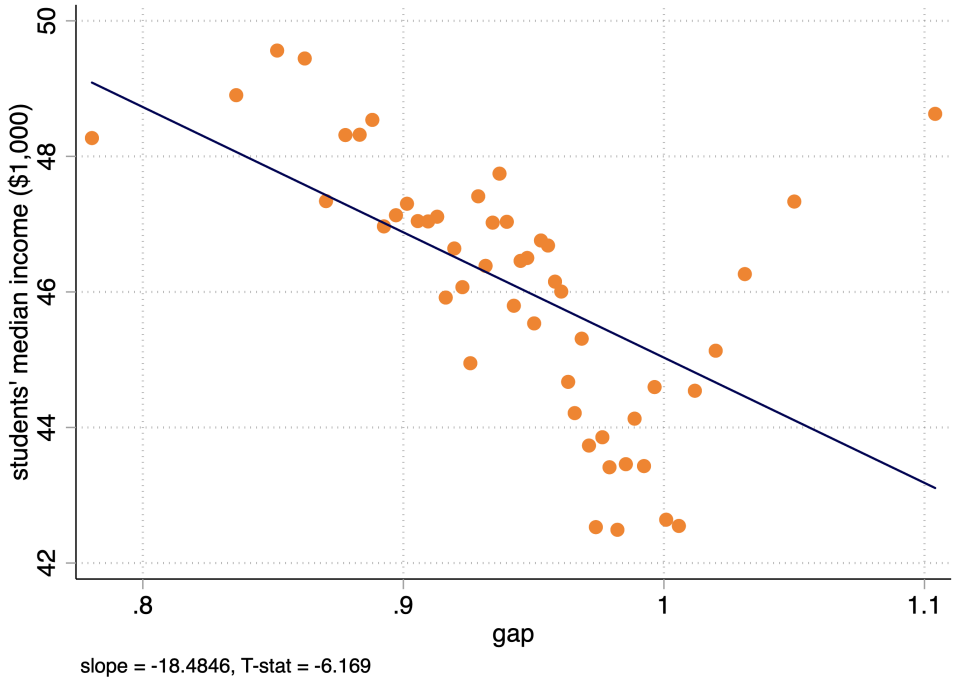
Notes: The figure shows estimates and 95-percent confidence intervals of the correlation between schools' graduation rates and the gap between syllabi and publications, separately by school tier. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained pooling data for the years 2002 and 2013, and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on median parental income at the school level is taken from Chetty et al. (2019).

Figure 15. Gap with Publications and Graduation Rate, By Schools' Median Parental Income



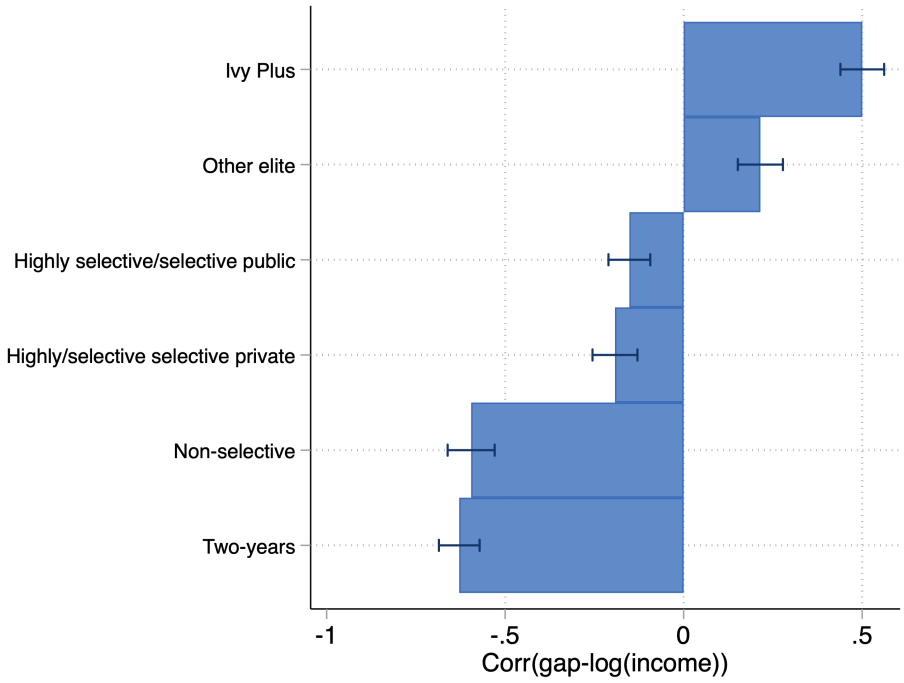
Notes: The figure shows estimates and 95-percent confidence intervals of the correlation between schools' graduation rates and the gap between syllabi and publications, separately by percentile of median parental income at the school level. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Percentiles are calculated using the distribution of median parental incomes across all schools. Estimates are obtained pooling data for the years 1992 to 2018, and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on median parental income at the school level is taken from Chetty et al. (2019).

Figure 16. Gap with Publications and Income



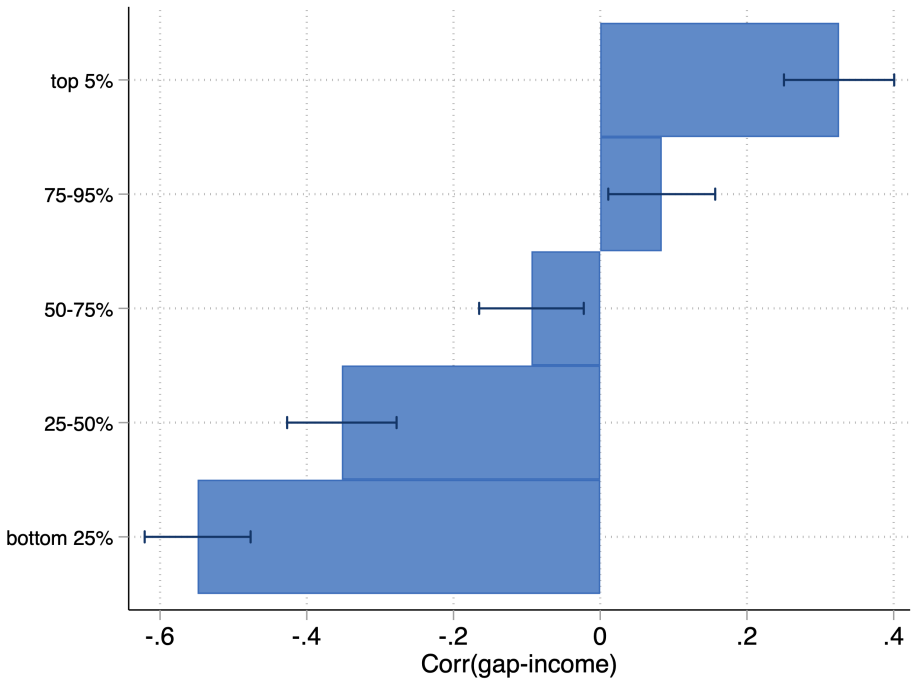
Notes: The figure shows a binned scatterplot of 2014 incomes of students who graduated from each school between 2002 and 2007 and the gap between syllabi and publications. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained using syllabus data from 2004 and controlling for field fixed effects. Information on student incomes at the school level is taken from Chetty et al. (2019).

Figure 17. Gap with Publications and Income, By School Tier



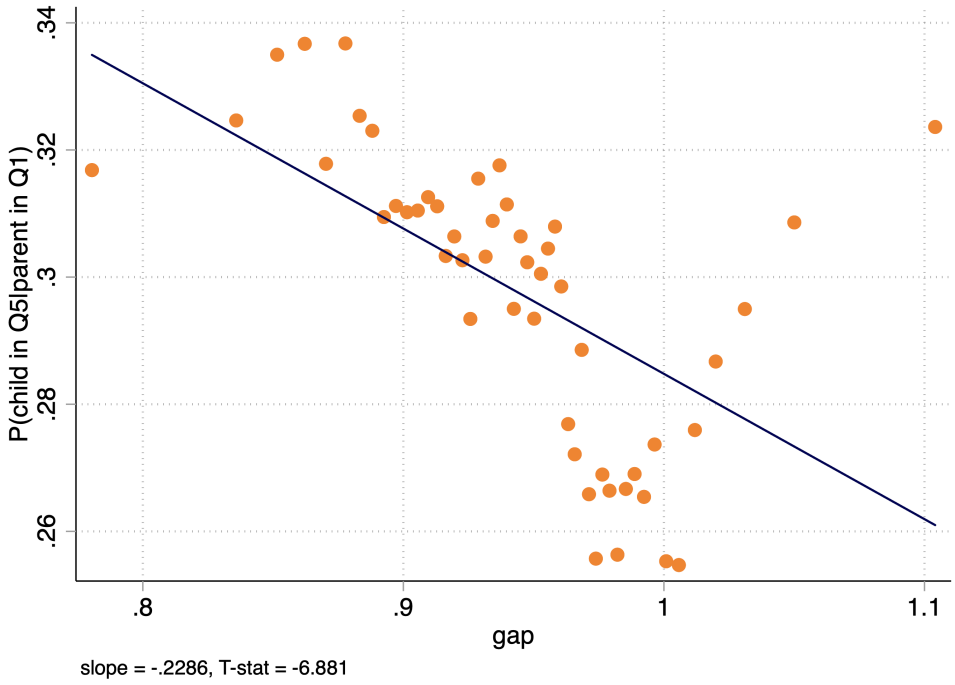
Notes: The figure shows estimates and 95-percent confidence intervals of the correlation between 2014 incomes of students who graduated fromn each school between 2002 and 2007 and the gap between syllabi and publications, separately by school tier. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained using syllabus data from 20041 and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on student incomes at the school level is taken from Chetty et al. (2019).

Figure 18. Gap with Publications and Income, By Schools' Median Parental Income



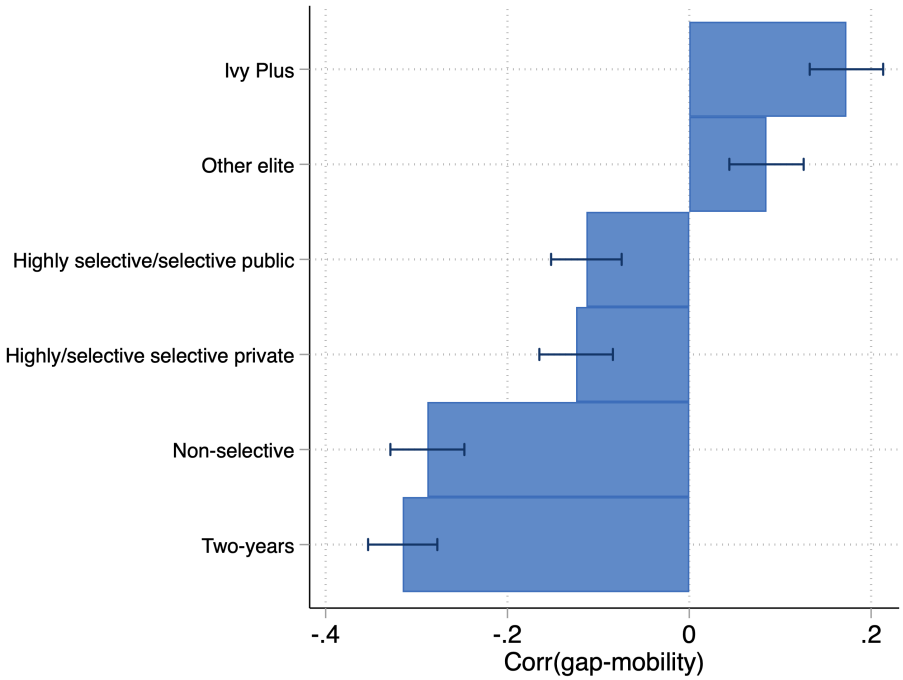
Notes: The figure shows estimates and 95-percent confidence intervals of the correlation between 2014 incomes of students who graduated from each school between 2002 and 2007 and the gap between syllabi and publications, separately by percentile of median parental income at the school level. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Percentiles are calculated using the distribution of median parental incomes across all schools. Estimates are obtained using syllabus data from 2004 and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on student incomes at the school level is taken from Chetty et al. (2019).

Figure 19. Gap with Publications and Intergenerational Mobility



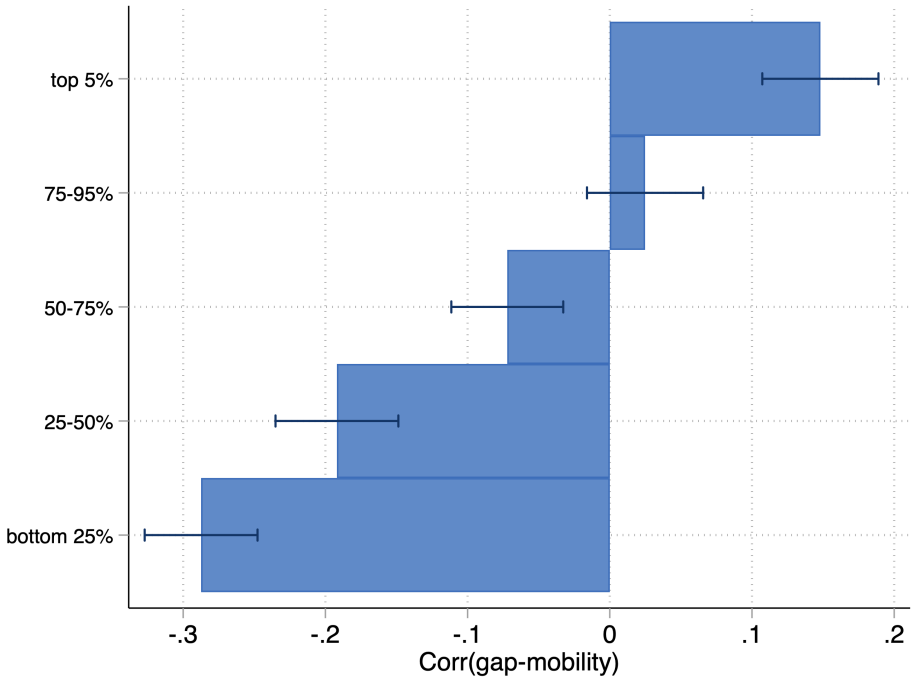
Notes: The figure shows a binned scatterplot of 2014 intergenerational mobility measures of students who graduated from each school between 2002 and 2007 and the gap between syllabi and publications. Intergenerational mobility is defined as the probability that a student with parents in the top 20 percent of the national income distribution reaches the top 20 percent of the income distribution. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained using syllabus data from 2004 and controlling for field fixed effects. Information on intergenerational mobility at the school level is taken from Chetty et al. (2019).

Figure 20. Gap with Publications and Intergenerational Mobility, By School Tier



Notes: The figure shows estimates and 95-percent confidence intervals of the correlation between 2014 intergenerational mobility measures of students who graduated from each school between 2002 and 2007 and the gap between syllabi and publications, separately by school tier. Intergenerational mobility is defined as the probability that a student with parents in the top 20 percent of the national income distribution reaches the top 20 percent of the income distribution. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Estimates are obtained using syllabus data from 2004 and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on intergenerational mobility at the school level is taken from Chetty et al. (2019).

Figure 21. Gap with Publications and Intergenerational Mobility, By Schools' Median Parental Income



Notes: The figure shows estimates and 95-percent confidence intervals of the correlation between 2014 intergenerational mobility measures of students who graduated from each school between 2002 and 2007 and the gap between syllabi and publications, separately by percentile of median parental income at the school level. Intergenerational mobility is defined as the probability that a student with parents in the top 20 percent of the national income distribution reaches the top 20 percent of the income distribution. The gap is defined as the ratio between syllabi's cosine similarity with publications 14 and 15 years prior to the syllabus date and the cosine similarity with publications one and two years prior. Percentiles are calculated using the distribution of median parental incomes across all schools. Estimates are obtained using syllabus data from 2004 and controlling for field and syllabus year fixed effects. Standard errors are clustered at the school-by-field level. Information on intergenerational mobility at the school level is taken from Chetty et al. (2019).

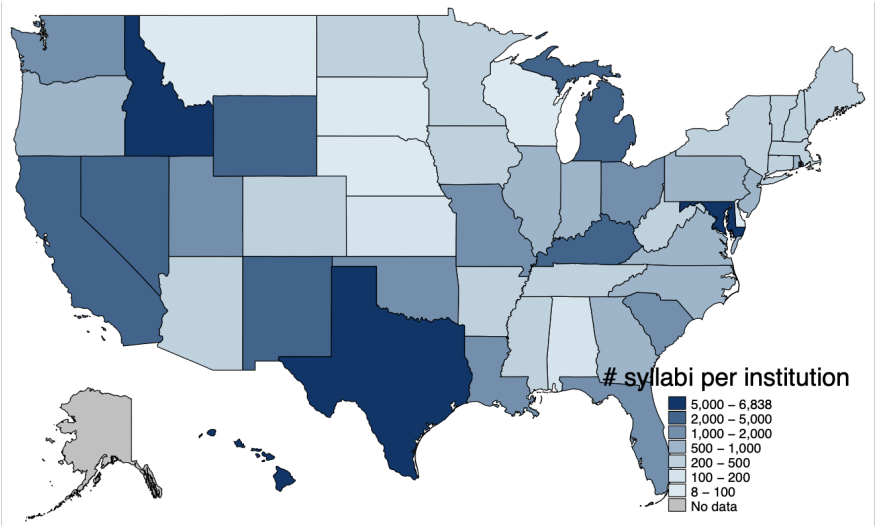
Appendix

For online publication only

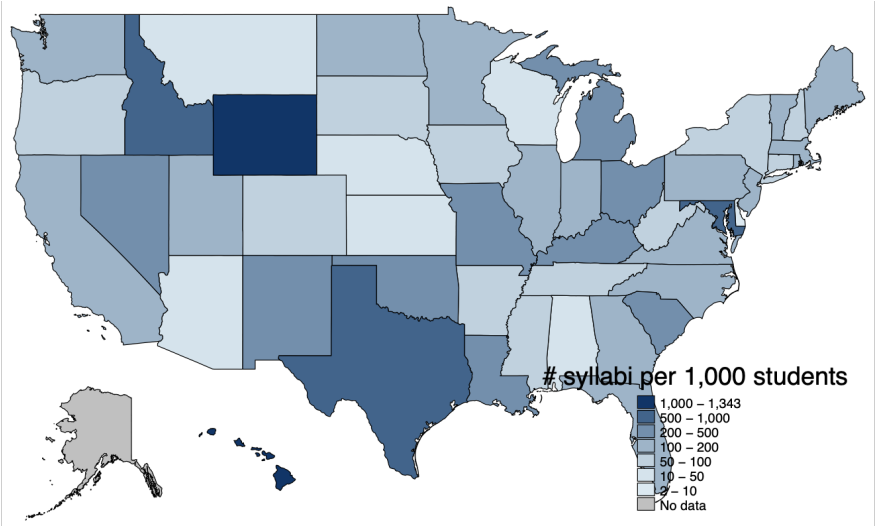
Additional Tables and Figures

Figure AI. Syllabi Across The United States

Panel a) Number of Syllabi Per Institution

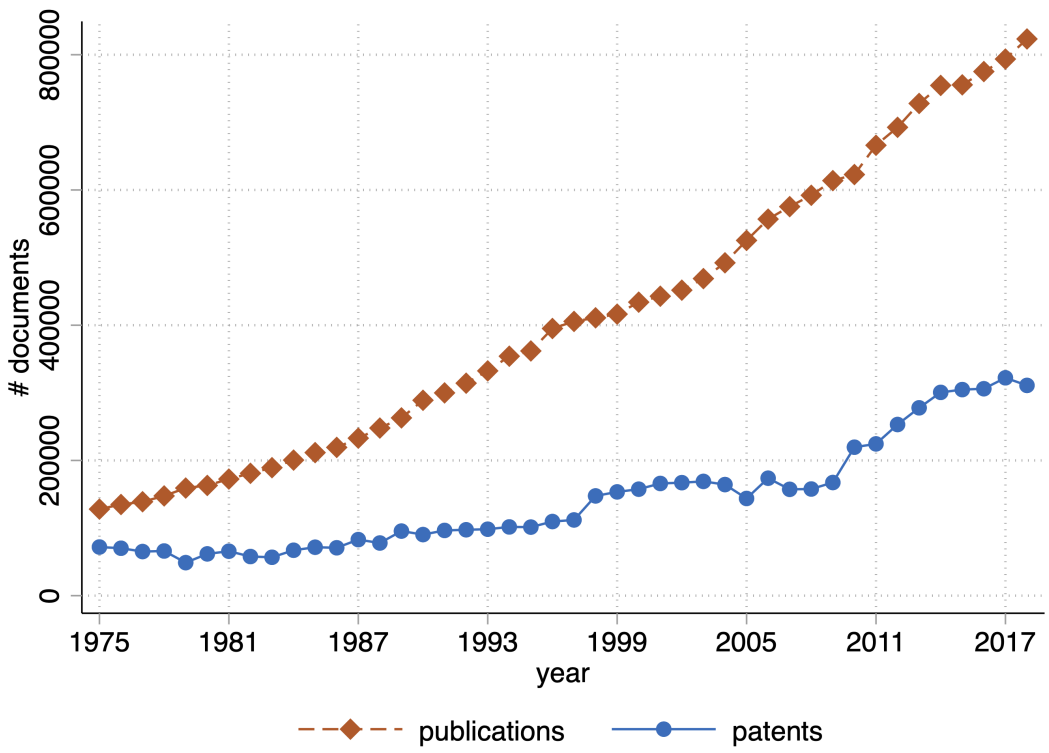


Panel b) Number of Syllabi Per Enrolled Student



Note: The map plots the number of syllabi taught between 1992 and 2018, per IPEDS institution (top panel) and per enrolled student (bottom panel). Enrollment figures from the year 2013 are taken from Chetty et al. (2019).

Figure AII. Academic Publications And Patents Across Time



Note: Counts of patents and academic publications per year.